Mind the Dialect: NLP Advancements Uncover Fairness Disparities for Arabic Users in Recommendation Systems

Abdulla Alshabanah **Murali Annavaram**
University of Southern California
{aalshaba, annavara}@usc.edu

Abstract

Recommendation systems play a critical role in shaping user experiences and access to digital content. However, these systems can exhibit unfair behavior when their performance varies across user groups, especially in linguistically diverse populations. Recent advances in NLP have enabled the identification of user dialects, allowing for more granular analysis of such disparities. In this work, we investigate fairness disparities in recommendation quality among Arabic-speaking users, a population whose dialectal diversity is underrepresented in recommendation system research. By uncovering performance gaps across dialectal variation, we highlight the intersection of NLP and recommendation system and underscore the broader social impact of NLP. Our findings emphasize the importance of interdisciplinary approaches in building fair recommendation systems, particularly for global and local platforms serving diverse Arabic-speaking communities. The source code is available at https://github.com/alshabae/FairArRecSys.

1 Introduction

Recommendation systems shape the way we interact with digital content and, by extension, influence how we perceive the world around us in the digital age. From the content provider's perspective, they have become integral to online platforms, enhancing user experience through personalization (Naumov et al., 2019; Covington et al., 2016) and driving significant economic activity for major technology companies (Linden et al., 2003; Greenstein-Messica and Rokach, 2018; Gomez-Uribe and Hunt, 2016). Given their societal impact, it is critical to uncover biases that could negatively affect user experience or undermine equitable access to content.

One unexplored source of bias in recommendation system research is related to Arabic-speaking users. Arabic, spoken by over 420 million people worldwide, comprises numerous dialects that differ significantly in grammar and vocabulary (Bergman and Diab, 2022). While these dialects share a common root in Modern Standard Arabic (MSA), they have diverged across regions, often resulting in limited mutual intelligibility. This divergence creates a form of diglossia, where MSA is used as the formal, high-resource language, while regional dialects are less represented in digital and formal contexts (Ferguson, 1959). Recent work and shared tasks in Arabic NLP has focused primarily on dialect identification (Abdul-Mageed et al., 2024, 2023; Keleg et al., 2024, 2023), with only one study exploring the use of dialect information to improve recommendation performance (Alshabanah and Annavaram, 2025). These advancements in Arabic NLP have made it possible to examine a critical fairness issue: The quality of recommendations for Arabic-speaking users varies with the dialect they use. Our analysis shows that users using certain dialects receive lower-quality recommendations, leading to access disparities across dialect groups.

In this paper, we take a first step toward uncovering this fairness issue in recommendation systems for Arabic-speaking users. Our findings challenge the implicit but unexamined assumption that all Arabic speakers are treated equally by recommendation systems. By highlighting disparities tied to Arabic dialects, this work draws attention to the broader impact of NLP and its connection to other fields.

2 Preliminaries

2.1 Arabic Dialect Identification

Identifying Arabic dialects in written text, known as Arabic Dialect Identification (ADI), is a challenging task due to the wide range of dialects and the linguistic overlap between them (Althobaiti, 2020). In many cases, a sentence may be correctly interpreted as belonging to more than one

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15895–15903
November 4-9, 2025

dialect, which makes accurate classification difficult (Abdul-Mageed et al., 2023). To address this challenge, researchers have proposed models that better capture dialectal overlap and improve prediction accuracy (Keleg and Magdy, 2023). These efforts have been supported by a series of shared tasks that promote the development and benchmarking of robust ADI models (Abdul-Mageed et al., 2021, 2022, 2023, 2024).

In addition to ADI, the Arabic Level of Dialectness (ALDi) (Keleg et al., 2023) provides a more fine-grained view of language variation. Instead of assigning a single dialect label, ALDi measures how dialectal a sentence is on a continuous scale. This approach reflects a common practice among Arabic speakers of mixing MSA with dialectal Arabic in everyday writing.

In this work, we focus on grouping users based on the dialectal characteristics of their written language. To do this, we apply both ADI and ALDi techniques to curate dialect labels for users. Specifically, we use a BERT-based model introduced in (Keleg and Magdy, 2023) to predict the most commonly used dialect for each user, based on their written reviews. The model classifies text into the following dialects, aligned with country-level groupings: Egypt, Sudan, Bahrain, Iraq, Kuwait, Oman, Qatar, Saudi Arabia, United Arab Emirates, Yemen, Jordan, Lebanon, Palestine, Syria, Libya, Morocco, Algeria and Tunisia. Following (Keleg et al., 2024), we take an additional step by labeling any text with an ALDi score below 0.1 as MSA. This allows us to separate high-resource, formal Arabic from dialectal Arabic more accurately.

In our experiments, we analyze fairness disparities at three levels: (1) **country-level** dialect groups, which reflect fine-grained distinctions; (2) broader **regional** groups, which cluster similar dialects from neighboring countries; and (3) the **MSA vs. dialectal Arabic** split, which distinguishes users who primarily use Modern Standard Arabic from those who write in dialectal Arabic. The regional groups are defined as follows: the **Nile** region includes Egypt and Sudan; the **Gulf** region consists of the United Arab Emirates, Saudi Arabia, Qatar, Yemen, Bahrain, Iraq, Kuwait, and Oman; the **Maghreb (MAG)** region comprises Libya, Morocco, Tunisia, and Algeria; the **Levant (LEV)** includes Jordan, Lebanon, Palestine, and Syria; and finally, **Modern Standard Arabic (MSA)** is treated as a distinct category. This multi-level analysis helps reveal where disparities emerge and how they scale across different levels of language variation.

2.2 Recommendation Systems

To model user-item interactions, various recommendation architectures have been proposed, with the Two-Tower Neural Network (TTNN) being one of the most widely adopted in large-scale systems (Balasubramanian et al., 2024; Yi et al., 2019). TTNNs use two separate deep neural networks: one processes user features and the other processes item features. Each network transforms a combination of sparse and dense inputs into a fixed-size dense representation, one for the user and one for the item. The relevance score between a user and an item is then computed using a scoring function, such as a dot product or an additional neural layer.

In addition to TTNN, we also experiment with another popular model, DeepFM (Guo et al., 2017), a model that combines the strengths of factorization machines (FM) (Rendle, 2010) and deep learning. DeepFM captures both low-order feature interactions (via its FM component) and high-order interactions (through a deep neural network). It operates directly on raw feature inputs without requiring manual feature engineering, making it particularly effective in capturing complex user-item relationships.

TTNN and DeepFM model parameters are optimized using a max-margin ranking loss, which encourages the model to assign higher scores to positive examples (items the user has interacted with) than to negative examples (items the user has not interacted with). Specifically, the objective is to ensure that the score of a positive item is higher than that of a negative item by at least a predefined margin Δ. Given a user u, a positive item i^+, and a negative item i^-, with corresponding item representations $\mathbf{y_{i^+}}$ and $\mathbf{y_{i^-}}$, and a user representation $\mathbf{x_u}$, the loss function is defined as:

$$\mathcal{L}(\mathbf{x_u}, \mathbf{y_{i^+}}) = \max(0,\ SF(\mathbf{x_u}, \mathbf{y_{i^-}}) - SF(\mathbf{x_u}, \mathbf{y_{i^+}})\ + \Delta) \quad (1)$$

where $SF(\cdot, \cdot)$ denotes the scoring function used to measure the relevance of items to users.

3 Evaluation

3.1 Evaluation Setting

Datasets: We conduct our experiments using two publicly available benchmark datasets: the *Book*

Reviews in Arabic Dataset (BRAD) (Elnagar et al., 2018; Elnagar and Einea, 2016) and the *Large-scale Arabic Book Reviews Dataset* (LABR) (Aly and Atiya, 2013). As our focus is on the top-k recommendation task, we apply a standard preprocessing step to convert both datasets into an implicit feedback format. Following previous work (He et al., 2017; Alshabanah et al., 2025), we adopt the leave-one-out strategy to split the data into training, validation, and test sets. Additional details on datasets are provided in Appendix A.

Metrics: We adopt Hit Rate@k (HR@k) and Normalized Discounted Cumulative Gain@k (NDCG@k) as evaluation metrics to measure the effectiveness of our method. HR@k evaluates whether a relevant item is included among the top-k recommendations, whereas NDCG@k considers the ranking position of the relevant item, assigning higher importance to items ranked closer to the top. By using both metrics, we obtain a comprehensive evaluation of the recommendation quality.

Implementation Details: The implementation details can be found in Appendix B.

3.2 Uncovering Fairness Disparities

We analyze fairness disparities in recommendation performance across two models (TTNN and DeepFM) and two datasets (LABR and BRAD), using HR@10 and NDCG@10 as evaluation metrics. We report results at three levels: MSA vs. dialectal Arabic (Figure 1), regional dialect groups (Figures 2 and 5), and country-level dialects (Figures 3, 4, 6 and 7). We also include additional results showing HR@10 and NDCG@10 for a Matrix Factorization model (Koren et al., 2009) at all three dialect grouping levels in Appendix C. Scores of zero for user groups with no interaction data are excluded to avoid distortion in the results.

At the highest level, comparing MSA and dialectal Arabic, we observe generally consistent patterns within each dataset, though the extent of the differences in recommendation quality varies across datasets. In BRAD, MSA users mostly perform better across both models. In LABR, the gap is more visible with DeepFM (HR@10 of 0.079 for dialectal vs. 0.058 for MSA). A possible explanation of this gap could lie in the average of test item degrees: in BRAD, MSA items (95.1) had slightly higher interaction counts than Dialectal items (93.2), whereas in LABR the opposite happened, with Dialectal items (40.0) showing higher interaction counts than MSA items (36.2). Generally, the more interactions an item has the more visibility it gets which eventually leads to better recommendation quality (Balasubramanian et al., 2024). Nevertheless, looking at this high-level view, we can see that there is a difference in how user groups are treated based on the variety of Arabic they use.

At the regional level (Nile, Gulf, Levant, Maghreb, and MSA), we also observe noticeable variation in recommendation quality. Specific user groups tend to receive better performance than others, but the ranking of regions varies by dataset and model. For example, in LABR, Levant users achieve the highest HR@10, while Gulf and Maghreb users fall behind. In contrast, in BRAD, Maghreb users receive higher-quality recommendations, with HR@10 scores of 0.065 and 0.077 for TTNN and DeepFM, respectively. These trends are also reflected in NDCG@10 scores, indicating that regional variation in dialect correlates with disparities in recommendation performance. However, it is also worth noting that the behavior varies across models, which highlights the need for more careful consideration when designing recommendation systems for diverse Arabic-speaking user groups.

At the country level, the disparities are even more pronounced. Certain countries have high HR@10 values, often above 0.1. In contrast, countries like Saudi Arabia, Oman and Palestine show substantially lower scores, often under 0.05. These findings indicate that fine-grained dialectal differences, possibly shaped by data distribution or linguistic features, have a significant effect on recommendation quality. This again advocates for the need to design recommendation systems that are inclusive and responsive to Arabic dialectal diversity.

3.3 Quantifying Multi-group Fairness

To quantify fairness disparities across multiple user groups, we build on the concept of *User-oriented Group Fairness (UGF)* introduced in (Li et al., 2021). We extend this notion to settings where users are divided into more than two groups based on the dialect they commonly use. Let $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_k\}$ be a partition of users into k disjoint groups. We define the multi-group fairness gap as:

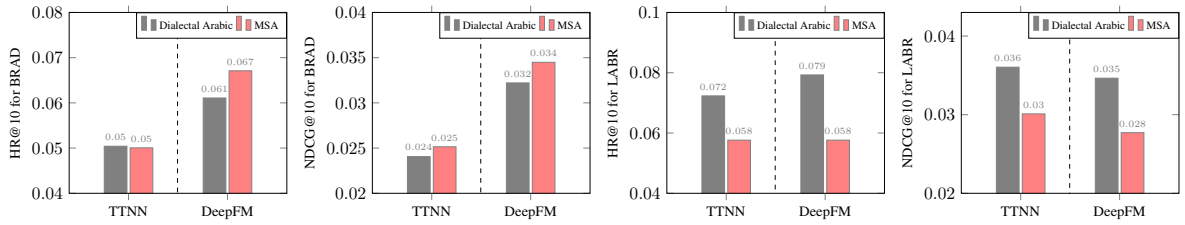

Figure 1: HR@10 and NDCG@10 for Dialectal Arabic and MSA.

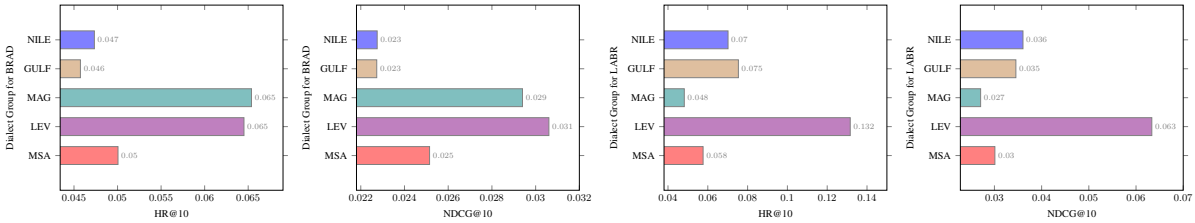

Figure 2: HR@10 and NDCG@10 for different dialects at the region level (TTNN).

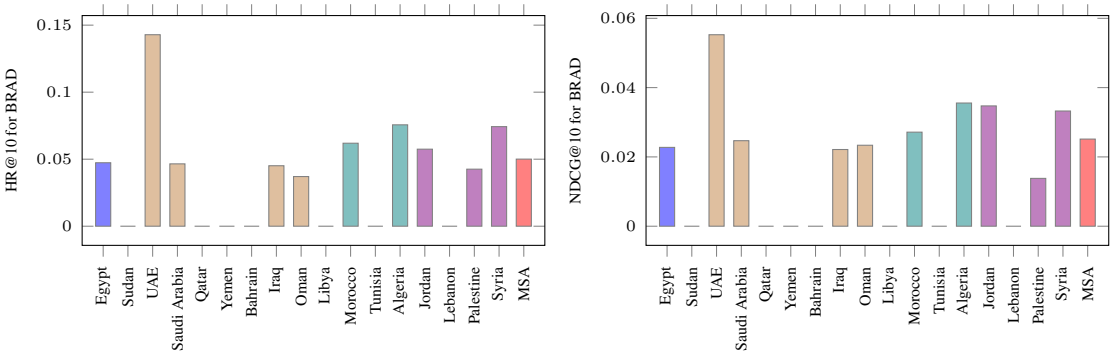

Figure 3: HR@10 and NDCG@10 for different dialects in BRAD dataset at the country level (TTNN).

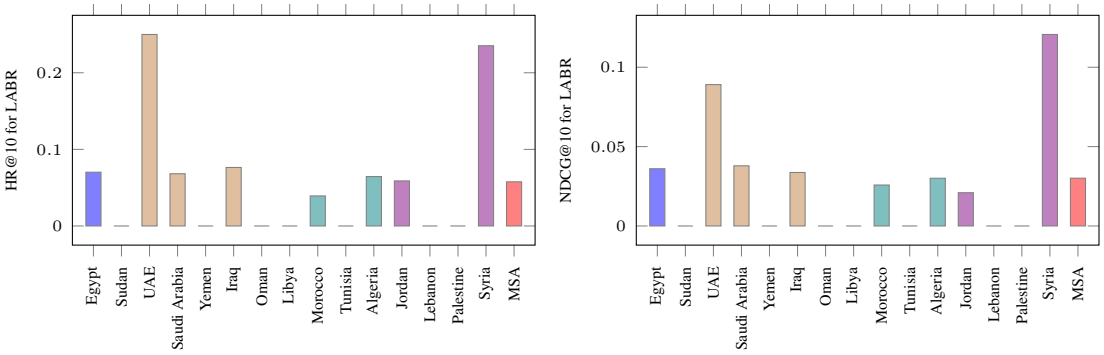

Figure 4: HR@10 and NDCG@10 for different dialects in LABR dataset at the country level (TTNN).

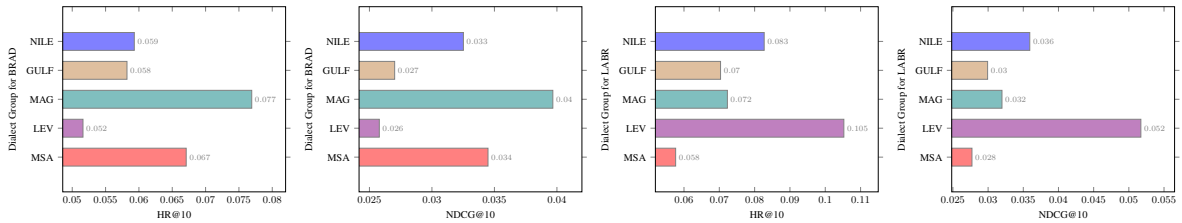

Figure 5: HR@10 and NDCG@10 for different dialect at the region level (DeepFM).

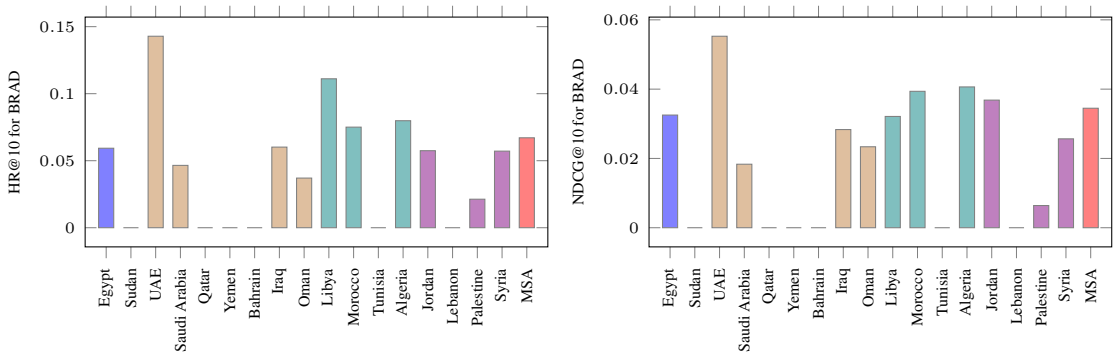

Figure 6: HR@10 and NDCG@10 for different dialects in BRAD dataset at the country level (DeepFM).

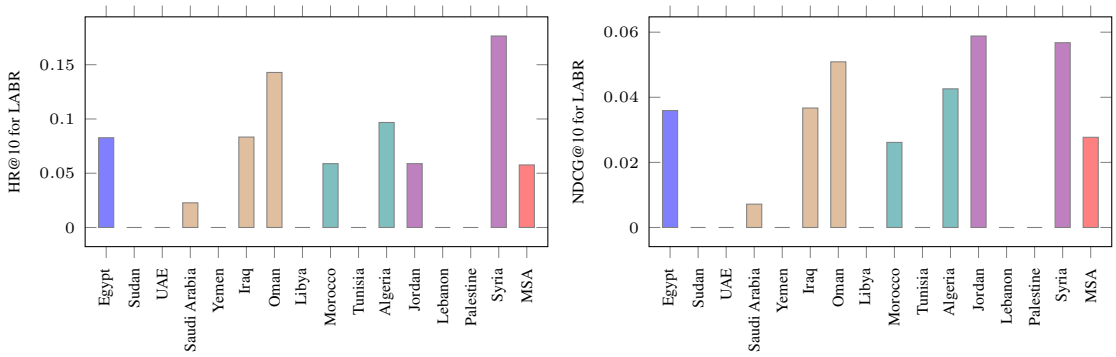

Figure 7: HR@10 and NDCG@10 for different dialects in LABR dataset at the country level (DeepFM).

$$\text{UGF}^*(\mathcal{Z}, W) = \max_{i,j} \left| \frac{1}{|Z_i|} \sum_{u \in Z_i} M(W_u) - \frac{1}{|Z_j|} \sum_{u \in Z_j} M(W_u) \right| \quad (2)$$

where $M(W_u)$ denotes the performance metric (e.g., HR@10 or NDCG@10) for user u, and W is the output of the recommendation model. Equation 2 computes the maximum performance gap between any pair of user groups in $\mathcal{Z}$.

In our experiments, we compute UGF* using both HR@10 and NDCG@10 across multiple grouping levels to measure fairness gaps in recommendation quality. The results, presented in Tables 1 and 2, show that the fairness gap increases as we move to finer-grained user groupings based on dialectal variation. This trend complements our earlier findings by highlighting that disparities become more pronounced when user groups are defined in a more fine-grained manner.

Overall, our multi-level analysis reveals a clear fairness issue: The quality of recommendations for Arabic-speaking users varies with the dialect they use, with some user groups benefiting more. This highlights the importance of considering linguistic diversity in evaluating and designing recommendation systems.

	BRAD		LABR	
	TTNN	DeepFM	TTNN	DeepFM
MSA vs. Dialectal	0.00036	0.00601	0.01469	0.02164
Regional	0.01965	0.02531	0.08339	0.04763
Country-level	0.14298	0.14683	0.25000	0.17651

Table 1: UGF* using HR@10 for BRAD and LABR.

	BRAD		LABR	
	TTNN	DeepFM	TTNN	DeepFM
MSA vs. Dialectal	0.00109	0.00228	0.00594	0.00693
Regional	0.00788	0.01388	0.03627	0.02400
Country-level	0.05526	0.05319	0.12064	0.05882

Table 2: UGF* using NDCG@10 for BRAD and LABR.

4 Conclusion

We uncover a fairness issue in recommendation systems for Arabic-speaking users, showing that performance varies significantly by dialect. Our analysis highlights the importance of accounting for linguistic diversity and show that NLP advancements can reveal hidden disparities in related fields.

5 Limitations

Our analysis is limited by the accuracy of dialect identification models and the availability of labeled data. We focus on written user reviews, which may not fully capture the variety of dialect use in other contexts such as spoken or informal communication. Additionally, we do not propose solutions, leaving fairness mitigation for future work.

6 Ethical Considerations

This study involves analyzing user-generated text to infer dialect, which may raise privacy concerns. All data used is publicly available and anonymized. We are careful not to associate dialect with identity, and our goal is to highlight disparities, not to profile users.

Acknowledgment

We sincerely thank all the reviewers for their time and constructive comments. This material is based upon work supported by REAL@USC-Meta center, and a Broadcom gift. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of our sponsors.

References

Muhammad Abdul-Mageed, Abdelrahim Elmadany, Chiyu Zhang, Houda Bouamor, Nizar Habash, et al. 2023. Nadi 2023: The fourth nuanced arabic dialect identification shared task. In *Proceedings of Arabic-NLP 2023*, pages 600–613.

Muhammad Abdul-Mageed, Amr Keleg, Abdelrahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. Nadi 2024: The fifth nuanced arabic dialect identification shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 709–728.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdelrahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. Nadi 2021: The second nuanced arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdelrahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. Nadi 2022: The third nuanced arabic dialect identification shared task. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97.

Abdulla Alshabanah and Murali Annavaram. 2025. On using Arabic language dialects in recommendation systems. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2178–2186, Albuquerque, New Mexico. Association for Computational Linguistics.

Abdulla Alshabanah, Keshav Balasubramanian, and Murali Annavaram. 2025. Meta-learn to unlearn: Enhanced exact machine unlearning in recommendation systems with meta-learning. *Proceedings on Privacy Enhancing Technologies*.

Maha Jarallah Althobaiti. 2020. Automatic arabic dialect identification systems for written texts: A survey. *International Journal of Computational Linguistics (IJCL)*, 11(3):61–89. Revised: 31-10-2020; Published: 01-12-2020.

Mohamed Aly and Amir Atiya. 2013. Labr: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498.

Keshav Balasubramanian, Abdulla Alshabanah, Elan Markowitz, Greg Ver Steeg, and Murali Annavaram. 2024. Biased user history synthesis for personalized long-tail item recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 189–199, New York, NY, USA. Association for Computing Machinery.

A. Bergman and Mona Diab. 2022. Towards responsible natural language annotation for the varieties of Arabic. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 364–371, Dublin, Ireland. Association for Computational Linguistics.

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, New York, NY, USA.

Ashraf Elnagar and Omar Einea. 2016. Brad 1.0: Book reviews in arabic dataset. In *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE.

Ashraf Elnagar, Leena Lulu, and Omar Einea. 2018. An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis. *Procedia computer science*, 142:182–189.

Charles A. Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

Carlos A. Gomez-Uribe and Neil Hunt. 2016. The netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manage. Inf. Syst.*, 6(4).

Asnat Greenstein-Messica and Lior Rokach. 2018. Personal price aware multi-seller recommender system: Evidence from ebay. *Know.-Based Syst.*, 150(C):14–26.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: A factorization-machine based neural network for CTR prediction. *CoRR*, abs/1703.04247.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. *CoRR*, abs/1708.05031.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Amr Keleg, Walid Magdy, and Sharon Goldwater. 2024. Estimating the level of dialectness predicts inter-annotator agreement in multi-dialect Arabic datasets. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 766–777, Bangkok, Thailand. Association for Computational Linguistics.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented fairness in recommendation. In *Proceedings of the web conference 2021*, pages 624–632.

G. Linden, B. Smith, and J. York. 2003. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80.

Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep learning recommendation model for personalization and recommendation systems. *CoRR*, abs/1906.00091.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pages 995–1000.

Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Ajit Kumthekar, Zhe Zhao, Li Wei, and Ed Chi, editors. 2019. *Sampling-Bias-Corrected Neural Modeling for Large Corpus Item Recommendations*.

A Datasets Details

Tables 3, 4, 5 and 6 show statistical information of BRAD and LABR datasets.

	BRAD	LABR
Number of Users	33,986	5891
Number of Items	4978	1333
Number of Interactions	455,912	45,681
Sparsity	99.73%	99.42%

Table 3: Statistics of the datasets.

	BRAD	LABR
Dialectal	24644	3872
MSA	9342	2019

Table 4: User group statistics (Dialectal vs. MSA)

	BRAD	LABR
Nile	14895	2130
Gulf	4451	973
Levant	1323	178
Maghreb	3975	591
MSA	9342	2019

Table 5: User group statistics by region

B Implementation Details

We implement all models using PyTorch (Paszke et al., 2019) and run our experiments on a machine with an Nvidia RTX 5000 GPU and an AMD EPYC 7502 CPU. Each model uses an embedding dimension of 96, GeLU activation, and layer normalization between layers. We follow a common tower design where the hidden layers get smaller as we go deeper, each higher layer has half as many neurons as the one below it. We use the Adam optimizer, and select hyperparameters through a grid search. The learning rate is chosen from 0.2, 0.02, 0.002, 0.0002, 0.00002, and the batch size from 32, 64, 128, 512, 1024. Our final setup uses a learning rate of 0.00002 and a batch size of 1024. Training uses a max-margin ranking loss, which pushes positive examples to have higher scores than negative ones. For each positive edge, we sample 20 negative edges. Each epoch takes about 2 minutes, and we train for up to 100 epochs. On average, training takes around 10 GPU hours. To ensure stable results, we report the average across three runs for each experiment. More details about the model implementation can be found in the link of the source code in the abstract.

	BRAD	LABR
Algeria	2235	400
Bahrain	3	0
Egypt	14889	2127
Iraq	3450	702
Jordan	485	89
Lebanon	38	3
Libya	100	9
MSA	9342	2019
Morocco	1614	180
Oman	341	95
Palestine	215	25
Qatar	2	0
Saudi Arabia	563	155
Sudan	6	3
Syria	585	61
Tunisia	26	2
UAE	58	11
Yemen	34	10

Table 6: User group statistics by country

C Uncovering Fairness Disparities (Matrix Factorization)

In this section, we report the HR@10 and NDCG@10 results for a Matrix Factorization model at three levels: MSA vs. dialectal Arabic (Table 7), regional dialect groups (Table 8), and country-level dialects (Table 9).

The reported results complement what we presented earlier in Section 3, showing that there is fairness disparity across all grouping levels.

	BRAD		LABR	
	HR@10	NDCG@10	HR@10	NDCG@10
Dialectal	0.03510	0.01621	0.07093	0.03313
MSA	0.03355	0.01487	0.09798	0.04408

Table 7: HR@10 and NDCG@10 for Dialectal Arabic and MSA (Matrix Factorization).

	BRAD		LABR	
	HR@10	NDCG@10	HR@10	NDCG@10
NILE	0.03138	0.01545	0.08521	0.03956
GULF	0.03638	0.01602	0.06030	0.02726
MAG	0.04359	0.01745	0.04819	0.02742
LEV	0.05484	0.02295	0.02632	0.00877
MSA	0.03355	0.01487	0.09798	0.04408

Table 8: HR@10 and NDCG@10 for Arabic varieties and MSA (Matrix Factorization).

	BRAD		LABR	
	HR@10	NDCG@10	HR@10	NDCG@10
Egypt	0.03138	0.01545	0.08521	0.03956
Sudan	0.00000	0.00000	0.00000	0.00000
UAE	0.28571	0.13295	0.00000	0.00000
Saudi Arabia	0.01550	0.00500	0.02273	0.00684
Qatar	0.00000	0.00000	–	–
Bahrain	0.00000	0.00000	–	–
Yemen	0.00000	0.00000	0.00000	0.00000
Iraq	0.03885	0.01734	0.07639	0.03558
Oman	0.00000	0.00000	0.00000	0.00000
Libya	0.11111	0.05556	0.00000	0.00000
Morocco	0.04315	0.01739	0.03922	0.01936
Tunisia	0.00000	0.00000	0.00000	0.00000
Algeria	0.04202	0.01615	0.06452	0.04158
Jordan	0.03448	0.01202	0.05882	0.01961
Lebanon	0.00000	0.00000	0.00000	0.00000
Palestine	0.02128	0.00823	0.00000	0.00000
Syria	0.07429	0.03246	0.00000	0.00000
MSA	0.03355	0.01487	0.09798	0.04408

Table 9: HR@10 and NDCG@10 for country-level Arabic varieties and MSA (Matrix Factorization).