Fairness and Bias in Recommender Systems

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

Introduction

Recommender systems shape what users see and interact with online, making their fairness a critical concern. While **collaborative filtering** (**CF**) methods are known to amplify popularity bias and create exposure inequalities, the fairness of **content-based recommenders** (**CBR**) is less understood.

This work addresses three key questions:

- **RQ1**: Do **content-based recommenders** exhibit lower unfairness than CF methods?
- **RQ2**: What are the **accuracy–fairness trade-offs** between CBR and CF?
- **RQ3**: How do **content feature choices** and **weightings** affect these trade-offs?

Methodology

Models Compared:

- **Random recommender** (baseline for fairness)
- Collaborative Filtering: **BPR**, **NeuMF**, **BERT4Rec**
- Content-Based: **MultiFuseCB** (proposed)

MultiFuseCB Details:

- Uses only **item-side features** (text, metadata)
- Feature extraction and selection using pretrained SentenceTransformer models (aka **SBERT**)
- **Feature embeddings** computed and evaluated individually, then fused with learnable weights
- **User embeddings** are an aggregation of interacted item embeddings

Datasets:

- **MovieLens 1M** (with enriched metadata via OMDb API)
- **Amazon Beauty Reviews**

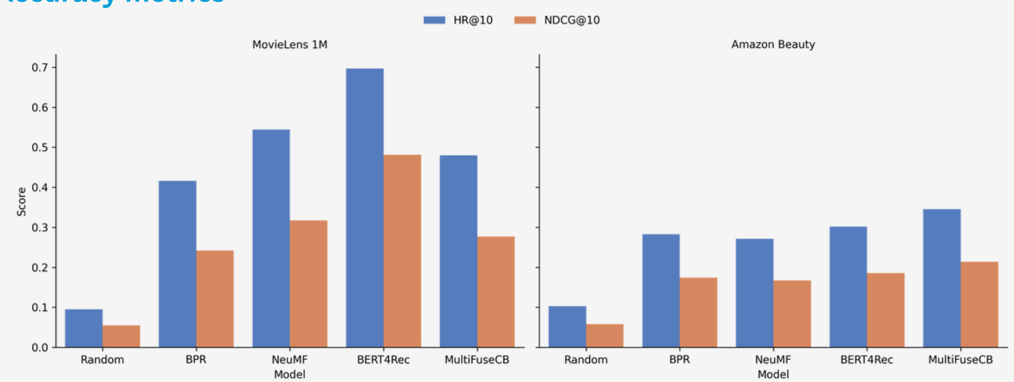
Evaluation Metrics:

- **Accuracy**: Hit Rate (**HR**), **NDCG**
- **Item fairness**: item coverage (**IC**), entropy (**Ent.**), Gini coefficient (**Gini**), average popularity (**Pop**), tail item exposure (**Tail**), head item exposure (**Head**),
- **User fairness**: standard deviation of gender group hit rates (**STD**) .

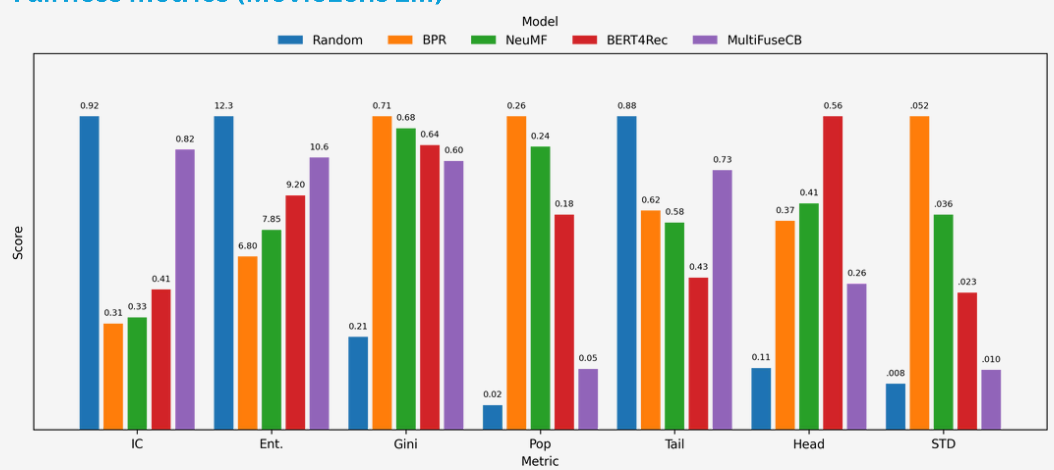
Discussion

- **Standard fairness metrics** (like exposure disparity and popularity bias) are useful for comparing models but are **limited**: they often miss **subjective**, **context-dependent** aspects of **fairness** and may not reflect how users actually **perceive** recommendation outcomes.
- Fairness is **sociotechnical**: User perceptions of fairness depend on **personal**, **cultural**, and **situational** factors. Even if metrics show improvement, some users or groups may still feel **unfairly treated**, especially if the system is opaque or inflexible.
- Beyond metrics: Future work should consider **user-centered dimensions** such as **perceived representation**, **agency**, and **explanation quality**. These are harder to quantify but essential for understanding real-world fairness, and can be explored through **user studies** and **qualitative feedback**.

Accuracy metrics



Fairness metrics (MovieLens 1M)



Results

RQ1: Are Content-Based Recommenders Fairer than CF?

- Our proposed **CBR** consistently achieved higher item coverage, lower popularity bias, and more equitable item exposure than **CF** baselines on both **MovieLens 1M** and **Amazon Beauty**.
- CF models (e.g., BERT4Rec, NeuMF) amplified popularity bias and concentrated exposure on a small set of popular items, while **CBR** distributed recommendations more broadly.

RQ2: What Are the Accuracy–Fairness Trade-offs?

- **CF** models achieved the **highest accuracy** (Hit Rate and NDCG), but at the cost of **significant fairness issues** (e.g., low coverage, high bias).
- **MultiFuseCB** delivered substantially improved fairness with only a modest reduction in accuracy compared to the best CF models.

RQ3: How Feature Choices and Weights Affect Trade-offs?

- **Feature selection** and **weighting** in **CBR** had a major impact on both **fairness** and **accuracy**.
- Incorporating diverse features (e.g., year, genre, plot) and optimizing their weights led to more balanced recommendations and **improved fairness metrics**.
- Embedding model choice also influenced results, with some text encoders yielding better trade-offs.

Conclusion & Future Work

- **Content-based recommenders**, when designed with careful feature selection and weighting, can achieve competitive accuracy while substantially **improving fairness** compared to **collaborative filtering models**—most notably by **reducing popularity bias** and **increasing item exposure**.
- Feature engineering and the use of advanced **embedding** models are essential for promoting **equitable recommendations** and **mitigating systemic biases**.
- **Future work** should expand evaluation to more diverse datasets, investigate richer and more varied item features, and incorporate user-centered fairness assessments to better understand real-world impacts and perceptions

Author: Filip Angheluta **Supervisor: Masoud Mansoury** **Email: f.angheluta-1@student.tudelft.nl**