Aligning Emotions with Lyrics for Personalized Recommendations
Project Category: Audio & Music

Diego Valdez Duran
Department of Computer Science, Stanford University
diegoval@stanford.edu

1. Introduction

Modern music streaming platforms primarily utilize content-based filtering and collaborative filtering techniques to generate music recommendations. While these methods leverage song metadata and user interaction data to predict preferences and offer recommendations, they often neglect critical real-time factors such as users' emotional states. Existing research has shown that a listener's mood—characterized by psychological factors like valence ("goodness") and arousal (alertness)—significantly influences music preferences. Traditional recommendation systems, by not accounting for these dynamic emotional and contextual states, often provide recommendations that lack emotional resonance and personalization.

This project aims to improve upon traditional recommendation systems by incorporating users' emotional states and situational context, along with sentiment analysis of song lyrics. It builds on the SiTunes dataset [5] across three context groups—*Objective* (Obj.), *Subjective* (Sub.), and their combination (*Obj. + Sub.*)—recorded in two settings. The Situational Data (Obj.) includes measurable environmental factors (e.g., weather, location) and physiological features (e.g., heart rate, activity intensity). The Emotional State Data (Sub.) incorporates user-reported emotional annotations (valence and arousal) before and after music listening. Additionally, we use Song Metadata from SiTunes and web-scraped lyrics for sentiment analysis, linking lyrical content to traditional song features.

We use classification models to predict mood changes across recorded contexts, using both objective and subjective features. Semantic features, such as narrative complexity, emotional sophistication, thematic elements, and temporal patterns, are extracted from lyrics using a fine-tuned LLaMa model. These lyrical features are integrated with user preference data, and hierarchical clustering groups users based on preference similarity to generate personalized recommendations. Songs with similar features are then recommended to users.

The project produces two main outputs: (1) Mood Prediction Models, which predict emotional changes based on situational and physiological data; and (2) User Clusters, which organize users by preferences for generating song recommendations based on lyrical sentiment and these preferences. User preferences are derived from interaction data, including rating-based and emotional features, for enhanced personalization.

2. Related Work

Recommender systems have improved recently by integrating contextual and semantic features for more personalized experiences. Context-Aware Recommender Systems (CARS) address traditional methods' limitations by incorporating temporal, environmental, and user-reported data. For example, [8] demonstrates the effectiveness of deep learning models, such as RNNs, in capturing sequential and contextual data, while [2] emphasizes the importance of selecting relevant contextual features to improve recommendation quality. In other industries, collaborative filtering is also common such as in [6].

Emotion-aware music recommendations have also become more common, with psychological features shaping user preferences. [3] highlights the use of mood-based user inputs in music therapy and session-based recommendations, aligning with our secondary objective. Similarly, [2] shows how contextual factors influence genre preferences, which supports our integration of psychological data with sentiment analysis of song lyrics. Sentiment analysis and song metadata used together have improved content-based filtering. [9] and [1] explore genre, style, and lyrics embeddings, with [1] using CRNNs for genre-based recommendations.

Large-scale datasets, such as the Yahoo! Music Dataset [4] and the Music Streaming Sessions Dataset [3], have driven advancements in recommendation models. These datasets highlight the importance of robust data representation, guiding our use of the SiTunes dataset. [10] discusses reward modification techniques to address biases in user feedback, though we do not incorporate these due to the nature of the SiTunes data.

The SiTunes dataset [5] forms the foundation of our project, offering unique physiological, psychological, and contextual data. Its three-stage user study captures both preferences and situational responses, enabling analysis of objective and subjective features. By combining SiTunes data with sentiment analysis, our

project predicts emotional changes and clusters users based on emotional alignment and lyrical content.

3. Dataset

This project uses the SiTunes dataset, a three-stage user study, to explore two complementary tasks: mood prediction and thematically resonant music recommendation. Situational features from Stages 2 and 3 are used in the prediction task, while Stage 1 data informs lyrical sentiment analysis and user preference learning.

3.1. Prediction Task: Mood Changes

The prediction task uses objective situational data to model emotional changes following music listening. Stage 2 includes 897 interactions from 30 users and 105 tracks, while Stage 3 provides 509 interactions from 10 users and 217 tracks. Each record contains metadata (e.g., user IDs, timestamps, ratings), physiological signals (e.g., heart rate, activity intensity), environmental context (e.g., weather, time of day), and pre- and post-listening emotional valence and arousal scores.

Preprocessing involved filtering interactions with ratings greater than 3 as positive examples. Physiological data was aggregated over 30-minute windows using mean and standard deviation, while activity type was represented by the majority activity within each window. Continuous features were normalized to the range [0, 1]. The dataset was then partitioned into three experimental settings to asses mood predictions.

3.2. Recommendation Task: Lyrical Sentiment and Clustering

The recommendation task integrates lyrical sentiment analysis, user preference learning, and clustering. Song lyrics were scraped from Genius using titles and artist information provided in the SiTunes metadata. A fine-tuned LLaMA model was used to extract lyrical features such as narrative complexity, emotional sophistication, thematic elements (e.g., love, life), and temporal focus. These lyrical features were indexed by song and visualized with t-SNE to uncover relationships among songs.

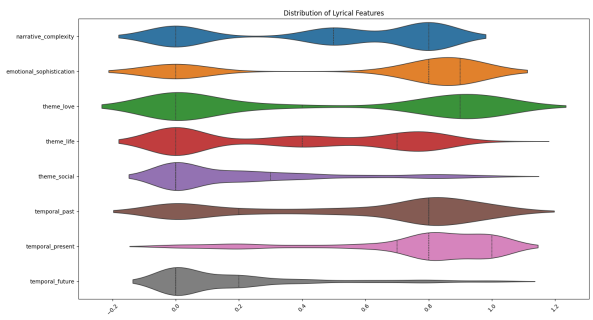

Figure 1. Feature Distributions from fine-tuned LLaMa model

User preferences, derived from Stage 1 data, included rating-based and physiological features. These were standardized using a StandardScaler and combined with lyrical features for clustering. Agglomerative Clustering with Ward linkage grouped user preferences based on emotional and thematic alignment. Dimensionality reduction using PCA aided visualization and feature importance analysis.

4. Methods

Our analysis comprised of two key components. First, we applied classification algorithms and ensemble techniques to predict mood changes after music listening, using situational and emotional features to model users' psychological satisfaction. Second, we utilized a fine-tuned LLaMA model to analyze song lyrics, extracting features such as narrative complexity, emotional sophistication, and thematic elements. These lyrical features, combined with user preferences learned from interaction data, were used in hierarchical clustering to group user preferences by semantic alignment, allowing for recommendations of songs with similar features.

4.1. Classification Algorithms for Predicting Mood Changes

Our primary classification task aimed to predict if users' emotions will lift, drop, or remain the same based on user ID, music metadata, and situational features. Specifically, we predict changes in valence, where positive valence generally indicates psychological satisfaction. The threshold for valence change is ±0.125, as it represents a perceptible difference based on SiTunes' annotations. We use three feature combinations to represent situational information in the classification task: Objective (Obj.), Subjective (Sub.), and a combined set of Objective + Subjective (Obj. + Sub.) features. These combinations allow the model to incorporate both situational context and user emotional states.

In the first setting, we use Stage 2 data (a field study with 30 users collecting situational context and feedback before and after music listening) as both the training and test set to assess model performance in a traditional recommendation environment, with and without situational information. In the second setting, we train on Stage 2 and test on Stage 3, which involves 10 users in a situation-aware recommendation setting, evaluating model generalization across traditional (Stage 2) and context-adaptive (Stage 3) environments. Each experiment is repeated 10 times, with average results reported.

4.2. Baseline Models and Ensemble Learning Strategies

To predict mood changes, we first implement four baseline classifiers: Logistic Regression (LR), Multi-Layer Perceptron (MLP), Random Forest (RF), and Gradient Boosted Decision Trees (GBDT). These models serve as standalone predictors and are evaluated individually on accuracy (Acc.), Macro F1 score, and Micro F1 score metrics. Each model uses a concatenation of user ID, item metadata, and situational features as input.

Given the complementary strengths of these classifiers, we also construct an ensemble model to try to improve classification accuracy. We employ a stacking ensemble strategy to combine the outputs of multiple base classifiers, aiming to capture more complex patterns and reduce individual model biases.

In the **Stacking Ensemble** [7], predictions from each base model (RF, GBDT, and MLP) serve as input features for a final meta-classifier, a Logistic Regression model. The meta-classifier learns to combine the outputs from each base model, capturing the relative strengths of each classifier and mitigating individual weaknesses. Formally, let $P_{RF}(x), P_{GBDT}(x)$, and $P_{MLP}(x)$ represent the predicted probabilities from RF, GBDT, and MLP, respectively. The stacked model then combines these as input to the Logistic Regression classifier to produce a final prediction:

$$\hat{y} = LR(P_{RF}(x), P_{GBDT}(x), P_{MLP}(x)).$$

4.3. Lyric Sentiment and Thematic Analysis

To extract semantic lyrical features, we utilize the **unsloth/Llama-3.2-3B-Instruct** model fine-tuned with **LoRA (Low-Rank Adaptation)**. LoRA introduces low-rank matrices into the model's weights, allowing updates to be made by modifying only a small subset of parameters. Specifically, weight updates are expressed as $W' = W + W_{\text{LoRA}}$, where $W_{\text{LoRA}} \in \mathbb{R}^{r \times d}$ and $r \ll d$. This approach maintains the model's performance while significantly reducing computational overhead, making it ideal for analyzing our scraped lyrical data.

Given a song's lyrics L, the fine-tuned model generates a feature vector $F(L) \in \mathbb{R}^n$, capturing our desired emotional and thematic dimensions. These dimensions include *narrative complexity*, which measures the structural sophistication of the lyrics, and *emotional sophistication* as the depth and subtlety of emotional content. The model also assigns scores for *thematic elements*, such as love, life, and social themes, as well as *temporal focus*, quantifying the emphasis on past, present, or future events. Formally, the feature vector is defined as:

$$F(L) = [f_1(L), f_2(L), f_{3:k}(L), f_{k+1:k+3}(L)]^\top,$$

where f_1 through f_{k+3} represent the respective lyrical charactersitics.

The extraction process involved feeding structured prompts into the fine-tuned model, instructing it to analyze the lyrics and assign scores to predefined categories. The model's output responses, $R(L)$, are parsed into numerical features using a custom parsing function. These features are then organized in a dictionary format indexed by song title and artist.

4.4. Song Clustering Using Obj + Sub and Lyric Features

Clustering is used to group users based on their preferences, based on song interactions and computed lyrical features as well as emotional preferences. The primary method used is **Agglomerative Clustering**, a hierarchical approach that progressively merges smaller clusters into larger ones using the **Ward variance minimization criterion**.

Before clustering, user features were extracted and scaled for a uniform representation of preferences. The feature set $X \in \mathbb{R}^{n \times d}$ consists of n users, each represented by d-dimensional vectors. These features include rating-based metrics, emotional features, lyrical preferences. The features are normalized to zero mean and unit variance, so there is equal weight for all dimensions during clustering.

To find the optimal number of clusters, we use a binary search on the *distance threshold*, τ, which controls granularity in clustering. Given a range $[\tau_{\text{min}}, \tau_{\text{max}}]$, the algorithm iteratively adjusts τ to get the desired k number of clusters. Formally, the clustering is:

$$\mathcal{C}(\tau) = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}, \text{ where } \|\mathbf{x}_i - \mathbf{x}_j\| \leq \tau \, \forall \, i, j \in C.$$

The Ward criterion minimizes the increase in within-cluster variance when merging clusters A and B:

$$d(A, B) = \frac{|A||B|}{|A| + |B|} \|\mu_A - \mu_B\|^2,$$

where $|A|$ and $|B|$ are the number of users in clusters A and B, and μ_A, μ_B are their centroids.

5. Experimental Results

We began by evaluating the effectiveness of various classification algorithms in predicting mood changes following music listening. Each model was tuned and tested individually on the SiTunes dataset using different situational feature sets (Obj., Sub., and Obj. + Sub.) to capture the effect of these contexts on prediction accuracy. The primary metrics used to evaluate model performance are accuracy, Macro F1 score, and Micro F1 score, as these give insights into both the overall accuracy and the balance of predictions across different mood classes (lifting, keeping, and dropping). The equations for accuracy and F1 score are given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

5.1. Hyperparameter Tuning and Model Selection

For each classification model (Logistic Regression, Multi-Layer Perceptron, Random Forest, and Gradient Boosted Decision Trees), we used hyperparameters tailored to the situational and emotional data in the SiTunes dataset. Key configurations are outlined below:

- **Logistic Regression (LR)**: We used $L1$-regularization with a regularization parameter $C = 10$, and the model was trained using the 'liblinear' solver. The maximum number of iterations was set to 1000 for convergence.
- **Multi-Layer Perceptron (MLP)**: The MLP model utilized the Adam solver and ReLU activation functions. For Setting 2, we used a learning rate of 0.005 and a batch size of 256, while for Setting 3, we adjusted these to a learning rate of 0.001 and a batch size of 512. The maximum number of iterations was set to 2000 for convergence.
- **Random Forest (RF)**: The Random Forest model was configured with 300 trees for Setting 2 and 200 trees for Setting 3, with a maximum depth of 3 and 5, respectively, balancing complexity with interpretability.
- **Gradient Boosted Decision Trees (GBDT)**: For Setting 2, we used a learning rate of 0.1 and 200 boosting stages, while for Setting 3, we reduced the learning rate to 0.05 with 100 stages. Each tree had a maximum depth of 5 to capture complexity while controlling overfitting.
- **Ensemble Model (Stacking)**: For the ensemble model, we used a stacking approach that combines the outputs of the base models (RF, GBDT, and MLP) with a meta-learner to improve overall prediction accuracy. Each base model was trained independently, and the results were aggregated to compute the final metrics.

5.2. Model Performance on Mood Change Prediction

The table below summarizes preliminary experiments that quantify performance of each model, evaluated on the test set using the Obj., Sub., and Obj. + Sub. feature sets. Across all models, the Obj. + Sub. feature combination consistently yielded the highest accuracy and F1 scores, indicating that combining objective situational features with subjective emotional states does indeed improve prediction. However, the results were more varied in Setting 2, which is likley due to training on different data distributions.

Model	Context	Setting 1 (Test Set)			Setting 2 (Test Set)		
		Acc.	Macro F1	Micro F1	Acc.	Macro F1	Micro F1
LR	ALL	**0.5919**	**0.4593**	**0.5919**	0.5759	0.4060	0.5759
	SUB	0.5775	0.4110	0.5775	**0.5775**	**0.4110**	**0.5775**
	OBJ	0.5349	0.3727	0.5349	0.5132	0.2999	0.5132
GBDT	ALL	**0.8777**	**0.8646**	**0.8777**	**0.8044**	**0.7710**	**0.8044**
	SUB	0.8292	0.8078	0.8292	0.7964	0.7629	0.7964
	OBJ	0.8063	0.7707	0.8063	0.6562	0.5495	0.6562
MLP	ALL	**0.5080**	0.2308	**0.5080**	**0.5072**	0.2243	**0.5072**
	SUB	0.4827	0.2504	0.4827	0.4785	**0.2463**	0.4785
	OBJ	0.4985	**0.2713**	0.4985	0.5000	0.2359	0.5000
RF	ALL	**0.6511**	**0.5216**	**0.6511**	0.7301	0.6565	0.7301
	SUB	0.6401	0.5173	0.6401	**0.7332**	**0.6646**	**0.7332**
	OBJ	0.5840	0.4085	0.5840	0.6564	0.5500	0.6564
Ensemble	ALL	**0.7565**	**0.5855**	**0.7565**	**0.7045**	**0.5293**	**0.7045**
	SUB	0.6767	0.5064	0.6767	0.6791	0.5039	0.6791
	OBJ	0.5076	0.2311	0.5076	0.5108	0.2571	0.5108

The results indicate that GBDT achieved the highest performance across all metrics, with an accuracy of 87.8% and a Macro F1 score of 0.8646 when using the Obj. + Sub. feature set. Notably, the Ensemble model also performed well, achieving with an accuracy of 75.7% and a Macro F1 score of 0.5855, though not surpassing results from an individual classifier.

5.3. Experimental Results for Sentiment Analysis and Clustering

The fine-tuning process used LoRA to adapt the LLaMa model for lyrical sentiment analysis. The figure below shows the validation loss with an approaching convergence over approximately 400 steps with 10 epochs. Early overfitting was addressed by adjusting the learning rate to $1e-4$ and using 0.01 dropout.

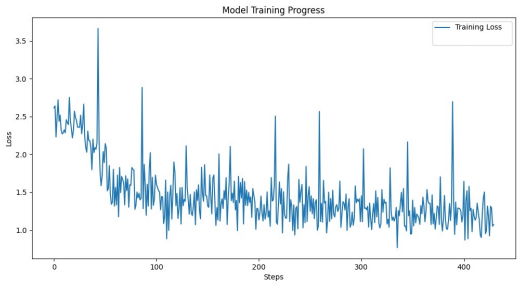

Figure 2. Validation loss of fine-tuned LLaMa model

5.4. Feature Extraction and Analysis

Using the fine-tuned LLaMa model, we extracted eight numerical features from song lyrics and the correlation matrix shown below highlights strong positive correlations, such as between narrative complexity and emotional sophistication ($r = 0.96$), indicating that these features frequently co-occur in lyrically complex songs. On the other hand, temporal focus on the present shows a negative correlation with both narrative complexity and emotional sophistication, suggesting that simpler themes often lack deep narrative structures, which objectively makes sense.

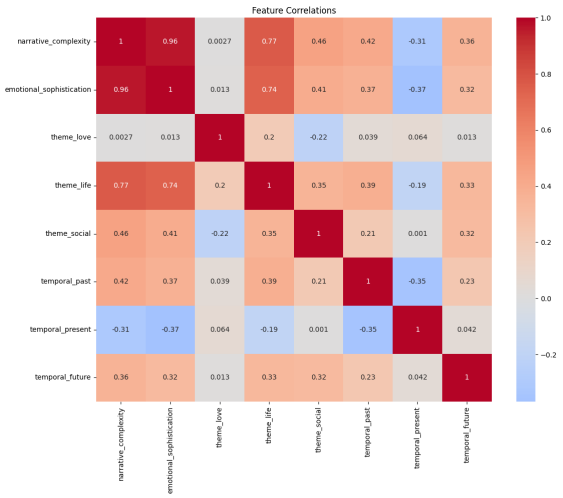

Figure 3. Lyric feature correlations in analyzed songs

The distribution of features, shown in Figure 1, reveals considerable variance across thematic and temporal elements, with ”love” and ”life” emerging as the most prominent themes. Temporal focus is skewed toward past events, reflecting the common introspective nature of lyrical content.

5.5. Clustering

The clustering process identified four distinct user groups based on emotional responses and lyrical preferences, as visualized in a 2D PCA plot capturing 71.9% of the total variance in the first two components. There's clear separation among some clusters like in clusters 3 and 1, which highlights meaningful differences in user behaviors and preferences. However, some clusters like 0 and 2 have less defined separations and so these groupings show more similar and less differentiable emotional and lyrical dimensions. A computed dendrogram further validated these clusters (not shown), and below we show the computed user clusters based on user preferences and lyrical alignment of those preferences.

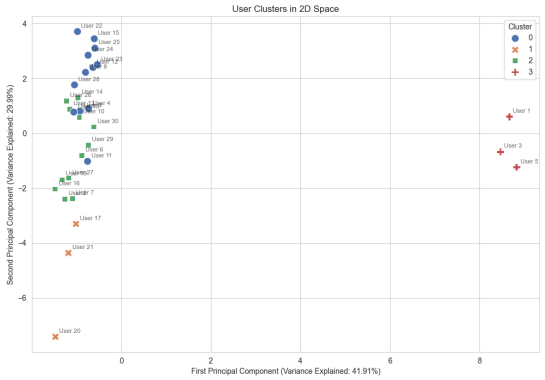

Figure 4. User clusters based on preferences, extracted lyrical features, and song metadata

The cluster profiles heatmap in Figure [5] provided additional insights into these groupings. Cluster 0 showed high emotional variability and a preference for songs with complex narratives and emotional sophistication. Cluster 1 had a more balanced feature profile, learning towards socially themed songs. Cluster 2 favored highly rated tracks, while Cluster 3 had a greater valence range and slightly higher emotional sophistication. Clusters showcase similar values across these categories; however, the subtle distinctions within each cluster showcase the complexity of capturing nuanced patterns in the data and variations in user preferences.

5.6. Recommendations

The recommendations for 30 users, based on cluster assignments, showed strong alignment with individual preferences, as indicated by high similarity scores ($\geq$ 0.95). To preserve space and as an example of a high-performing output, we show a table of recommendations for user 6, who is contained within cluster 2.

Cluster: 2		
Lyrical Preferences		
Preference	Value	Std. Dev.
Narrative Complexity	0.61	±0.26
Emotional Sophistication	0.77	±0.27
Thematic Preferences		
Theme	Value	Std. Dev.
Love	0.60	±0.43
Life	0.49	±0.35
Social	0.21	±0.21
Temporal Focus		
Timeframe	Value	Std. Dev.
Past	0.66	±0.26
Present	0.71	±0.28
Future	0.39	±0.31
Top Recommendations		
Rank	Song (Artist)	Similarity (Narrative, Emotional)
1	Alabaster (Foals)	1.00 (0.80, 0.90)
2	Battle (Colbie Caillat)	1.00 (0.80, 0.90)
3	Whereabouts Unknown (Rise Against)	1.00 (0.80, 0.90)
4	Dance Away (Roxy Music)	1.00 (0.80, 0.90)
5	Thinking Of You (Katy Perry)	0.99 (0.80, 0.90)

We see a a strong preference for emotional sophistication and a moderate inclination toward narrative complexity. Thematic preferences rank highest for love followed by life, and lowest for social themes. Temporal focus emphasizes the present and past, with less focus in the future. Top recommendations, such as *Alabaster* and *Battle*, achieve perfect similarity scores (1.00) and we're also checked manually, showing strong alignment with lyrical and thematic preferences.

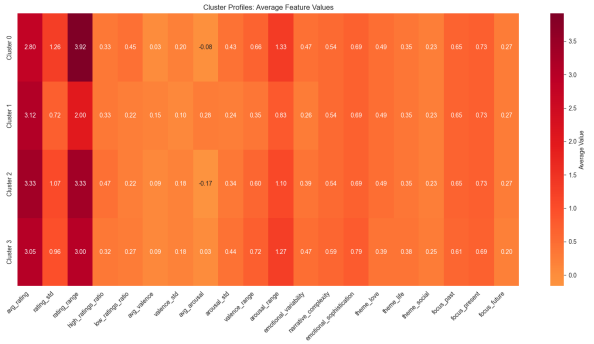

Figure 5. Average feature values for computed cluster profiles

6. Conclusion and Future Work

This project combined situational features, lyrical sentiment analysis, and user preference clustering to predict mood changes and generate personalized music recommendations. Gradient Boosted Decision Trees (GBDT) achieved 87.8% accuracy and a Macro F1 score of 0.8646 by leveraging objective and subjective features from the SiTunes dataset. Hierarchical clustering grouped songs into cohesive clusters based on lyrical and situational features, giving high-similarity recommendations closely aligned with user preferences. However, repeated songs across clusters highlighted the need for greater recommendation diversity, which we didn't discuss much.

Integrating diversity-aware ranking methods could reduce repetition and boost engagement, while advanced models like transformers trained on multimodal data could better capture complex relationships between the many complex features. More extensive data, including temporal user feedback, is needed for a more cohesive recommendation system and so we only built components of one. However, incorporating physiological and psychological data showed to significantly improve personalization, making the system more dynamic, accurate, and applicable in real-world settings.

References

[1] Adiyansjah, Alexander A S Gunawan, and Derwin Suhartono. Music recommender system based on genre using convolutional recurrent neural networks. *Procedia Computer Science*, 157:99–109, 2019. The 4th International Conference on Computer Science and Computational Intelligence (ICCSCI 2019) : Enabling Collaboration to Escalate Impact of Research Results for Society.

[2] Imen Ben Sassi and Sadok Ben Yahia. How does context influence music preferences: a user-based study of the effects of contextual information on users' preferred music. *Multimedia Syst.*, 27(2):143–160, Apr. 2021.

[3] Brian Brost, Rishabh Mehrotra, and Tristan Jehan. The music streaming sessions dataset. In *The World Wide Web Conference*, WWW '19, page 2594–2600, New York, NY, USA, 2019. Association for Computing Machinery.

[4] Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. The yahoo! music dataset and kdd-cup'11. In Gideon Dror, Yehuda Koren, and Markus Weimer, editors, *Proceedings of KDD Cup 2011*, volume 18 of *Proceedings of Machine Learning Research*, pages 3–18. PMLR, 21 Aug 2012.

[5] Vadim Grigorev, Jiayu Li, Weizhi Ma, Zhiyu He, Min Zhang, Yiqun Liu, Ming Yan, and Ji Zhang. Situnes: A situational music recommendation dataset with physiological and psychological signals. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, page 417–421, New York, NY, USA, 2024. Association for Computing Machinery.

[6] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16. International World Wide Web Conferences Steering Committee, Apr. 2016.

[7] Bohdan Pavlyshenko. Using stacking approaches for machine learning models. In *2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)*, pages 255–258, 2018.

[8] Igor André Pegoraro Santana and Marcos Aurelio Domingues. A systematic review on context-aware recommender systems using deep learning and embeddings, 2020.

[9] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. Lfm-2b: A dataset of enriched music listening events for recommender systems research and fairness analysis. CHIIR '22, page 337–341, New York, NY, USA, 2022. Association for Computing Machinery.

[10] Xiao Zhang, Sunhao Dai, Jun Xu, Zhenhua Dong, Quanyu Dai, and Ji-Rong Wen. Counteracting user attention bias in music streaming recommendation via reward modification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2504–2514, New York, NY, USA, 2022. Association for Computing Machinery.