Unfairness in Recommender Systems

To what extent do content-based recommendation models suffer from unfairness, and how does this differ from collaborative filtering?

Filip Angheluta
Supervisor(s): Masoud Mansoury

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: Filip Angheluta
Final project course: CSE3000 Research Project
Thesis committee: -

Abstract

Fairness in recommender systems is an increasingly critical concern as these models mediate access to information, opportunities, and visibility. While collaborative filtering (CF) approaches have been extensively scrutinized for popularity bias and unfair exposure, the fairness properties of content-based recommendation (CBR) models remain underexplored. In this work, we present a comparative evaluation of CF and CBR models—introducing a modular, feature-fused content-based recommender (MultiFuseCB)—on MovieLens 1M and Amazon Beauty datasets. We systematically analyze how the selection and weighting of content features, as well as the choice of embedding models, affect both recommendation accuracy and fairness, using metrics such as item coverage and popularity bias. Our results show that CBR models, with appropriate feature engineering, can achieve competitive accuracy while substantially improving fairness relative to CF baselines. We further demonstrate that certain features (e.g., year, genre, plot) and embedding choices can be leveraged to promote more equitable item exposure. These findings provide actionable insights for designing fairer content-based recommenders and highlight the importance of feature selection and model tuning in achieving both accuracy and fairness.

1 Introduction

Recommendation systems play a central role in shaping online user experiences, influencing what media we consume, which products we see, and whose content gains visibility. As such systems gain influence, concerns over their fairness have come to the forefront of research and policy discussions [3,6,17]. A large body of work has studied collaborative filtering (CF) methods—known for their reliance on historical user-item interactions—for their propensity to amplify popularity bias and reinforce inequalities in exposure [8, 16, 20].

Less attention, however, has been given to content-based recommendation (CBR) approaches, which rely on item attributes such as text, metadata, or categories. While often assumed to be more "neutral," these models can encode systemic bias present in underlying content features or metadata distributions [4, 12]. Moreover, CBR systems may reflect implicit value judgments about which content characteristics matter, leading to underexplored fairness implications. Importantly, CBR models are structurally independent from collaborative signals, which may constrain predictive accuracy—especially when content features are sparse—but may also offer fairness advantages by reducing susceptibility to feedback loops or historical interaction biases.

Recent advances in natural language processing, particularly the development of powerful pre-trained Sentence Transformer models [11], provide new opportunities to enhance content-based recommenders. By producing rich, semantically meaningful embeddings from diverse textual and metadata features, these models enable more nuanced and flexible representations of items, thereby improving the expressiveness and effectiveness of content-based approaches.

In this paper, we aim to systematically explore the fairness properties of content-based recommendation systems relative to collaborative filtering methods. To guide our analysis, we pose the following research questions:

- **RQ1:** *To what extent do content-based recommenders exhibit lower unfairness—such as reduced popularity bias or more equitable exposure—compared to collaborative filtering methods?*
- **RQ2:** *What are the trade-offs between accuracy and fairness in content-based versus collaborative filtering recommenders?*
- **RQ3:** *How does the design of a content-based recommender, particularly the choice and weighting of content features, affect fairness and accuracy trade-offs?*

To investigate these questions, we developed a custom content-based recommender system that relies exclusively on item-side information. Our approach computes item representations using multiple content features—including textual descriptions, genres, and metadata fields—while explicitly excluding any collaborative interaction data. For each feature, we generate embeddings using state-of-the-art pre-trained SentenceTransformer models from `sbert.net` [11], and then fuse them into a unified item embedding via a weighted combination. The weights of these features are treated as tunable parameters, optimized to improve recommendation quality on a validation set.

We compare this system, *MultiFuseCB*, against several collaborative filtering baselines (e.g. BPR, NeuMF) on two benchmark datasets: MovieLens 1M (enriched with external metadata via the OMDb API) and Amazon Beauty Reviews. These datasets allow us to assess the generalizability of findings across domains with varying content richness.

To evaluate performance and fairness, we adopt a multi-dimensional evaluation protocol. Accuracy is measured using standard top-K ranking metrics (Hit Rate and NDCG). For fairness, we examine both user-side metrics (e.g., intra-group ranking parity, individual consistency) and provider-side metrics (e.g., exposure disparity, popularity bias) [1, 13, 18]. This allows us to assess not only whether content-based recommenders are fairer than their CF counterparts, but also how different design choices influence fairness–accuracy trade-offs.

Through this work, we aim to shed light on the comparative strengths and weaknesses of content-based and collaborative filtering models from a fairness perspective, and to offer insights into the conditions under which content-based models may serve as more equitable alternatives.

2 Related Work

2.1 Fairness in Collaborative Filtering

Collaborative filtering (CF) models have long dominated recommender systems research and practice, yet they are increasingly scrutinized for perpetuating popularity bias and limiting the exposure of less popular or niche items [16, 20].

This bias is largely attributed to feedback loops inherent in user interaction data, where popular items receive more recommendations and thus become even more dominant over time. To address these issues, researchers have developed a variety of mitigation strategies, including model-centric approaches [8], ranking-based interventions [1, 13], and counterfactual techniques [16]. These methods aim to rebalance recommendation outcomes by explicitly accounting for exposure disparities and promoting more equitable item visibility.

2.2 Fairness in Content-Based Recommendation

Content-based recommendation (CBR) systems, which rely on item attributes rather than user interaction histories, are often assumed to be less susceptible to the feedback loops that plague CF. However, recent research has shown that CBR models can also propagate or even amplify biases present in item metadata, such as genre, author, or category distributions [4, 5, 9]. These biases may stem from the overrepresentation of certain content features or from the ways in which features are selected and weighted. As highlighted by [12] and [4], fairness in algorithmic systems is not solely a technical challenge—it also involves how social values and structural inequalities are reflected in feature design and data collection. A persistent challenge for CBR research is the lack of rich, high-quality item data in many domains, which limits the ability to develop and evaluate fair content-based recommenders at scale [9]. Despite their different mechanisms, both CF and CBR models can encode and amplify different forms of unfairness, necessitating careful attention to both model architecture and feature engineering.

2.3 Fairness Metrics and Evaluation Frameworks

A broad spectrum of fairness metrics has been proposed to evaluate recommender systems from multiple perspectives. These include exposure disparity and ranking parity for users [1,13], as well as item coverage, popularity bias, and provider fairness for items [18,19]. Multi-objective optimization frameworks have been developed to balance accuracy with fairness for diverse stakeholders [2, 17]. Despite these advances, purely quantitative metrics often fail to capture the nuanced perceptions of fairness among real users [3, 7], underscoring the importance of complementary qualitative or user-centered evaluation methods [10, 15].

2.4 Recent Advances and Open Challenges

Recent advances in fairness-aware recommendation have emphasized the need for systematic, empirical comparisons between CF and CBR models [5]. While the literature has predominantly focused on CF, there is growing recognition of the unique challenges and opportunities for fairness in CBR, particularly as new embedding techniques and feature engineering approaches emerge. However, one of the main obstacles to advancing fairness research in CBR remains the scarcity of datasets with sufficiently rich and diverse item features, which limits the generalizability and practical impact of many studies. Large-scale empirical studies of fairness trade-offs in CBR remain scarce, highlighting a critical gap in the literature and an important direction for future research.

3 Methodology

We evaluate fairness and accuracy across collaborative filtering (CF) and content-based recommendation (CBR) paradigms using the RecBole framework. As there are limited established CBR models available for fair comparison, we benchmark our proposed content-based model against representative CF models.

3.1 Recommender Models

We evaluate a mix of collaborative filtering (CF) baselines, a content-based model, and a random recommender. All CF baselines are implemented using the RecBole framework, while our content-based model, *MultiFuseCB*, is implemented independently as described in the Experimental Setup section.

Random Recommendation:

- **Random:** A non-personalized baseline that recommends items by sampling uniformly at random from the entire item pool. While it exhibits extremely low accuracy, it is theoretically the most fair model in terms of item exposure and popularity bias, serving as a useful lower-bound for both accuracy and fairness evaluations.

Collaborative Filtering (CF) Baselines:

- **Bayesian Personalized Ranking (BPR):** An implicit-feedback CF model that optimizes pairwise ranking by encouraging observed (positive) interactions to be ranked higher than unobserved (negative) ones. It is particularly effective for top-N recommendation tasks.
- **Neural Matrix Factorization (NeuMF):** A neural extension of matrix factorization that integrates both generalized matrix factorization and multi-layer perceptrons to capture both linear and nonlinear user-item interaction patterns.
- **BERT4Rec** [14]: A state-of-the-art sequential recommender model based on the BERT architecture. It predicts future interactions by modeling sequences of user behavior through masked item prediction. Although it incorporates item embeddings, it learns solely from interaction histories and is considered a CF method in this study.

Content-Based Recommendation (CBR):

- **MultiFuseCB (proposed):** A modular content-based recommender that generates item embeddings by fusing multiple textual and numeric content features using pretrained transformers and learned feature weights. User embeddings are constructed by aggregating past interacted item embeddings using various weighting strategies. Unlike CF models, MultiFuseCB operates without collaborative signals, enabling a focused analysis of fairness implications from content-based recommendations.

3.2 Content-Based Model: *MultiFuseCB*

To investigate fairness in content-only systems, we developed *MultiFuseCB*, a content-based recommender that exclusively uses item features without collaborative signals.

Feature Extraction. On the MovieLens 1M dataset, we extracted seven features: *title*, *genre*, *year*, *plot*, *actors*, *director*, and *IMDb rating*. All features except *IMDb rating* were treated as text and encoded using pretrained **sentence transformers** from SBERT.net [11]. On the Amazon Beauty dataset, we used the following features: *title*, *description*, *details*, and *price*. Here, *price* was numeric, while the others were semantic.

Feature Selection. For both datasets, we followed a principled feature selection procedure. For each feature, we generated embeddings using several sentence transformers:

`'all-mpnet-base-v2'`,
`'all-roberta-large-v1'`,
`'paraphrase-MiniLM-L6-v2'`,
`'all-MiniLM-L12-v2'`,
`'all-distilroberta-v1'`,
`'multi-qa-mpnet-base-cos-v1'`, and
`'jinaai/jina-embeddings-v3'`.

Each model-feature pair was evaluated individually using HR@10 on a validation split. We retained features for which at least one embedding model significantly outperformed a naive baseline (HR@10 > 0.1). For example, we found that features such as movie length did not provide meaningful signal and were thus excluded.

Model	Actors	Director	Genre	Plot	Title	Year	IMDb
distilroberta-v1	0.210	0.174	0.238	0.170	0.240	0.267	-
MiniLM-L12-v2	0.214	0.170	0.252	0.151	0.214	**0.287**	-
mpnet-base-v2	**0.241**	**0.215**	0.239	0.166	**0.299**	0.265	-
roberta-large-v1	0.234	0.196	0.243	0.150	0.247	0.258	-
jina-v3	0.238	0.188	**0.260**	0.185	0.246	0.281	-
multi-qa-mpnet	0.239	0.213	0.234	**0.196**	0.242	0.286	-
paraphrase-MiniLM-L6	0.161	0.139	0.245	0.190	0.187	0.277	-
Harmonic	-	-	-	-	-	0.265	**0.203**

Table 1: HR@10 of each model-feature combination for MovieLens 1M.

Model	Title	Description	Details	Price
distilroberta-v1	0.240	0.170	0.238	-
MiniLM-L12-v2	0.214	0.151	0.252	-
mpnet-base-v2	**0.299**	0.166	0.239	-
roberta-large-v1	0.247	0.150	0.243	-
jina-v3	0.246	0.185	**0.260**	-
multi-qa-mpnet	0.242	**0.196**	0.234	-
paraphrase-MiniLM-L6	0.187	0.190	0.245	-
Harmonic	-	-	-	**0.126**

Table 2: HR@10 of each model-feature combination for Amazon Beauty.

Feature Encoding. The best-performing model for each retained feature was selected. The *year* feature was encoded semantically (e.g., treating "1994" as a token), as this approach outperformed numeric normalization. Numeric features such as *IMDb rating* and *price* were embedded using a **harmonic embedding** inspired by positional encodings in transformers. Given a normalized scalar value $x_{\text{norm}} \in [0, 1]$, we computed:

$$\mathbf{e}_x = \begin{bmatrix} \sin(2^0 \pi x_{\text{norm}}),\ \cos(2^0 \pi x_{\text{norm}}), \\ \sin(2^1 \pi x_{\text{norm}}),\ \cos(2^1 \pi x_{\text{norm}}), \\ \dots \\ \sin(2^{d-1} \pi x_{\text{norm}}),\ \cos(2^{d-1} \pi x_{\text{norm}}) \end{bmatrix}$$

We used $d = 64$, resulting in a 128-dimensional vector to balance the dimensionality of semantic features.

Item Embedding Construction. Let $\mathbf{e}_i^{(k)}$ denote the embedding of item i for feature k. We compute the final item embedding by applying a learnable scalar weight α_k to each feature embedding and concatenating the results, followed by normalization:

$$\mathbf{v}_i = \text{Normalize}\left(\bigoplus_{k=1}^{K} \alpha_k \cdot \mathbf{e}_i^{(k)} \right)$$

where $\bigoplus$ denotes concatenation. This allows the model to adjust the relative importance of each feature during training.

User Embedding Construction. User embeddings are computed by aggregating the embeddings of previously interacted items $\mathbf{v}_j$. We evaluated four aggregation schemes:

- **Simple Average**: Treats all items equally:

$$\mathbf{u} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{v}_j$$

- **Rating-Based Weighting**: Uses user-provided ratings r_j:

$$\mathbf{u} = \frac{1}{\sum_{j=1}^{N} r_j} \sum_{j=1}^{N} r_j \cdot \mathbf{v}_j$$

- **Linear Recency Weighting**: Normalizes interaction timestamps t_j:

$$a_j = \frac{t_j - t_{\min}}{t_{\max} - t_{\min} + \epsilon}, \quad \mathbf{u} = \frac{1}{\sum_{j=1}^{N} a_j} \sum_{j=1}^{N} a_j \cdot \mathbf{v}_j$$

- **Exponential Decay Recency Weighting**: Applies exponential bias:

$$w_j = \exp(-\lambda \cdot (1 - a_j)), \quad \mathbf{u} = \frac{1}{\sum_{j=1}^{N} w_j} \sum_{j=1}^{N} w_j \cdot \mathbf{v}_j$$

In all cases, candidate items are ranked based on the cosine similarity between the user embedding $\mathbf{u}$ and item embeddings $\mathbf{v}_i$. The exponential decay method with tuned λ yielded the highest recommendation accuracy across both datasets.

Table 3: Statistics of the Datasets after Preprocessing

Dataset	#Users	#Items	#Interactions
MovieLens 1M	6,040	3,859	991,994
Amazon Beauty	4,374	4,761	16,359

3.3 Datasets

We evaluate our models on two publicly available datasets: MovieLens 1M and Amazon Beauty Reviews. These datasets differ significantly in content richness and sparsity, enabling evaluation across diverse recommendation contexts.

MovieLens 1M: A widely-used benchmark dataset containing one million explicit ratings from over 6,000 users across nearly 4,000 movies. We retained only movies with a valid IMDb ID and enriched them using the OMDb API, collecting a comprehensive set of content features including title, genre, year, plot, actors, director, and IMDb rating. MovieLens is notably rich in metadata, making it well suited for content-based recommendation.

Amazon Beauty Reviews: A subset of Amazon product reviews in the beauty domain, containing user-item interactions, textual reviews, and sparse item metadata such as *brand* and *category*. Compared to MovieLens, this dataset is considerably more sparse-both in terms of metadata availability and user interaction density. A large number of items lacked sufficient metadata and many users had very few interactions. To ensure consistency across datasets, we filtered out items without adequate metadata and removed users with fewer than 3 interactions. This filtering had a significant effect on Amazon Beauty due to its high sparsity.

3.4 Accuracy Metrics

To assess predictive performance, we adopt standard top-K ranking metrics:

- **Hit Rate@10 (HR@10)**: Measures whether the ground-truth item appears in the top-10 recommendations.
- **Normalized Discounted Cumulative Gain@10 (NDCG@10)**: Captures ranking quality by assigning higher scores to hits that appear earlier in the ranked list.

We used a leave-two-out evaluation strategy: for each user, the most recent interaction is held out for testing, the second-most recent for validation, and the remainder for training. During evaluation, models rank a candidate set of 100 items—one ground-truth item and 99 randomly sampled negatives.

3.5 Fairness Metrics

To evaluate fairness in recommendations, we employ a set of top-k metrics that assess exposure and accuracy disparities from both the user and item/provider perspectives. All metrics are normalized to ensure comparability across datasets. However, for readability, we report *Average Popularity (AvgPop)* as a percentage of total interactions.

Item Fairness Metrics:

- **Item Coverage:** Fraction of unique items that appear in the top-k recommendations across all users; higher values indicate broader exposure.
- **Entropy:** Measures the uncertainty of the item exposure distribution; high entropy suggests more uniform exposure across items.
- **Gini Coefficient:** Quantifies inequality in item exposure; lower values indicate a fairer distribution.
- **Average Popularity:** The average percentage of total interactions that the recommended items account for. Higher values indicate a stronger bias toward globally popular items, while lower values suggest more balanced exposure across the item catalog.
- **Tail%:** Proportion of recommendations assigned to long-tail (less popular) items.
- **Head%:** Proportion of recommendations assigned to head (most popular) items.

User Fairness Metrics:

- **STD (Standard Deviation):** Standard deviation of per-group accuracy; lower values reflect greater fairness across groups. We only calculate this for MovieLens 1M, as it gives access to gender group.

These metrics enable a nuanced evaluation of fairness by examining both who gets recommended what (user fairness) and which items gain visibility (item fairness), shedding light on how different modeling strategies influence recommendation equity.

3.6 Evaluation Protocol

We adopt a leave-two-out strategy: for each user, we retain the two most recent interactions as validation and test points, respectively, and use the remainder for training. During evaluation, we construct a candidate set consisting of the ground-truth item and 99 randomly sampled negative items the user has not interacted with. Models rank these 100 items, and accuracy is measured by the position of the ground-truth item.

Hyperparameters for each model were tuned based on validation performance. MultiFuseCB's transformer encoders were selected per feature using validation NDCG. The recency bias parameter λ for exponential decay was also tuned based on validation accuracy.

4 Results

4.1 Recommendation Accuracy

The comparative performance of each model on MovieLens 1M and Amazon Beauty is summarized in Table 4. On MovieLens 1M, the baseline Random model yields the lowest scores, as expected. Among collaborative filtering approaches, BERT4Rec demonstrates the strongest performance, achieving the highest HR@10 and NDCG@10, reflecting its capacity to model sequential user behavior effectively. The BPR and NeuMF models also perform well, though they are surpassed by BERT4Rec. Notably, our content-based model, *MultiFuseCB*, achieves competitive results-outperforming BPR and approaching NeuMF, despite relying solely on item features rather than user interaction data.

On the sparser Amazon Beauty dataset, *MultiFuseCB* leads all considered models, outperforming both BPR and NeuMF,

as well as the sequential model BERT4Rec. This result highlights the robustness of content-based recommendation in scenarios where user interaction data is limited or sparse.

These findings suggest that content-based recommendation can be highly effective when high-quality item features are available, especially in data-sparse or cold-start settings where collaborative filtering models may struggle. Across both datasets, *MultiFuseCB* demonstrates a strong ability to surface relevant items and maintain high-quality rankings, confirming the practical value of feature-rich content-based approaches in diverse recommendation contexts.

Table 4: Top-10 recommendation accuracy (HR@10 and NDCG@10) for each model on MovieLens 1M and Amazon Beauty.

Model	HR@10	NDCG@10
MovieLens 1M		
Random	0.095	0.055
BPR	0.416	0.242
NeuMF	0.544	0.317
BERT4Rec	0.697	0.481
MultiFuseCB	**0.480**	**0.277**
Amazon Beauty		
Random	0.103	0.058
BPR	0.283	0.174
NeuMF	0.271	0.167
BERT4Rec	0.302	0.186
MultiFuseCB	**0.345**	**0.214**

4.2 Fairness Results (RQ1)

We present fairness outcomes for all evaluated models using a comprehensive set of user- and item-centric metrics, as defined in Section 3.5. The results, reported in Table 5, reveal distinct patterns across models and datasets.

On MovieLens 1M, the Random baseline achieves high item coverage (IC = 0.921), entropy (12.31), and low popularity bias (AvgPop = 0.021), but this comes at the cost of weak recommendation quality. In contrast, collaborative filtering models (BPR, NeuMF, BERT4Rec) exhibit much lower item coverage and entropy, alongside higher popularity bias and a stronger focus on head items. *MultiFuseCB* strikes a notable balance: it offers high item coverage (0.823) and entropy (10.69), while significantly reducing popularity bias (AvgPop = 0.052) compared to collaborative approaches. MultiFuseCB also achieves a lower Gini index (0.609) and lower standard deviation (0.0099) than BPR, NeuMF, and BERT4Rec, indicating a more equitable and less variable distribution of recommendations.

The standard deviation (STD) reported in the table reflects the variability of hit rate (HR@10) across gender groups. A lower standard deviation suggests that the model provides recommendations with similar effectiveness to users of different genders. The results show that *MultiFuseCB* achieves a lower standard deviation than the collaborative filtering baselines, indicating more equitable treatment of gender groups.

On Amazon Beauty, similar trends are observed, with the Random model again showing the fairest distribution (IC = 0.938, AvgPop = 0.016), but with limited practical utility. Collaborative filtering models (BPR, NeuMF, BERT4Rec) display low item coverage, high popularity bias, and a strong preference for head items. *MultiFuseCB* stands out with exceptional item coverage (0.908), moderate entropy (9.08), and a substantially lower Gini index (0.434) compared to collaborative models. *MultiFuseCB*'s popularity bias (AvgPop = 0.152) is also markedly lower than that of BPR, NeuMF, and BERT4Rec, and it provides a more balanced exposure between tail and head items.

Across both datasets, *MultiFuseCB* consistently outperforms collaborative filtering models on fairness metrics, providing higher item coverage, lower popularity bias, and a more equitable distribution of recommendations. While collaborative filtering models, especially BERT4Rec, offer strong accuracy as shown in previous results, they tend to reinforce popularity bias and limit exposure to less popular items. *MultiFuseCB* demonstrates that content-based recommendation can achieve both competitive accuracy and superior fairness, particularly in scenarios where user interaction data is sparse or limited. These results affirm the value of content-based approaches for promoting fairness in recommender systems.

Table 5: Fairness metrics (@10) for all models. ↑ means higher is better; ↓ means lower is better. IC = Item Coverage, Ent. = Entropy, Gini = Gini Index, Pop = AvgPop, Tail = Tail%, Head = Head%, STD = Std. Dev.

Data	Model	IC ↑	Ent. ↑	Gini ↓	Pop ↓	Tail ↑	Head ↓	STD ↓
ML-1M	Random	0.921	12.31	0.210	0.021	0.888	0.112	0.0076
	BPR	0.312	6.80	0.710	0.268	0.621	0.379	0.0519
	NeuMF	0.330	7.85	0.683	0.242	0.587	0.411	0.0356
	BERT4Rec	0.412	9.20	0.645	0.184	0.431	0.569	0.0227
	MultiFuseCB	0.823	10.69	0.609	0.052	0.735	0.265	0.0099
Beauty	Random	0.938	12.89	0.205	0.016	0.905	0.095	-
	BPR	0.218	5.96	0.735	0.295	0.662	0.338	-
	NeuMF	0.225	6.43	0.721	0.274	0.628	0.372	-
	BERT4Rec	0.276	7.68	0.683	0.241	0.478	0.522	-
	MultiFuseCB	0.908	9.08	0.434	0.152	0.695	0.305	-

4.3 Trade-offs (RQ2)

To evaluate the trade-offs between accuracy and fairness, we examine how each model balances recommendation performance (HR@10, NDCG@10) with both item- and user-level fairness metrics, as reported in Tables 4 and 5.

Item fairness. Collaborative filtering models, particularly BERT4Rec and NeuMF, achieve strong accuracy—e.g., BERT4Rec reaches HR@10 of 0.697 and NDCG@10 of 0.481 on MovieLens 1M. However, this performance comes at the cost of fairness. These models exhibit low item coverage (e.g., 0.412 for BERT4Rec), high average popularity (0.184), and high Gini coefficients, indicating a concentration of exposure on a small set of popular items. In contrast, *MultiFuseCB* improves fairness by significantly increasing item coverage (0.823), boosting entropy (10.69), and lowering average popularity (0.052), which signals broader and more balanced exposure-including a higher proportion of long-tail items (Tail% = 0.735).

User fairness. Beyond item-level exposure, we also consider the variability in recommendation accuracy across dif-

ferent user groups. Lower values of standard deviation (STD) suggest more equitable treatment. On MovieLens 1M, *MultiFuseCB* shows reduced disparity in accuracy across users (STD = 0.0099) compared to BPR (STD = 0.0519) and NeuMF (STD = 0.0356).

These results are largely in line with expectations from the literature and the known properties of content-based and collaborative filtering approaches. Content-based recommenders, by design, do not rely on user interaction histories, and thus are less prone to reinforcing popularity bias and other feedback loop effects that often arise in collaborative filtering systems. As a result, content-based models are generally expected to exhibit fairer item exposure and more equitable treatment of user groups. However, the accuracy of content-based methods is fundamentally limited by the richness and quality of the available item features. If item metadata is sparse, uninformative, or poorly aligned with user preferences, the model's ability to make relevant recommendations diminishes. Furthermore, the practical effectiveness of content-based recommendation varies by domain: for example, in settings like Amazon Beauty, users are unlikely to purchase many highly similar items (e.g., multiple variants of the same product), which inherently restricts the achievable accuracy of content-based approaches. In contrast, domains like movies or music-where users often consume a wide variety of items with overlapping features-are more amenable to content-based recommendation. Thus, while content-based models can mitigate certain fairness concerns, their performance ceiling is shaped by both the quality of item features and the specific patterns of user behavior in the application domain.

4.4 Impact of Content Feature Design on Fairness and Accuracy (RQ3)

Our analysis reveals how different content features influence recommendation accuracy and fairness metrics. Table 6 presents the performance of each feature based on the best-performing embedding model selected for accuracy (HR@10), as per our methodology.

Feature	HR@10	IC	AvgPop
Plot	0.196	0.81	0.024
Actors	0.241	0.58	0.047
Genre	0.260	0.88	0.026
Title	0.299	0.65	0.055
Year	0.287	0.94	0.026
Director	0.215	0.24	0.040
IMDb Rating	0.203	0.91	0.031

Table 6: Performance of each feature using the best embedding model for accuracy (HR@10). IC: Item Coverage; AvgPop: Average Popularity.

Key Findings

Accuracy vs. Fairness Trade-offs

Several features achieve high accuracy (HR@10) while maintaining strong fairness. For example, *Year* and *Genre* both demonstrate high HR@10 (0.287 and 0.260, respectively), excellent Item Coverage (IC: 0.94 and 0.88), and low Average Popularity (AvgPop: 0.026 for both). *IMDb Rating* also performs well in fairness (IC: 0.91, AvgPop: 0.031), though with slightly lower accuracy (HR@10: 0.203). *Plot* stands out for its strong fairness (IC: 0.81, AvgPop: 0.024) despite moderate accuracy (HR@10: 0.196). In contrast, *Director* exhibits the weakest performance overall, with low accuracy (HR@10: 0.215) and fairness (IC: 0.24).

Notably, the *Title* feature achieves high accuracy (HR@10: 0.299), but its Average Popularity (AvgPop: 0.055) is higher than that of other features. This may be due to the fact that famous movie titles are more likely to be mentioned in the training data of the sentence transformers used for embeddings, leading to a bias toward recommending popular items. As a result, while *Title* is effective for accurate recommendations, it may inadvertently reinforce popularity bias, highlighting the importance of considering both accuracy and fairness when selecting and weighting features in content-based recommenders.

Embedding Model Selection

For each feature, we considered multiple embedding models and selected the one that maximized accuracy (HR@10). However, experiments showed that alternative embedding models could yield better fairness at the cost of some accuracy. For instance, a different embedding model for *Actors* achieved HR@10 of 0.233, IC of 0.65, and AvgPop of 0.040. Although this configuration improved fairness, our study prioritized accuracy in the final model selection.

Implications for Fairness-aware Recommender Design

The results demonstrate that both the choice of content features and the embedding strategy play a critical role in shaping the trade-off between accuracy and fairness. Certain types of features can contribute more effectively to fair item exposure and balanced user treatment, while others may offer limited utility for either objective. Additionally, the observation that some embedding models can enhance fairness—albeit sometimes at a minor cost to accuracy—suggests an important direction for future research: developing and tuning embedding techniques that explicitly account for fairness considerations, alongside traditional performance goals.

5 Responsible Research

We are committed to conducting research that adheres to principles of transparency, fairness, and reproducibility. In this section, we outline the steps taken to ensure that our study meets responsible research standards.

5.1 Fairness Considerations

Fairness is a central focus of this study. We explicitly examine how content-based and collaborative filtering recommender systems may exhibit disparities in exposure and treatment of users and content providers. Our evaluation employs multiple fairness metrics that cover both user-level and item-level perspectives, such as Exposure Disparity, Popularity Bias, and Provider Fairness. These metrics are selected based on prior literature and are intended to reveal trade-offs between accuracy and equitable treatment.

We recognize that fairness is context-dependent and can be influenced by societal structures reflected in data. While our study uses public datasets (MovieLens 1M and Amazon Beauty Reviews), we acknowledge that these datasets may embed historical biases. Our findings are interpreted within the limitations of these data sources, and we avoid making normative claims about fairness beyond the scope of the datasets analyzed.

5.2 Transparency and Reproducibility

To promote reproducibility, we document all preprocessing steps, models used, and evaluation protocols in detail. The custom model introduced in this paper (*MultiFuseCB*) is described in a modular fashion to facilitate replication and adaptation. We used publicly available tools and frameworks such as RecBole and SentenceTransformers [11].

5.3 Limitations and Ethical Impact

Our study is limited to offline evaluations on two datasets, which may not capture real-world deployment dynamics such as long-term user feedback loops or strategic behavior. Additionally, our fairness analysis focuses on quantifiable metrics; it does not fully capture deeper social or cultural harms that may arise from algorithmic recommendation.

We stress that content-based approaches, while often considered more "neutral," can still propagate embedded biases in metadata or textual descriptions. Our work should not be interpreted as a complete solution to fairness in recommendation, but rather as a step toward a more nuanced understanding of its challenges and trade-offs.

5.4 Data Usage

Both MovieLens 1M and Amazon Beauty Reviews are publicly available and widely used for academic research. We filtered and enriched the MovieLens dataset using publicly accessible APIs (OMDB) and adhered to data usage policies. No personal or sensitive information was used in our experiments.

6 Discussion

6.1 Limitations of Standard Fairness Metrics

Metrics such as exposure disparity, popularity bias, and intra-group ranking parity are widely used to evaluate fairness in recommender systems [1, 13, 20]. These metrics offer structured and repeatable ways to measure disparities in item visibility and user treatment. However, they also come with limitations. In particular, they tend to be rigid and context-agnostic, often assuming that fairness can be captured through objective numerical criteria.

This framing may overlook more nuanced or subjective aspects of fairness that vary across users and domains. For instance, what counts as "fair" exposure may differ between a movie platform and a job-matching site. Moreover, such metrics typically focus on outcomes while neglecting how users perceive and interpret those outcomes [3]. These gaps raise questions about the sufficiency of current fairness evaluations and whether additional dimensions should be considered.

6.2 A Sociotechnical Perspective

From a sociotechnical perspective, fairness is not solely a property of algorithmic output but rather something that emerges from the interplay between models, users, and context [4, 12]. Recommender systems operate within social and institutional frameworks, and user interpretations of fairness can vary based on cultural, personal, or situational factors.

For example, a system that personalizes based on user history might be seen as helpful by some users and restrictive by others, particularly if personalization leads to the reinforcement of past behaviors. Similarly, what appears as fair treatment in aggregate may still be perceived as unfair by particular subgroups, especially if the system's logic remains opaque or difficult to contest. Exploring these kinds of tensions may provide a richer understanding of fairness beyond what metrics alone can offer.

6.3 Alternative Fairness Dimensions

To expand the lens of fairness evaluation, we consider user-centered dimensions that capture more experiential qualities of interaction. These include:

- **Perceived Representation:** Do users feel that the recommended items reflect their identities, preferences, or values?
- **Agency and Control:** Do users have meaningful ways to influence what is recommended to them or to correct misrepresentations?
- **Explanation Quality:** Are users able to understand why items were recommended? Does the system provide transparency or interpretability?

These dimensions may be particularly relevant in settings where personalization affects users' sense of autonomy or inclusion. While harder to quantify, such factors could be investigated through qualitative studies or hybrid methods that combine metric-based evaluation with user feedback.

6.4 Open Questions and Future Directions

The interplay between fairness metrics, user perception, and social context presents a complex space for further research. Questions remain about how best to integrate subjective notions of fairness into system design and evaluation, and whether trade-offs between different fairness goals can be meaningfully resolved. Developing broader frameworks that incorporate both quantitative and qualitative insights might offer one path forward, though doing so introduces its own methodological challenges.

By exploring these perspectives, we aim to contribute to an ongoing dialogue about fairness in recommendation. Rather than proposing a definitive model or framework, this discussion highlights the importance of remaining attentive to multiple definitions and experiences of fairness, especially as recommender systems continue to evolve and shape user interactions in diverse domains.

7 Future Work

There are several promising directions for extending this study. First, while *MultiFuseCB* achieves competitive fair-

ness outcomes, its accuracy may be further improved by incorporating richer and more consistent metadata. For instance, many items in the Amazon Beauty dataset lack detailed descriptive fields; curating or inferring this information could enhance model quality.

Second, the choice of sentence encoder significantly impacts the representation of textual item features. Future work could explore a broader range of sentence transformers, or fine-tune them specifically for the task of extracting meaningful item representations in recommender contexts.

Additionally, large language models (LLMs) offer opportunities to enrich item metadata with higher-level attributes. In the movie domain, for example, LLMs could generate fields such as tone, mood, or target audience, enabling deeper semantic modeling of user preferences.

Beyond technical improvements, future work could expand the empirical scope by applying the *MultiFuseCB* framework to additional datasets from diverse domains, such as Steam (video games), Last.fm (music), or Book-Crossing (books). These datasets present varying patterns of sparsity, metadata richness, and user-item interactions, offering a broader testbed for evaluating both fairness and recommendation quality.

Finally, to complement quantitative fairness metrics, future studies could include qualitative investigations—such as user surveys or interviews—to explore how different stakeholders (e.g., users, item providers) perceive the fairness of content-based, collaborative, and hybrid recommender systems. These insights would help ground fairness optimizations in real-world expectations and values.

8 Conclusion

This work provides a systematic comparison of content-based and collaborative filtering recommender systems from a fairness perspective. By evaluating a modular content-based approach across multiple datasets and a range of accuracy and fairness metrics, we demonstrate that content-based models can achieve a favorable balance between accuracy and fairness, often outperforming collaborative filtering baselines on key fairness indicators such as item coverage and exposure diversity.

Our analysis further shows that both the selection of content features and the choice of embedding models play a significant role in shaping these trade-offs. While this study prioritized accuracy in model selection, we observed that alternative configurations could improve fairness outcomes, suggesting that content-based recommenders offer flexible levers for fairness-aware system design. These findings highlight the importance of considering both feature engineering and fairness objectives when developing content-based recommendation systems, and open promising avenues for future work that explicitly optimizes for fairness alongside accuracy.

References

[1] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys 2019)*, pages 48–56, 2019.

[2] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. Balanced neighborhoods for multi-sided fairness in recommendation. In *Proceedings of the FATREC Workshop on Responsible Recommendation (FATREC)*, 2017.

[3] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*, pages 275–285. ACM, 2019.

[4] Mateusz Dolata, Stefan Feuerriegel, and Gerhard Schwabe. A sociotechnical view of algorithmic fairness. *Information Systems Journal*, 31(4):514–538, 2021.

[5] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Joshua D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All the cool kids, how do they fit in? popularity and demographic bias in recommender systems. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[6] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268. ACM, 2015.

[7] Florian Lemmerich Claudia Wagner Markus Strohmaier Georg Ahnert, Ivan Smirnov. The fairceptron: A framework for measuring human perceptions of algorithmic fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 1–9. ACM, 2021.

[8] Zhongzhou Liu, Yuan Fang, and Min Wu. Mitigating popularity bias for users and items with fairness-centric adaptive recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 2023.

[9] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the ACM Conference on Web Search and Data Mining (WSDM)*, pages 691–699, 2018.

[10] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. Towards involving end-users in interactive human-in-the-loop ai fairness. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 12(3):Article 18, 2022.

[11] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Pro-*

ceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2019.

[12] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* '19)*, pages 59–68. ACM, 2019.

[13] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. ACM, 2018.

[14] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1441–1450. ACM, 2019.

[15] Evdoxia Taka, Yuri Nakao, Ryosuke Sonoda, Takuya Yokota, Lin Luo, and Simone Stumpf. Human-in-the-loop fairness: Integrating stakeholder feedback to incorporate fairness perspectives in responsible ai. *arXiv preprint arXiv:2312.08064*, 2023.

[16] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 2021.

[17] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation. *ACM Transactions on Information Systems (TOIS)*, 41(1):1–29, 2023.

[18] Chen Xu, Sirui Chen, Jun Xu, Weiran Shen, Xiao Zhang, Gang Wang, and Zhenghua Dong. P-mmf: Provider max-min fairness re-ranking in recommender system. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 2023.

[19] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)*. ACM, 2017.

[20] Ziwei Zhu, Yun He, Xing Zhao, Yin Zhang, Jianling Wang, and James Caverlee. Popularity-opportunity bias in collaborative filtering. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2021.