An Analysis of the Evolution of Music Listening Data and the Need for Task Discernment

Shahrzad Shashaani[1,*], Pavan Seshadri[*] and Peter Knees[1]

[1]TU Wien, Faculty of Informatics, Vienna, Austria

Abstract

With the availability of music streaming platforms, listening behavior has seen fundamental changes in the past two decades, going from mere consumption of and recommendation within personal collections to an exploration of massive catalogs. As part of this trend, collaborative filtering algorithms that exploit consumption data, user feedback, and, most recently, the sequential order of music consumption, have become indispensable.

In prior work, it has been shown that the incorporation of negative feedback (skipped track information) via contrastive learning can be applied to and improve existing sequential recommendation models. In this work, we extend previous findings by investigating two notable aspects of music listening data in detail. First, we analyze popular public datasets used in music recommender systems research (LFM-1k, LFM-2B, and the Music Streaming Sessions Dataset) with respect to the evolution of consumption activity and track skipping behavior, and show strongly deviating patterns based on data creation context. Second, focusing on LFM-2B, we further study the impact of data and skipping information availability on sequential and non-sequential recommendation algorithms over the different years available in the data set. We observe deviating model performance using time-based subsets of LFM-2B compared to experiments on the entire dataset. In conclusion, we argue for more careful discernment and understanding of listening tasks and user intents leading to creating datasets, as well as explicitly modeling different types of interactions.

Keywords

Sequential Recommendation, Music Recommendation, Contrastive Learning

1. Introduction

Music recommender systems have become central in shaping how users interact with streaming platforms, significantly influencing music listening and discovery. These systems have evolved over the past two decades from simple personal collection recommendations to sophisticated tools capable of navigating vast music catalogs. This evolution has been driven by the increasing availability of streaming data and the continuous advancement of recommendation algorithms.

Recent studies have highlighted the importance of understanding user behavior to improve the effectiveness of music recommender systems. For instance, Hidasi et al. (2020) emphasized the role of contextual information in music recommendation, by modeling the whole session, to achieve more accurate results [1]. While Quadrana et al. (2020) explored how multiple user-item interactions influence recommendation quality in a sequence-aware recommender system [2]. Furthermore, research by Wen et al. (2019) has shown that incorporating user feedback, such as track skips and short plays, can significantly enhance recommendation accuracy [3]. Dai et al. (2024) modeled user attention prediction as a positive-unlabeled learning problem, where active feedback is treated as positive samples and passive feedback is treated as unlabeled samples to increase user engagement [4].

In parallel, the availability of extensive datasets has provided valuable resources for analyzing user interactions and behavior on a large scale. For example, Yao et al. (2020) utilized these datasets to study session-based recommendations, highlighting the need for models that can adapt to evolving user preferences [5].

MuRS 2025: 3rd Music Recommender Systems Workshop, September 22nd, 2025, Prague, Czech Republic

*Corresponding author.

shahrzad.shashaani@tuwien.ac.at (S. Shashaani); pavanseshadri1@acm.org (P. Seshadri); peter.knees@tuwien.ac.at (P. Knees)

0000-0003-4344-2696 (S. Shashaani); 0009-0008-7838-9614 (P. Seshadri); 0000-0003-3906-1292 (P. Knees)

Furthermore, incorporating real negative feedback, such as user skips, significantly enhances recommender systems. Mei et al. (2024) have shown that using explicit negative samples reduces training time and improves test accuracy [6]. Pan et al. (2023) showed that, in sequential recommendation tasks, leveraging passive negative feedback, like video skips, provides crucial insights into user preferences [7]. Methods combining positive and passive-negative feedback through sub-interest encoders have demonstrated superior performance, highlighting the importance of diverse feedback types for improving accuracy and user satisfaction. In a related direction, we proposed a contrastive learning framework that directly incorporates skip behavior as informative negative signals. This method improves sequential music recommendation by aligning user representations with positively preferred tracks and pushing away tracks associated with negative feedback, thus leveraging fine-grained temporal dynamics of skips for better modeling [8].

In this work, we aim to identify trends in music consumption and track skipping behavior across public datasets used in music recommender systems research. By focusing on the LFM-1k, LFM-2B, and Music Streaming Sessions Dataset (MSSD), we analyze the evolution of these behaviors and their implications for recommender system design. Focusing on LFM-2B for in-depth evaluation, for each contained year individually, we explore SASRec [9] as a sequential model, negative feedback enhanced SASRec [8], and two non-sequential baseline algorithms, Weighted Regularized Matrix Factorization (WRMF) [10] and Bayesian Personalized Ranking (BPR) [11], using the methodology introduced by Wen et al. [3] to incorporate negative feedback.

This work aims to fill a gap in existing research by focusing on how music listening behavior has evolved over the past two decades and how these changes impact recommender systems. Unlike previous studies that have largely ignored this aspect, our study examines trends across public datasets used in music recommender systems research, specifically looking at consumption activity and track-skipping behavior. By comparing sequential and non-sequential recommendation algorithms incorporating negative feedback against feedback-agnostic baselines, we demonstrate the increasing importance of integrating different forms of interaction into recommender models. This underscores the necessity of understanding listening tasks and user intents to create better datasets and explicitly model various types of interactions.

2. Datasets

For this study, we use three real-world music recommendation datasets: the Music Streaming Sessions Dataset (MSSD) [12] using data from Spotify,[1] the LFM-2B dataset [13], and the LFM-1k dataset [14, 15], both using data from Last.fm.[2]. These datasets have been instrumental in uncovering patterns in music consumption and track-skipping behavior, which are critical for refining recommendation algorithms.

2.1. Music Streaming Sessions Dataset

The MSSD contains 160 Million user sessions of 10 to 20 consecutively listened songs (<60 seconds between listens), which are uniformly sampled from a variety of contexts, such as the user's personally selected collections, expertly selected playlists, contextual non-personalized recommendations, and personalized recommendations. As this dataset is pseudonymized and lacks user labels, we can treat each session as a new user for recommendation tasks.

Each listening event contains a skip label from 0-3, with 0 denoting "no skip" and 1-3 denoting the length of time before a user skipped a given track. This is defined as "played very briefly", "played briefly", and "played mostly (but not completely)", respectively for labels 1 to 3. In this study, we would mainly assume skip label 3 not to be a strong indicator of negative preference for the given session, so we only assume labels 1 and 2 as true skips.

[1] https://open.spotify.com
[2] https://last.fm

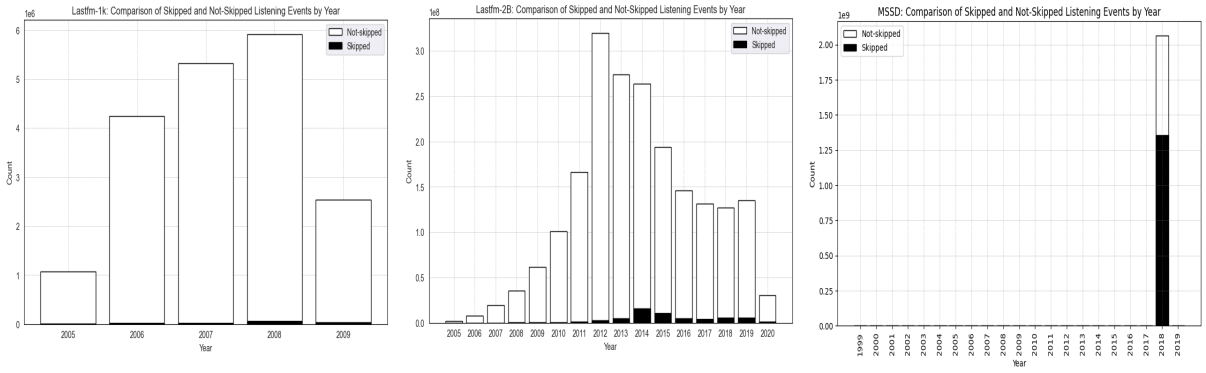

Figure 1: Skip counts per year for LFM-1k, LFM-2B, and Spotify(MSSD) datasets, respectively. Each stack plot presents the total number of listening events for each year, with skips represented by the black segments and not-skipped events by the white segments. The entirety of each bar illustrates the total number of listening events recorded in that particular year across the datasets.

2.2. LFM-1k

The LFM-1k dataset contains 19M discrete listening events for 1000 users, containing time stamps, user IDs, track IDs, and track names for each event. Therefore, we create implicit sessions, such that for each user, we consider a sequence of chronological listening events with less than 20 minutes between any individual event as a "session". We consider skips to be any prior listening event with less than 30 seconds between its subsequent event. We additionally prune any sessions with fewer than 5 events for a total of 650K sessions, with around 1M unique tracks. In line with the MSSD, we can discard user labels and treat each individual session as a new user in recommendation tasks.

2.3. LFM-2B

The LFM-2B dataset contains more than 2 billion listening events for 120,322 users and 50,813,373 tracks, which is collected over 15 years (from 2005 until 2020). The dataset includes demographic details of users (such as age, country, and gender), metadata related to music (such as artist and track names), and timestamps indicating the exact time when a user listened to a specific track [13]. Similarly to the other datasets, we discard metadata and demographic information. Similar to the LFM-1k dataset, we define implicit "sessions" for each user by grouping together a chronological sequence of listening events where the time interval between any two consecutive events is less than 20 minutes. We define skips as any previous listening event followed by another event within a time difference of less than 30 seconds. In addition, we remove user labels and treat each session as a distinct user.

We applied the same session extraction process to all yearly subsets of the data. However, the numbers of sessions per year are significantly larger than in the 2020 subset reported earlier in [8] — with over 250 million interactions recorded annually between 2012 and 2014, as shown in Figure 1. Therefore, after creating the sessions as described earlier, we randomly sampled 100k sessions from each year. These sampled sessions were then used to evaluate the different methods.

3. Dataset Analysis and Comparison

For clarity, we note that the following analyses are performed on the complete datasets and not the evaluation subsets used for impact analysis (sec. 4). Figure 1 illustrates the number of listening events and skips per year for each dataset separately. It is notable that a significant portion of the Spotify (MSSD) dataset's listening events was gathered in 2018, showing an uneven distribution over the dataset's collection period. In contrast, the data collection distribution for the LFM-1k and LFM-2B datasets appears more even within the 5 and 15 years data collection period, respectively. In addition, the skip percentages are relatively low, showing a maximum skip percentage of 1.5% for LFM-1k and

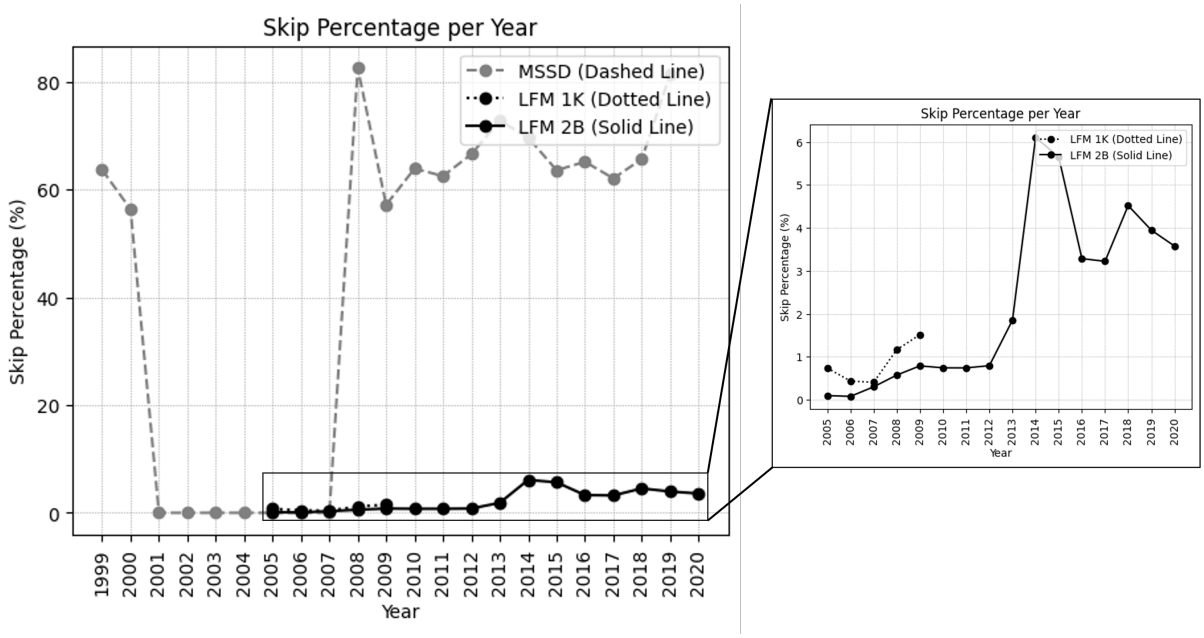

Figure 2: Comparison of skips relative to total listening events per year for LFM-1k, LFM-2B, and Spotify-MSSD datasets.

6.1% for LFM-2B datasets. However, the skip percentage is considerably higher for the Spotify (MSSD) dataset, which reached ~ 66% in 2018 and higher in other years when less data has been gathered. The high skip rate in the data may be related to the different types of skip behavior that were represented in the data [12].

The overall skip comparison between the three datasets is illustrated in Figure 2. As expected, the Spotify (MSSD) dataset has a higher skip rate than others in the same time interval. For the LFM-2B dataset, there is no clear pattern indicating whether the number of skips increases along with the number of listening events, or if the number of skips increases/decreases in the corresponding years. However, the overall skip percentage for LFM-2B is lower than LFM-1k within the same time interval. Regarding the LFM-1k dataset, there appears to be a trend of increasing skips over the years. However, since the data is presented for only a five-year sequence, it is not conclusive evidence of a continuous increase. While the overall skip trend for LFM-1k and LFM-2B datasets is similar, there are considerable differences between them, particularly shown in the subplot in Figure 2. Notably, for the LFM-2B dataset, the overall skip rate increased after 2012, despite a decrease in overall listening events. This observation motivates our current study, in which we focus on examining the LFM-2B dataset from 2012 onwards in more detail.

4. Experiments

Previously, Seshadri et al. conducted experiments on a selection of well-established sequential and non-sequential recommendation models across different datasets to evaluate the impact of integrating negative feedback [8]. Building on this, we extend the analysis by examining the performance of selected models across multiple yearly subsets of the LFM-2B dataset. We focus on three representative models: Bayesian Personalized Ranking (BPR) [11], Weighted Regularized Matrix Factorization (WRMF) [10], and Self-Attentive Sequential Recommendation (SASRec) [9], where the first two models (BPR, WRMF) are non-sequential approaches. Each model was selected for its distinct approach to modeling user preferences—BPR and WRMF rely on matrix factorization techniques, while SASRec employs a self-attention mechanism for sequential modeling. All models were configured according to the original implementations and subsequently modifications incorporating negative feedback.

We aim to investigate how the integration of negative feedback influences model performance across

these different approaches and whether temporal variation (i.e., using different yearly subsets) leads to noticeable differences in results.

For SASRec, we use the same training strategies as in [8]. This includes modifications to standard practices to incorporate negative feedback effectively into sequential recommender systems. We use a sampled softmax approach with negative log-likelihood to handle the extensive item space, which ranges from 300,000 to 500,000 items. During each training iteration, we sample 1,000 unseen items to rank against the target items. This approach allows the model to adjust to an expanding subset of items, thereby enhancing its ability to rank items accurately as training progresses.

To ensure consistency, we set the sequence length limit to 20 items, splitting longer sessions into multiple segments. Model embeddings and hidden layers are uniformly set to a dimension of 128. For SASRec, we use two layers of self-attention with eight attention heads each. Initial parameters are sampled from a truncated normal distribution with $\mu = 0$ and $\sigma = 1$ within a range of $[-0.02, 0.02]$. We optimize the α and β parameters from the set [0.1, 0.2, 0.5, 0.75, 1], choosing $\alpha = 1$ and $\beta = 0.2$ for the LFM-2B dataset. The ADAM optimizer [16] is used with a learning rate of 0.005.

The training process varies by model. For SASRec, we reserve the final and penultimate items in each session for testing and validation purposes. For BPR and WRMF, we focus on predicting the next item in a sequence. We integrate negative feedback into these models, following the methods proposed in [3]. These methods include *-BL*, which adjusts preference labels based on post-click feedback, and *-NR*, which probabilistically samples items across different feedback types.

5. Discussion of Results

The results of evaluating on year-based subsets are shown in Table 1. As described earlier, we randomly sampled 100k sessions for each year and calculated the skip percentage in both the sampled subset and the original dataset. As shown in the table, the overall skip occurrence is preserved.

In line with previous results we can see that sequential models outperform the non-sequential baselines consistently. However, we can also observe less consistency regarding the impact of incorporating negative feedback. While we can see some improvements on the sequential model in the earlier years (2012–2014) and the last years (2019–2020), negative feedback improves the baseline models in all subsets, albeit at a much lower performance level and dependent on the strategy chosen. For the sequential model, there seems to be no clear trend of impact based on the skip ratio of the corresponding year alone, indicating that different years exhibit different patterns and that a generalization of results on individual subsets is not possible.

To further investigate this indication, we assess the effect of the proportion of skip events relative to total interactions in a dataset for one of the subsets. The results are shown in Table 2. Changing the sampling method increases the proportion of skips in our used dataset. Consequently, methods that incorporate negative feedback (skips) benefit from this change and achieve better performance. This becomes more pronounced when comparing with our previous work [8], where we applied an oversampling processing on the 2020 LFM-2B subset that resulted in a much higher skip ratio (around ~14% compared to 3.86% in our current sampling method) and leading to a much better outcome—e.g., HR@1 for SASRec (original and negative-feedback versions) reached .190 and .221, resp. in the oversampled setup, a result which could only be achieved by current sampling process at HR@20. The two non-sequential models showed a similar pattern, while oversampling led to better results for them, they were more robust to changes and showed smaller performance differences in comparison. These findings highlight how important maintaining a higher ratio of skip interactions is for this approach, once again confirming the conclusions in [8].

6. Conclusion

From the results obtained, we can see the potential of incorporating negative feedback, however with a high sensitivity of algorithms regarding the underlying data. While the effect is more robustly seen

Table 1
Performance of LFM-2B from 2012-2020 using *Hit Ratio @ [1, 5, 10, 20]* for the sequential model (SASRec) and the non-sequential baselines (WRMF, BPR). Bold faced entries in the SASRec Neg., -BL, and -NR columns indicate an improvement of the negative feedback incorporating approach over the feedback agnostic versions. The Skip Ratio column shows the percentage of skip events in our sampled subset (with the overall percentage of skips in that year in parentheses for reference).

Year	Skip	Metric	SASRec		WRMF			BPR		
			Orig.	Neg.	Orig.	-BL	-NR	Orig.	-BL	-NR
2012	0.59% (0.81%)	HR@1	.129	.116	.000	**.006**	**.004**	.000	**.009**	**.008**
		HR@5	.158	.151	.000	**.015**	**.013**	.000	**.026**	**.024**
		HR@10	.167	**.168**	.001	**.025**	**.023**	.001	**.042**	**.040**
		HR@20	.177	**.181**	.007	**.079**	**.075**	.009	**.087**	**.083**
2013	1.84% (1.88%)	HR@1	.165	**.172**	.000	**.003**	**.006**	.000	**.007**	**.007**
		HR@5	.205	**.210**	.000	**.011**	**.014**	.000	**.022**	**.020**
		HR@10	.222	**.224**	.001	**.022**	**.029**	.001	**.041**	**.038**
		HR@20	.234	.234	.006	**.071**	**.080**	.007	**.084**	**.080**
2014	6.36% (6.14%)	HR@1	.303	**.320**	.012	**.018**	.012	.013	**.014**	**.022**
		HR@5	.354	**.365**	.039	.037	**.042**	.039	**.047**	**.067**
		HR@10	.369	**.378**	.061	**.086**	**.066**	.065	**.075**	**.099**
		HR@20	.384	**.391**	.133	**.163**	**.138**	.137	**.150**	**.170**
2015	5.50% (5.67%)	HR@1	.268	.101	.006	**.008**	**.010**	.012	**.014**	.012
		HR@5	.304	.154	.013	**.016**	**.029**	.025	**.043**	**.028**
		HR@10	.315	.180	.028	**.029**	**.039**	.063	**.071**	**.071**
		HR@20	.327	.212	.082	**.086**	**.086**	.122	**.143**	**.129**
2016	3.57% (3.31%)	HR@1	.276	.084	.019	**.020**	.015	.019	**.021**	.018
		HR@5	.315	.136	.051	**.058**	.047	.040	**.063**	**.051**
		HR@10	.329	.167	.089	**.100**	.089	.091	**.102**	**.094**
		HR@20	.315	.204	.161	**.168**	.159	.162	**.171**	**.164**
2017	3.18% (3.26%)	HR@1	.213	.047	.019	.018	.018	.014	**.017**	**.018**
		HR@5	.256	.083	.059	**.076**	.036	.047	.036	.035
		HR@10	.277	.104	.100	**.129**	.089	.073	**.089**	**.089**
		HR@20	.301	.129	.166	**.270**	**.178**	.157	**.159**	**.165**
2018	4.24% (4.56%)	HR@1	.062	.032	.011	.010	**.012**	.010	**.012**	**.011**
		HR@5	.153	.058	.025	.025	.025	.022	**.038**	.022
		HR@10	.198	.072	.058	.037	**.069**	.055	**.065**	**.058**
		HR@20	.240	.093	.119	.078	**.121**	.103	**.136**	**.106**
2019	3.79% (3.99%)	HR@1	.059	**.063**	.018	**.020**	.017	.016	**.017**	**.019**
		HR@5	.149	.140	.047	**.055**	.042	.028	**.038**	**.054**
		HR@10	.191	.174	.081	**.087**	.079	.087	**.089**	**.096**
		HR@20	.229	.204	.162	**.173**	.155	.159	**.166**	**.188**
2020	3.86% (3.62%)	HR@1	.082	.065	.010	**.013**	.010	.021	**.060**	**.071**
		HR@5	.138	**.139**	.021	**.025**	.021	.039	**.120**	**.135**
		HR@10	.165	**.174**	.060	**.074**	**.067**	.094	**.156**	**.168**
		HR@20	.194	**.208**	.152	**.165**	**.154**	.168	**.202**	**.213**

with non-sequential models, their overall performance is limited. For the sequential algorithm, we see a much higher dependency on the choice of subsets, availability of negative data, and sampling method due to the contrastive learning approach.

Other conclusions concern the bias in the data distribution over different years represented in the data—not only in the descriptive analysis and across datasets but also impacting the predictive capabilities of models. The overall performance on LFM-2B based on the originally provided 2020

Table 2
Performance comparison on LFM-2B dataset across different skip numbers for 2019.

Year	Metric	SASRec (100k sampled)	# of skips	SASRec (20 int./sess. sampled)	# of skips
		Negative-feedback		Negative-feedback	
2019	HR@1	.063	5928	.033	62396
	HR@5	.140		.119	
	HR@10	.174		.185	
	HR@20	.204		.270	

subset, as shown in [8], is not necessarily indicative of individual annual subsets, which by themselves should represent different trends in consumption. The observed trends in the data likely reflect the dataset creation process in addition to the consumption patterns of these years.

Extrapolating from these findings also to other datasets, it is inherent that individual datasets represent only some aspects of the larger picture and overall trends of shifting consumption patterns. The models learned from a single dataset are therefore limited in terms of validity and generalization. However, even with the zoo of music recommender datasets available (or partly not available anymore, e.g. [13]), one can not easily discern listening trends and listening modalities. While certain preferences and temporal phenomena are only represented in some datasets capturing data from the respective time, they are strongly linked to the platforms, applications, and recommendation paradigms of that time. This presents a conundrum as one can not simply "add up" different datasets to create a larger and more "complete" pool. Instead, we first need to understand the individual backgrounds, foci, tasks, and intents captured in and connected to the individual datasets, before devising strategies to craft a more holistic picture of music listening preferences—if that is considered a goal worthwhile and a relevant research question.

Acknowledgments

This research was funded in whole or in part by the Vienna Science and Technology Fund (WWTF) [Grant ID: 10.47379/DCDH001]. For open access purposes, the author has applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk, Session-based recommendations with recurrent neural networks, arXiv preprint arXiv:1511.06939 (2015).

[2] M. Quadrana, P. Cremonesi, D. Jannach, Sequence-aware recommender systems, ACM computing surveys (CSUR) 51 (2018) 1–36.

[3] H. Wen, L. Yang, D. Estrin, Leveraging post-click feedback for content recommendations, in: Proceedings of the 13th ACM Conference on Recommender Systems, RecSys '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 278–286. URL: https://doi.org/10.1145/3298689.3347037. doi:10.1145/3298689.3347037.

[4] S. Dai, N. Shao, J. Zhu, X. Zhang, Z. Dong, J. Xu, Q. Dai, J.-R. Wen, Modeling user attention in music recommendation, in: 2024 IEEE 40th International Conference on Data Engineering (ICDE), IEEE, 2024, pp. 761–774.

[5] H. Yao, J. Hu, W. Xie, Y. Huang, W. Xie, Session-based recommendation model based on multiple neural networks hybrid extraction feature, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 5315–5322.
[6] M. J. Mei, O. Bembom, A. F. Ehmann, Negative feedback for music personalization, in: Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, 2024, pp. 195–200.
[7] Y. Pan, C. Gao, J. Chang, Y. Niu, Y. Song, K. Gai, D. Jin, Y. Li, Understanding and modeling passive-negative feedback for short-video sequential recommendation, in: Proceedings of the 17th ACM conference on recommender systems, 2023, pp. 540–550.
[8] P. Seshadri, S. Shashaani, P. Knees, Enhancing sequential music recommendation with negative feedback-informed contrastive learning, in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 1028–1032.
[9] W.-C. Kang, J. McAuley, Self-attentive sequential recommendation, in: 2018 IEEE international conference on data mining (ICDM), IEEE, 2018, pp. 197–206.
[10] Y. Hu, Y. Koren, C. Volinsky, Collaborative filtering for implicit feedback datasets, in: 2008 Eighth IEEE international conference on data mining, Ieee, 2008, pp. 263–272.
[11] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI 2009), 2009. doi:10.48550/arXiv.1205.2618.
[12] B. Brost, R. Mehrotra, T. Jehan, The music streaming sessions dataset, in: L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, L. Zia (Eds.), The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, ACM, 2019, pp. 2594–2600. URL: https://doi.org/10.1145/3308558.3313641. doi:10.1145/3308558.3313641.
[13] M. Schedl, S. Brandl, O. Lesota, E. Parada-Cabaleiro, D. Penz, N. Rekabsaz, Lfm-2b: A dataset of enriched music listening events for recommender systems research and fairness analysis, in: Proceedings of the 2022 Conference on Human Information Interaction and Retrieval, 2022, pp. 337–341.
[14] Ò. Celma, Music Recommendation and Discovery in the Long Tail, Springer, 2010.
[15] Ò. Celma, lastfm music recommendation dataset, 2022. URL: https://doi.org/10.5281/zenodo.6090214. doi:10.5281/zenodo.6090214.
[16] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1412.6980.