A Comparative Fairness Study in News Recommendation Systems

Marta Salcedo and Alejandro Bellogín$^{[0000-0001-6368-2510]}$

Universidad Autónoma de Madrid, Madrid, Spain
marta.salcedos@estudiante.uam.es, alejandro.bellogin@uam.es

Abstract. Recommender systems have transformed how users access information, particularly in the news domain. While these systems enhance personalization, they also introduce fairness concerns, potentially limiting users' exposure to diverse perspectives. This experimental study aims to evaluate fairness across various recommendation algorithms by comparing collaborative and neural network-based approaches. To assess performance, both classical evaluation metrics, together with bias and fairness-aware metrics, such as Bias Disparity, Ranking-based Equal Opportunity, and Average Recommendation Popularity, are considered. The Adressa dataset is used to analyze recommendation behavior and their impact when using various user and item attributes. The results highlight important variations in fairness distribution across different algorithms, underscoring the necessity of incorporating fairness considerations into recommender system design.

Keywords: News recommendation · Fairness · Algorithmic bias.

1 Introduction

News articles are continuously published, while less-recent ones rapidly become not relevant or interesting for users. With a vast volume of news emerging daily, users often struggle to efficiently find content that aligns with their interests. At the same time, Recommender Systems (RS), while designed to enhance user experience by personalizing content, often inadvertently propagate media bias [21]. This occurs as these systems tend to favor content that aligns with users' existing beliefs, leading to an echo chamber effect that restricts exposure to diverse viewpoints [6]. Empirical studies indicate that advanced news RS are more likely to suggest biased news articles to users who exhibit a preference for such content, thereby amplifying the issue of media bias in digital news consumption [7].

Thus, developing fair RS in general, but especially in the news domain, is a critical step towards improving equal opportunities among users, items, and other stakeholders. However, there are several definitions of what fairness means [9], and many proposals of fair RSs use various fairness metrics focused on specific criteria [29]. Because of these issues, there is a lack of a standardized evaluation framework and benchmarking results across metrics [16]. Moreover, while it is important to improve (or guarantee some minimum) fairness, it might be possible that the overall recommendation performance may be negatively affected when introducing mitigation strategies with that goal. Hence, a balance between fairness and recommendation accuracy is recognized as necessary for these systems to be practical and useful in the real world [17,5].

In this context, we study the performance of various state-of-the-art collaborative filtering and neural network recommendation algorithms in the news domain, where these challenges are critical. Indeed, as stated in a recent paper [26], news RS should inform users about the existence of opinions they are not familiar with (diversity), without favoring some of them (fairness); however, no previous study has considered several fairness and bias metrics while comparing their results. Even though our main focus is on bias and fairness, we also measure classical accuracy, to understand the overall impact on all dimensions. For this, we use a real-world benchmark dataset, Adressa, where various user and item attributes can be extracted and analyzed. The results demonstrate the well-known tradeoff between fairness and accuracy, but they also evidence a less obvious outcome: that it might be more difficult to guarantee fairness for some attributes, whereas for others (probably because they are linked to other signals, such as popularity), classical and simple techniques already provide *fair* results.

2 Fairness and bias evaluation

When discussing fairness within RS, it is important to recognize the distinction between fairness and bias, two concepts that usually converge as synonyms or closely related [9]. **Bias** is usually associated (especially when referring to *item bias*) to the propensity of RS to disproportionately favor popular items, thereby creating a feedback loop that limits exposure to a diverse range of content [19]. Conversely, **fairness** concerns both users and items, and addresses disparities of recommendations received by different user groups while impacting on items of different characteristics [10]. Here we consider the following metrics to measure bias and fairness. First, regarding bias and, hence, mostly focused on popularity:

Average Coverage of Long Tail Items (ACLT) measures the extent to which items in the long tail (less popular items) are recommended. Higher ACLT indicates better representation of niche or less popular items [2].

Average Percentage of Long Tail Items (APLT) This metric reflects the recommender system's ability to diversify its suggestions beyond mainstream, popular items [1].

Average Recommendation Popularity (ARP) Captures the average popularity of recommended items, with lower ARP indicating a preference for recommending less popular items [31].

Popularity-based Ranking-based Equal Opportunity (PopREO) Focuses on equalizing the exposure of popular and less popular items in the top ranks, ensuring balanced representation regardless of item popularity [33]. Lower values are better.

Popularity-based Ranking-based Statistical Parity (PopRSP) Ensures that the proportion of popular and less popular items in recommendation lists is statistically similar across different user groups, promoting parity and mitigating popularity bias [33].

Expanding on the notions of item and user fairness, it becomes evident that fairness in RS extends beyond the interactions between users and items to encompass a more comprehensive ecosystem. This ecosystem includes both consumers—the users who interact with recommendations—and providers—the creators or organizations whose content is being recommended [22]. Understanding fairness through these dual lenses of consumer fairness and provider fairness provides a more holistic approach to addressing biases and disparities in recommendation outcomes. Consumer fairness ensures equitable user experiences by safeguarding against biases that disproportionately affect specific user groups, while provider fairness focuses on creating balanced opportunities for content creators to reach diverse audiences. Together, these dimensions underscore the interconnectedness of fairness concerns, where neglecting one can amplify inequities in the other, ultimately shaping the broader fairness landscape of recommender systems [10].

With growing concerns regarding fairness in recommendations, specific metrics have been developed to evaluate how equitably algorithms serve different user or item groups. The following will be used for the evaluation of the methods tested in this work:

Bias Disparity (BD) measures the difference in recommendation performance between user or item groups, highlighting systemic biases in the model [27]. A value close to 0 indicates the recommendation model mirrors the bias found in the data, meaning it neither amplifies nor mitigates existing biases, whereas a positive (negative) value suggests the model introduces a higher (lower) bias than what is present in the original data.

Item Mean Absolute Deviation Ranking-based (MADR) This metric evaluates how evenly ranking performance is distributed in recommendation lists among the different (item) groups, helping to reduce item unfairness [8].

Ranking-based Equal Opportunity (REO) ensures that items from different groups have an equal probability of appearing at the top ranks of recommendation lists [33].

3 Experiments and results

In this section, we address our main research question: **how do state-of-the-art RS based on collaborative filtering and neural networks perform in terms of fairness and accuracy in the news domain?** For this, in Section 3.1 we present the data used in our study and how the user and item groups were identified; later, Sections 3.2, 3.3, and 3.4 show the results obtained when considering accuracy, bias, and fairness metrics. In Sections 3.5 and 3.6 we discuss the main outcomes and limitations derived from our study.

3.1 Experimental settings

The dataset used in this study is the Adressa Dataset (Light Version 1), a publicly available dataset designed for research in news RS[1]. Provided by Adresseavisen, a Norwegian newspaper based in Trondheim, the dataset was collected using Cxense APIs and has been widely used in news recommendation research to analyze user behavior and content personalization strategies [12]. The dataset consists of user interactions with news articles recorded over one week (January 1–7, 2017), while providing insights into user demographics, device usage, and content distribution. Specifically, the dataset comprises 640,502 users and 20,428 items, with a total of 3,101,991 recorded transactions. The high sparsity indicates that user interactions with items are relatively infrequent, highlighting the challenges in generating personalized recommendations effectively.

To analyze user and item fairness, we identified and clustered in groups user and item attributes from the dataset. Item attributes selected included `access` type (distinguishing subscriber-only from open-access content), `author` name (which was further processed to infer gender through the use of Genderize API[2]), `category` (defining the topic of the article), `classification` (originally containing multiple values related to automatically assigned topics but later concatenated for consistency), and `sentiment` (derived from article metadata). User attributes selected included `country` (obtained from metatada and estimated from IP address), `device` type (such as mobile, desktop, or tablet), and operating system `OS` (e.g., Windows, iOS, Android).

The following recommendation algorithms were considered in the experiments, all of them available in the Elliot library [3]: Random produces random suggestions and serves as a non-personalized baseline; MostPop is also non-personalized but recommends according to how popular items are; ItemKNN is a nearest-neighbor method based on items [20]; BPRMF is a Matrix Factorization (MF) method that uses Bayesian Personalized Ranking optimization to handle pairwise preferences [18]; EASE$^{\mathrm{R}}$ is a linear autoencoder-based algorithm [25] popular to perform well across datasets [4]; NeuMF is a neural approach that extends traditional MF [14]. Finally, experiments were executed at different ranking cutoffs, but for the sake of space, only results at cutoff 10 are reported.

3.2 Accuracy evaluation

Table 1 presents the evaluation results in terms of accuracy metrics for different recommendation models at cutoff 10. The models were assessed using multiple evaluation metrics, including Normalized Discounted Cumulative Gain (nDCG), Recall, Hit Rate (HR), Precision, Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR) [13]. These results provide insights into the overall performance of each model in terms of ranking accuracy effectiveness.

The Random model, as expected, performs the worst across all metrics, reinforcing the importance of personalized over non-personalized approaches. The

[1] `https://reclab.idi.ntnu.no/dataset/`

[2] `https://genderize.io/documentation`

Table 1: Accuracy evaluation, higher values, better. Subindices denote relative rank of the algorithm with respect to that column (metric).

Model	**nDCG**	**Recall**	**HR**	**Precision**	**MAP**	**MRR**
Random	0.0011_6	0.0017_6	0.0054_5	0.0005_5	0.0006_5	0.0017_5
MostPop	0.0065_4	0.0111_4	0.0216_4	0.0022_4	0.0027_4	0.0075_4
ItemKNN	0.0396_2	0.0701_2	0.0943_2	0.0099_2	0.0125_2	0.0360_2
BPRMF	0.0085_3	0.0139_3	0.0278_3	0.0029_3	0.0034_3	0.0099_3
EASE$^{\mathrm{R}}$	$\mathbf{0.0478}_1$	$\mathbf{0.0888}_1$	$\mathbf{0.1234}_1$	$\mathbf{0.0129}_1$	$\mathbf{0.0147}_1$	$\mathbf{0.0419}_1$
NeuMF	0.0014_5	0.0028_5	0.0052_6	0.0005_5	0.0005_6	0.0015_6

MostPop model, despite being a simple popularity-based approach, maintains moderate performance and consistently outperforms the Random baseline. However, it remains weaker compared to ItemKNN and BPRMF in capturing personalized recommendations, although it outperforms NeuMF.

ItemKNN demonstrates significantly better performance in Recall and Hit Rate. EASE$^{\mathrm{R}}$ emerges as one of the strongest models, outperforming both ItemKNN and BPRMF in terms of nDCG and Recall. It exhibits a particularly strong ability to retrieve relevant items efficiently, making it a competitive alternative to traditional collaborative filtering models. Interestingly, NeuMF, despite being a deep learning-based approach, performs worse than traditional collaborative filtering techniques in Recall and Hit Rate. This may be due to dataset sparsity or suboptimal hyperparameter tuning, which could affect its ability to generalize effectively.

3.3 Bias evaluation

Table 2 presents the bias evaluation measurements according to the metrics presented in Section 2. These metrics provide insight into how recommendation algorithms handle popularity bias and content exposure when suggesting news articles. First, the Random model exhibits the lowest level of popularity bias, based on its high ACLT and APLT values and relatively low ARP values, evidencing it does not prioritize recommending popular content. Additionally, its PopREO and PopRSP values are significantly lower than those of personalized models. While this model minimizes bias, it lacks personalization, making it less suitable for real-world recommendation scenarios. In the other extreme, the MostPop model, as expected, strongly favors popular content, reflected in its consistently high ARP values, and maximum PopREO and PopRSP values. The BPRMF model exhibits a strong popularity bias – an observation already known in the area for other domains [32,11,23] –, as evidenced by all the values that mimic those from the MostPop algorithm. The ItemKNN model, on the other hand, demonstrates lower popularity bias compared to MostPop and BPRMF, as indicated by its higher ACLT and APLT values. This suggests that ItemKNN distributes recommendations across a broader range of items, mitigating bias to

Table 2: Bias evaluation, better (less biased) values indicated by ↑ (higher) and ↓ (lower) arrows, depending on the metric.

Model	**ACLT ↑**	**APLT ↑**	**ARP ↓**	**PopREO ↓**	**PopRSP ↓**
Random	$\mathbf{8.689}_1$	$\mathbf{0.955}_1$	$\mathbf{84.056}_1$	$\mathbf{0.020}_1$	$\mathbf{0.004}_1$
MostPop	0.000_5	0.000_5	4604.950_5	1.000_6	1.000_6
ItemKNN	1.396_3	0.131_3	2722.296_3	0.739_3	0.985_3
BPRMF	0.000_5	0.000_5	4671.479_6	1.000_6	1.000_6
EASE$^{\mathbf{R}}$	0.498_4	0.050_4	3351.067_4	0.860_4	0.995_4
NeuMF	7.274_2	0.727_2	423.002_2	0.181_2	0.784_2

some extent. Similarly, the EASE$^{\text{R}}$ model introduces a slight reduction in popularity bias compared to BPRMF, as reflected in its nonzero ACLT and APLT values, which indicate some level of long-tail recommendation. However, its ARP value remains relatively high, suggesting that it continues to prioritize popular items, something that should be taken into account when considering its high accuracy performance (see Table 1). The NeuMF model shows moderate bias compared to the other models; while it achieves slightly better long-tail coverage than MostPop and BPRMF, its PopREO and PopRSP values suggest a strong preference for popularity, though to a lesser extent than the others.

These results highlight how different recommendation models handle popularity bias in news recommendation. While some models reduce popularity bias by promoting diverse content, they often do so at the cost of engagement and relevance. Among the evaluated approaches, ItemKNN appears to offer the most balanced tradeoff, effectively reducing bias while maintaining competitive recommendation performance.

3.4 Fairness evaluation

The Bias Disparity evaluates disparities in exposure and representation by analyzing combinations of user clusters (`device type`, `OS`, `country`) and item clusters (`access` type, `author` gender, `category`, `classification`, `sentiment`). Due to the large number of possible user-item attribute combinations, individual values for each pair cannot be analyzed in detail. Instead, an averaged Bias Disparity value has been computed for each combination to provide a clearer understanding of the general trends present in the data. As an example, Table 3 presents specific Bias Disparity values for different `device` and `sentiment` clusters[3], whereas Table 4 presents the aggregated Bias Disparity values for all user and item attributes. It can be noted how the aggregated values do not capture all the nuances that appear in Table 3; for example, for MostPop, even though

[3] Note that for this example, 6 columns are obtained, due to the combination of a user cluster with 3 values (`device`: mobile, desktop, or tablet) and an item cluster with 2 values (`sentiment`: no sentiment or negative). For other clusters or combinations, this way of presenting the results is unfeasible.

Table 3: BD for `device` (D) and `sentiment` (S) clusters.

Model	**D0-S0**	**D0-S1**	**D1-S0**	**D1-S1**	**D2-S0**	**D2-S1**
Random	0.147	-0.426	0.085	-0.312	0.133	-0.405
MostPop	-0.387	1.122	-0.443	1.634	-0.396	1.207
ItemKNN	-0.128	0.372	-0.088	0.325	-0.119	0.361
BPRMF	-0.255	0.740	-0.315	1.160	-0.266	0.811
EASE$^\mathbf{R}$	-0.052	0.151	-0.037	0.137	-0.052	0.157
NeuMF	0.151	-0.437	0.087	-0.322	0.135	-0.411

Table 4: Aggregated BD for user (device, OS, country) and item (access type, author gender, category, classification, sentiment) clusters.

Model	**Device**					**OS**					**Country**				
	Access	Author	Category	Class.	Sent.	Access	Author	Category	Class.	Sent.	Access	Author	Category	Class.	Sent.
Random	5.129	0.672	15.577	4.755	-0.130	2.522	0.532	4.697	1.741	-0.068	0.627	0.259	0.556	0.367	0.085
MostPop	-0.621	-0.251	-0.439	-0.240	0.456	-0.419	-0.059	-0.166	-0.032	0.407	-0.104	0.143	0.032	0.193	0.690
ItemKNN	-0.097	-0.096	0.410	0.185	0.121	-0.120	0.001	0.121	0.156	0.090	0.071	0.123	0.088	0.308	0.160
BPRMF	-0.374	-0.205	-0.414	-0.261	0.313	-0.218	-0.023	-0.141	-0.054	0.289	0.078	0.178	0.064	0.172	0.539
EASE$^\mathbf{R}$	-0.375	-0.162	-0.249	-0.159	0.051	-0.247	-0.048	-0.088	0.040	0.037	-0.006	0.118	0.055	0.218	0.123
NeuMF	5.071	0.211	0.132	0.081	-0.133	2.460	0.159	0.428	0.132	-0.072	0.648	0.183	0.747	0.094	0.033

3 of the values (those related to S0) are negative, the aggregated value in Table 4 (0.456) is positive.

These results offer a broader perspective on how each model interacts with different demographic and content groups. By averaging the results, it is possible to compare the overall fairness of recommendation models, highlighting which models distribute recommendations more equitably across diverse user and item characteristics, providing insights into how recommendation models favor certain user groups or content types, indicating disparities in exposure. In particular, the Random model consistently exhibits the highest absolute BD values across most attributes (all except `sentiment`). Since its values are usually positive, this shows that it tends to increase the bias (for that pair of attributes) with respect to the original data. The rest of the models, except for NeuMF, seem to simply replicate the biases that already exist in the training data (with values very close to zero). NeuMF, on the other hand, produces very high values for the `access` attribute, comparable to the Random model; however, for the rest of the attributes, even though it increases the bias with respect to the data, its personalization component, produces more limited disparities than Random.

Table 5 presents the MADR evaluation metric, assessing how different recommendation models distribute ranking performance across various item attributes, hence, highlighting disparities in ranking fairness. One key observation is that the Random model exhibits the lowest values across all attributes excepts for `category`, where the MostPop model obtains slightly better performance. This indicates that the performance of the Random recommender is the same independently of the item attribute, however, since its performance is, by definition, very low, this good result in fairness does not translate in the model being practical for real-world recommendation tasks. At the other side of the spectrum we

Table 5: MADR values for item attributes.

Model	**Access**	**Author**	**Category**	**Classification**	**Sentiment**
Random	$\mathbf{0.0009}_1$	$\mathbf{0.0027}_1$	0.0011_2	$\mathbf{0.0012}_1$	$\mathbf{0.0009}_1$
MostPop	0.0026_2	0.0033_2	$\mathbf{0.0006}_1$	0.0016_2	0.0019_2
ItemKNN	0.0270_5	0.0162_4	0.0212_6	0.0135_6	0.0024_3
BPRMF	0.0293_6	0.0271_6	0.0027_4	0.0055_4	0.0543_6
EASE$^\mathbf{R}$	0.0189_4	0.0263_5	0.0167_5	0.0117_5	0.0030_4
NeuMF	0.0036_3	0.0038_3	0.0011_2	0.0025_3	0.0054_5

Table 6: REO values for item attributes.

Model	**Access**	**Author**	**Category**	**Classification**	**Sentiment**
Random	$\mathbf{0.2726}_1$	$\mathbf{0.1384}_1$	$\mathbf{2.9657}_1$	$\mathbf{2.1699}_1$	0.0193_2
MostPop	1.4142_6	0.9365_5	11.2957_6	5.7810_6	0.5178_5
ItemKNN	0.4775_2	0.4050_2	3.4245_2	2.6967_2	0.1675_3
BPRMF	1.3251_5	0.8653_4	9.4444_4	4.2546_5	0.3425_4
EASE$^\mathbf{R}$	1.0804_3	0.5441_3	4.1486_3	2.9322_3	$\mathbf{0.0030}_1$
NeuMF	1.3001_4	0.9882_6	9.5606_5	4.1440_4	0.7668_6

find BPRMF and ItemKNN, which seem to perform quite differently depending on the item attribute (in particular, `access`, `author` gender, and `sentiment` for BPRMF and `access`, `category`, and `classification` for ItemKNN). NeuMF and EASE$^\mathrm{R}$, as a consequence, obtain medium values of this metric for most of the attributes; considering the high performance achieved by EASE$^\mathrm{R}$ in terms of accuracy (see Table 1), this model could be a good option to address the accuracy-fairness tradeoff.

Finally, Table 6 presents the REO performance values, assessing whether items from different groups have an equal probability of appearing at the top ranks of recommendation lists, promoting fairness according to item visibility across various item attributes. Thus, this metric helps determine whether certain content types receive disproportionate exposure. Again, we observe the Random model exhibits the lowest values for this metric except, in this case, for the `sentiment` attribute. The worst-performing models according to this metric correspond to MostPop (`access`, `category`, `classification`, `author` gender, and `sentiment`) and NeuMF (`author` gender, `sentiment`, and `category`).

3.5 Discussion

In this section we analyze how various personalized and non-personalized algorithms perform with respect to different accuracy, bias, and fairness metrics. The first obvious observation is that, as expected, there is not a clear winner with respect to the *accuracy-fairness* (or *accuracy-bias*) tradeoff. However, what also stems from these results is the **bias-fairness** tradeoff. Since the selected bias metrics are mainly related to popularity measurements, any fairness metric not aligned with this attribute will result in incompatible conclusions. For example, MADR ranks MostPop as the first or second best algorithm for any item attribute, where this method is, by definition, the most biased one when considering popularity-biased metrics. A similar observation can be made between REO and bias metrics.

However, the fact that fairness metrics may behave very differently depending on the user or item attribute being considered, makes this analysis richer and with far more implications for this dimension. According to the relative ranking of the recommendation algorithms for MADR, we may group the item attributes in 3 groups: `access` and `author` gender (although `category` behaves in a very similar way as these two attributes), `classification` is more similar to `category` than to `access` or `author` gender, and `sentiment`. At the same time, the performance of the attributes according to the REO metric, although being slightly different (for example, MostPop is the worst performing one in REO for `access` attribute, but the second-best in MADR), in terms of the derived groups according to the relative performance of the algorithms remains the same: `category` and `classification` are very similar to each other, `sentiment` is the least similar to the rest, and `access` and `author` gender correlate to some extent to each other and with `category`. This highlights that, even though the recommendation algorithms were not trained with any information about these attributes whatsoever, the inherent signals available in the data allow to produce recommendations that can be interpreted as more or less fair with respect to these attributes, depending on the considered fairness definition.

A little bit more complex to analyze is the bias disparity (BD) metric, since it takes into account the original bias in the data and, as a consequence, it is not bounded and the sign is important. However, this is the only metric, among the ones considered in this study, that can integrate both user and item attributes at the same time. In this case, the `sentiment` attribute again stands out with a distinctive pattern. Also `country`, as user attribute, performs differently than `device` and `OS`, obtaining almost in every case positive BD values, meaning the algorithms are increasing the bias with respect to the original data.

3.6 Limitations

While this study provides valuable insights into fairness in news recommender systems, several limitations must be acknowledged. First, the choice of the Adressa dataset, a local Norwegian news outlet, constrains the generalizability of the findings. As a regional newspaper, Adressa does not reflect the diversity or editorial plurality of a global news aggregator, where balancing across multiple news sources and ideological perspectives is more critical. Datasets like MIND[4], which include a broader range of publishers and user demographics, may be more appropriate for evaluating fairness in such contexts. Additionally,

[4] `https://msnews.github.io/`

the use of ‘country’ as a fairness-relevant user attribute is limited in this setting, as most users are likely Norwegian, and international interactions may still represent Norwegian expatriates, thus reducing the attribute’s discriminatory power.

Moreover, the evaluation relies on standard recommendation metrics that may not fully capture the normative and democratic roles of news recommenders. Metrics such as nDCG and Recall, while useful for general performance assessment, overlook domain-specific concerns like viewpoint diversity, civic engagement, and stakeholder equity. As highlighted in recent work [28,15,24], fairness in news recommendation requires more context-aware evaluation approaches. Future work should better align methodological choices with the unique ethical and societal demands of the news domain.

Additionally, this work is focused on classical collaborative filtering approaches or neural network-based recommendation algorithms. Other algorithmic families, such as those based on content or knowledge, are not considered and the effects on fairness and bias metrics by those methods remain unexplored.

4 Conclusions

This study conducted a comparative assessment of fairness in news RS by evaluating multiple algorithms using classical performance metrics, bias measurements, and fairness indicators. The findings reveal that recommendation algorithms inherently introduce bias, with certain models disproportionately favoring specific item categories, user demographics, or sentiment-driven content. Traditional algorithms like MostPop and BPRMF exhibited a strong tendency to prioritize popular articles, reinforcing the visibility of widely consumed content. EASE$^{\mathrm{R}}$ demonstrated a more balanced item distribution, though biases remained, influenced by both user attributes and content characteristics. Fairness-specific metrics further revealed systematic biases in content ranking. For example, REO and BD metrics indicated that certain models disproportionately favored sentiment-driven news, whereas long-tail exposure metrics (ACLT and APLT) confirmed that less popular items were consistently underrepresented, underscoring the ongoing challenge of achieving equitable content distribution. Despite incorporating fairness-aware evaluation, no model successfully eliminated bias. Instead, this study highlights the *accuracy-fairness tradeoff*, where efforts to mitigate bias often come at the cost of reduced recommendation performance. This finding suggests that multi-objective optimization approaches to balance accuracy, diversity, and equity are essential for fairness-aware recommender systems [30], and more work, specifically tailored to the news domain, is missing to measure and mitigate these aspects.

Acknowledgments. This work was supported by Grant PID2022-139131NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by “ERDF, a way of making Europe.”

References

1. Abdollahpouri, H., Burke, R., Mobasher, B.: Controlling popularity bias in learning-to-rank recommendation. In: Proceedings of RecSys, Como, Italy, August 27-31, 2017. pp. 42–46. ACM (2017)
2. Abdollahpouri, H., Burke, R., Mobasher, B.: Managing popularity bias in recommender systems with personalized re-ranking. In: Proceedings of FLAIRS, Sarasota, Florida, USA, May 19-22 2019. pp. 413–418. AAAI Press (2019)
3. Anelli, V.W., Bellogín, A., Ferrara, A., Malitesta, D., Merra, F.A., Pomo, C., Donini, F.M., Noia, T.D.: Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. In: Proceedings of SIGIR, Virtual Event, Canada, July 11-15, 2021. pp. 2405–2414. ACM (2021)
4. Anelli, V.W., Bellogín, A., Noia, T.D., Jannach, D., Pomo, C.: Top-n recommendation algorithms: A quest for the state-of-the-art. In: Proceedings of UMAP, Barcelona, Spain, July 4 - 7, 2022. pp. 121–131. ACM (2022)
5. Buijsman, S.: Navigating fairness measures and trade-offs. AI Ethics **4**(4), 1323–1334 (2024)
6. Cinelli, M., Morales, G.D.F., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber effect on social media. Proc. Natl. Acad. Sci. USA **118**(9), e2023301118 (2021), last accessed: Nov 2024
7. Cinus, F., Minici, M., Monti, C., Bonchi, F.: The effect of people recommenders on echo chambers and polarization. In: Proceedings of ICWSM, Atlanta, Georgia, USA, June 6-9, 2022. pp. 90–101. AAAI Press (2022)
8. Deldjoo, Y., Anelli, V.W., Zamani, H., Bellogín, A., Noia, T.D.: A flexible framework for evaluating user and item fairness in recommender systems. User Model. User Adapt. Interact. **31**(3), 457–511 (2021)
9. Deldjoo, Y., Jannach, D., Bellogín, A., Difonzo, A., Zanzonelli, D.: Fairness in recommender systems: research landscape and future directions. User Model. User Adapt. Interact. **34**(1), 59–108 (2024)
10. Ekstrand, M.D., Das, A., Burke, R., Diaz, F.: Fairness in recommender systems. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 679–707. Springer US (2022)
11. Guíñez, F., Ruiz, J., Sánchez, M.I.: Quantification of the impact of popularity bias in multi-stakeholder and time-aware environments. In: Proceedings of BIAS, Lucca, Italy, April 1, 2021. Communications in Computer and Information Science, vol. 1418, pp. 78–91. Springer (2021)
12. Gulla, J.A., Zhang, L., Liu, P., Özgöbek, Ö., Su, X.: The adressa dataset for news recommendation. In: Proceedings of ICWI, Leipzig, Germany, August 23-26, 2017. pp. 1042–1048. ACM (2017)
13. Gunawardana, A., Shani, G., Yogev, S.: Evaluating recommender systems. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 547–601. Springer US (2022)
14. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.: Neural collaborative filtering. In: Proceedings of WWW, Perth, Australia, April 3-7, 2017. pp. 173–182. ACM (2017)
15. Helberger, N.: On the democratic role of news recommenders. Digital Journalism **7**(8), 993–1012 (2019)
16. Jin, D., Wang, L., Zhang, H., Zheng, Y., Ding, W., Xia, F., Pan, S.: A survey on fairness-aware recommender systems. Inf. Fusion **100**, 101906 (2023)

17. Karimi, S., Rahmani, H.A., Naghiaei, M., Safari, L.: Provider fairness and beyond-accuracy trade-offs in recommender systems. CoRR **abs/2309.04250** (2023)
18. Koren, Y., Rendle, S., Bell, R.M.: Advances in collaborative filtering. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 91–142. Springer US (2022)
19. Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., Burke, R.: Feedback loop and bias amplification in recommender systems. In: Proceedings of CIKM, Virtual Event, Ireland, October 19-23, 2020. pp. 2145–2148. ACM (2020)
20. Nikolakopoulos, A.N., Ning, X., Desrosiers, C., Karypis, G.: Trust your neighbors: A comprehensive survey of neighborhood-based methods for recommender systems. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 39–89. Springer US (2022)
21. Ruan, Q., Xu, J., Leavy, S., Namee, B.M., Dong, R.: Rewriting bias: Mitigating media bias in news recommender systems through automated rewriting. In: Proceedings of UMAP, Cagliari, Italy, July 1-4, 2024. pp. 67–77. ACM (2024)
22. Sacharidis, D., Mouratidis, K., Kleftogiannis, D.: A common approach for consumer and provider fairness in recommendations. In: Proceedings of ACM RecSys Late-Breaking Results, Copenhagen, Denmark, September 16-20, 2019. CEUR Workshop Proceedings, vol. 2431, pp. 1–5. CEUR-WS.org (2019)
23. Sánchez, P., Bellogín, A., Boratto, L.: Bias characterization, assessment, and mitigation in location-based recommender systems. Data Min. Knowl. Discov. **37**(5), 1885–1929 (2023)
24. Smets, A., Hendrickx, J., Ballon, P.: We're in this together: A multi-stakeholder approach for news recommenders. Digital Journalism **10**(10), 1813–1831 (2022)
25. Steck, H.: Embarrassingly shallow autoencoders for sparse data. In: Proceedings of WWW, San Francisco, CA, USA, May 13-17, 2019. pp. 3251–3257. ACM (2019)
26. Treuillier, C., Castagnos, S., Özgöbek, Ö., Brun, A.: Beyond trade-offs: Unveiling fairness-constrained diversity in news recommender systems. In: Proceedings of UMAP, Cagliari, Italy, July 1-4, 2024. pp. 143–148. ACM (2024)
27. Tsintzou, V., Pitoura, E., Tsaparas, P.: Bias disparity in recommendation systems. In: Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments, Copenhagen, Denmark, September 20, 2019. CEUR Workshop Proceedings, vol. 2440. CEUR-WS.org (2019)
28. Vrijenhoek, S., Bénédict, G., Granada, M.G., Odijk, D., de Rijke, M.: Radio - rank-aware divergence metrics to measure normative diversity in news recommendations. In: Proceedings of RecSys, Seattle, WA, USA, September 18 - 23, 2022. pp. 208–219. ACM (2022)
29. Wang, Y., Ma, W., Zhang, M., Liu, Y., Ma, S.: A survey on the fairness of recommender systems. ACM Trans. Inf. Syst. **41**(3), 52:1–52:43 (2023)
30. Wu, H., Ma, C., Mitra, B., Diaz, F., Liu, X.: A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation. ACM Trans. Inf. Syst. **41**(2), 47:1–47:29 (2023)
31. Yin, H., Cui, B., Li, J., Yao, J., Chen, C.: Challenging the long tail recommendation. Proc. VLDB Endow. **5**(9), 896–907 (2012)
32. Zhu, Z., He, Y., Zhao, X., Zhang, Y., Wang, J., Caverlee, J.: Popularity-opportunity bias in collaborative filtering. In: Proceedings of WSDM, Virtual Event, Israel, March 8-12, 2021. pp. 85–93. ACM (2021)
33. Zhu, Z., Wang, J., Caverlee, J.: Measuring and mitigating item under-recommendation bias in personalized ranking systems. In: Proceedings of SIGIR, Virtual Event, China, July 25-30, 2020. pp. 449–458. ACM (2020), last accessed: Nov 2024