Leveraging children's music preferences to enhance the recommendation process

Msc Thesis Computer Science

Ilias Papadimitriou

TU Delft

Leveraging children's music preferences to enhance the recommendation process

Msc Thesis Computer Science

Thesis report

by

Ilias Papadimitriou

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on TO Update!

Thesis committee:

Chair:	Sole Pera
Supervisor:	Sole Pera
Daily Supervisor:	Sole Pera
External examiner:	Maliheh Izadi
Place:	Faculty of Electrical Engineering, Mathematics, Computer Science, Delft
Project Duration:	From May 2024 - To December 2024
Student number:	5847079

Faculty of EEMCS · Delft University of Technology

Delft University of Technology

Abstract

Children spend a significant amount of time listening to music. Music has a significant cognitive and developmental effect on them. Because of their unique behavioral characteristics and their emotion regulation skills, their music preferences differ significantly from adults. However, limited study has been conducted on their music preferences and no music recommendation system have been designed to cater for their specific needs and preferences. With this study, we conduct an empirical exploration of the music preferences of children in terms of audio and sentiment characteristics, readability and topics discussed in the songs listened by children at different ages. We utilize the outcomes of our empirical exploration to adapt a recommender system to cater for the music preferences of children. Both the empirical exploration and the evaluation of the recommender system is based on the well-known extensive music dataset LastFM-2b. Outcomes from this work showcase that grade school students prefer more positive and joyful sentiments expressed in the lyrics, simpler language and higher acousticness. Older children prefer songs that convey sadness or anger, higher language complexity and songs they can dance to. Incorporating the song features into a recommender system leads to an increase in terms of all evaluation metrics in children compared to the RS trained only on user-item interaction data.

Preface

Contents

List of Figures

List of Tables

Part I

Introduction and Contextual Overview

1 Introduction

Children spend a lot of time listening to music, a habit that has a significant cognitive and developmental impact on them [1]. Music experiences, besides being a source of enjoyment for them, have also been proven to enhance the perception of language, mathematics proficiency and intellectual development in children [2]. Especially when children systematically engage with musical instruments, significant changes have been observed in their brain images and cognitive behaviour [3]. However, children have some cognitive and emotional characteristics that separate them from adults, such as their vocabulary understanding and emotion regulation skills, which also form their unique listening preferences [2].

The individual music preferences should not be neglected when designing a Recommender System (RS) for children. Childhood is an age where children go through different developmental stages, where their social relationships, competencies and interests change [4]. Therefore, they exhibit different behaviour and characteristics at each stage of their development.

Specifically regarding their music preferences, previous research shows that children in grade school (GS), middle school (MS) and high school (HS) have significant differences in audio characteristics of the songs they prefer [1], as well as the genre they prefer [5]. However, children's music preferences regarding lyric content have not been investigated. The lyrics include valuable insights regarding the readability, the sentiment and the thematic elements of the tracks. Analyzing these aspects through the lens of age can reveal crucial information about the preferences of children in different developmental stages. Additionally, no music RSs have been specifically developed for children, utilizing the unique music preferences of children at different ages [1]. Previous research shows that children at different ages react differently to different music characteristics [6]. Therefore, we conduct the empirical exploration based on this research objective:

Research Objective

Explore in depth the audio and lyric features that children of different educational levels prefer

By identifying the key discriminative music features between children of different ages, the similarity between songs can be captured even with limited user-interaction data. Therefore we aim to integrate the discriminative features into an RS, in a way that it enhances the recommendation process for children.

Therefore we will employ an RS, aiming for this objective:

Research Objective

Utilize the discriminative features of an RS to deliver music suggestions specifically tailored to meet the preferences and needs of children.

With our work, we address the following research questions:

Research Question 1

Which song lyric and audio features help capture the music preferences of children across different educational levels?

Research Question 2

How can the song features be leveraged to enhance the recommendation process?

Similar to previous research [1], users are categorized by educational levels, due to the limited sample size among grade school students. The educational levels we examine are grade school for children between 6 and 11 years old, middle school for children between 12 and 14 and high school for children between 15 and 17 years old. We investigate the characteristics of songs that are preferred by children of different educational levels. The lyrics will be analyzed regarding their sentiment, readability and prevalent words. Specifically, the distinct Listening Events (LEs) corresponding to children of each educational level are captured and analyzed for each song characteristic.

The analysis reveals key track features that facilitate the identification of differences across age groups of children, or between children and adults. These features are employed for the recommendation process, along with the user-interaction data. Factorization Machines (FMs) are employed for the recommendation process.

The empirical exploration and the evaluation of the RS is conducted on the LastFM-2b dataset, which includes 2 billion Listening Events (LEs). For each song, the lyrics and the audio features are extracted for our analysis. The song lyrics are extracted using the Genius API and are used to calculate the readability and the sentiment of each song, as well as the prevalent words by children at different educational levels . The audio features are gathered using Spotify API [7]. Spotify API offers multiple features that can be extracted for each track, including acousticness, danceability or tempo. The diverse range of features that are gathered for each song enables us to capture sufficient details about a song, making it possible to recommend songs with similar audio or lyric characteristics, even if they belong to different genres or were performed by different artists.

The findings of the study indicate that there are significant differences in the audio and the lyric characteristics children prefer at different stages of their development. Utilizing the features for which a significant difference is observed leads to an enhancement of some performance metrics for the overall children population and a significant performance in all evaluation metrics for grade school students.

The key contributions of this study are outlined as follows.

- We explore the music characteristics children prefer at different educational levels. Specifically, the audio, sentiment and readability characteristics are analyzed as well as the prevalent words in the songs listened by children of different age groups are analyzed.
- We propose FMkids, an adaptation of FMs including the song attributes that capture differences between children at different educational levels.
- We identify the age groups of children for which FMkids can enhance the recommendation process and the comparative advantages of FMkids compared to an RS that relies only on user-item interaction data.

2 Background and Related Work

This section reviews the background our work builds upon and discusses key studies on the music preferences of children and music RS for children. Specifically, we introduce the song characteristics that we leverage for the empirical exploration of the music preferences of children and the RS research we build upon. Then we explore previous studies, identify research gaps and build a framework for the systematic analysis of the music preferences of children. We analyze existing music RSs and evaluate their limitations in addressing the unique needs of child users. Finally, we provide a detailed overview of the lyric characteristics that form the foundation of our subsequent analyses.

2.1. Background

This section provides an overview of the foundational background required for analyzing the characteristics of song lyrics. Additionally, it explores the current state-of-the-art categories of RSs, offering an understanding of their methodologies and applications.

2.1.1. Sentiment Analysis

Sentiment Analysis (SA) is employed to identify and categorize the sentiments expressed in a piece of text. It has been used in various domains to extract people's opinions, thoughts, and impressions regarding a topic.

Sentiment Analysis can be performed with the following methods:

- **Lexicon-based**: It is a knowledge-based approach where the sentiment of each word is predefined in a lexicon and the sentiment of a document is the average or weighted average of the words the document contains [8]. They are intuitive and simple to interpret and do not require any training.
- **Machine Learning-based**: Machine learning classifiers are utilized to identify the sentiment of a document [9]. Support Vector Machines and Naive Bayes are among the most frequently utilized machine learning algorithms for this type of classification. They train on labeled training data with sentiment annotations.
- **Deep Learning-based**: The textual data is preprocessed and then encoded using pre-trained embeddings such as GloVe and word2vec. These embeddings are then fed into deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for classification [9]. Unlike the other cases, the context is also assessed.

SA can be performed through capturing the polarity of the content, or by capturing the basic and prototypical emotions. Basic and prototypical emotions are the following: sadness, anger, fear, trust, disgust, surprise, and anticipation, as suggested by Plutchick in 1980 and was later widely used for SA [10]. In the case of lexicon-based SA, the NRC Word-Emotion Association Lexicon is the largest used lexicon for Sentiment Analysis that provides an association between words and the primary emotions [11]. NRC also offers a hashtag sentiment lexicon that captures the positivity and the negativity of the text.

2.1.2. Readability

Readability is a metric that expresses the expected difficulty for a reader based on a text [12]. Multiple metrics have been introduced to assess readability. The most popular metrics are Flesch-Kincaid Reading

Ease and Flesch Kincaid Grade Level [13], the Dale-Chall Readability Formula [14] and the Spache Readability Formula [15]. Specifically, Flesch Kincaid grade level provides an estimation of the class a child is following based on the number of words, number of sentences, and number of syllables [13]. It allows to investigate whether the readability of the text aligns with the readability skills of the people that read it or listen to it. The estimation of the readability grade level is also possible with Spache readability metric [15]. Spache readability is calculated based on the average sentence length and the percentage of words that are included in the Spache difficult words dictionary. For older students, the Dale-Chall is most appropriate for calculating readability [14]. This formula is based on the number of words, the number of sentences and the percentage of difficult words. A word is considered difficult if it does not appear in a list of common words for which it was found that more than 80% of fourth graders are familiar.

2.1.3. Recommender System

Recommender systems function by analyzing user, item, or user-item interaction data to predict preferences and suggest items that users are likely to enjoy. There are six general recommendation approaches: content-based (CB), collaborative-filtering (CF), community-based, demographic, knowledge-based, and hybrid recommenders [16]. Each of them is briefly explained:

- **Content-Based**: recommends items that are similar to the ones that the user liked in the past. The similarity of items is calculated based on the features associated with the compared items.
- **Collaborative Filtering**: recommends the items that other users with similar tastes liked in the past. The similarity in taste of two users is calculated based on the similarity in the rating history of the users.
- **Demographic**: recommends an item to a user based on the demographic profile of the user
- **Knowledge-based**: uses a similarity function between the user and the item that is based on specific domain knowledge. This similarity function is also the utility of the recommendation to the user.
- **Community-based**: This type of system recommends items based on the preferences of the user's friends. This is useful in the cases of social networks and performs better than recommending based on similar unknown users [17]
- **Hybrid**: These RSs are based on the combination of any of the above mentioned techniques.

2.2. Related Work

Music plays an important role in the lives of children. They spend 20% of their time listening to music and consider it more important compared to adults [18]. Besides being a source of enjoyment, music serves also as a way for children to develop a bond with their peers [18]. Listening to music offers significant benefits for children, such as enhancing their cognitive abilities [2]. Understanding children's music listening preferences is essential because it can help parents, educators, and policymakers tailor experiences that maximize these benefits, ensuring music serves as a tool for their social, emotional, and intellectual growth. Therefore, investigating these preferences can be a valuable tool to leverage music as a resource for children's development.

2.2.1. Children's music preferences

Importantly, children's music preferences evolve with age [1], highlighting the necessity of considering developmental stages when examining the distinct patterns of their music listening behaviors. Previous research has examined the music preferences of children at different ages and has drawn valuable conclusions on the music preferences of children at different ages [19, 1, 20, 21, 22]. Children younger than 4 years old prefer fast and loud music regardless of gender [20]. Children until the age of 11 tend to enjoy a wide variety of different songs, exhibiting what is called open-earedness, which declines when children enter adolescence [21]. Significant differences have been observed in the audio characteristics preferred by children at different ages. Older children and adolescents tend to prefer more complex musical structures, compared to younger ones who prefer simpler melodies [22].

The existing body of work provides a strong foundation for understanding how children's music preferences vary by age and developmental stage. However, a notable gap remains in systematically linking these developmental patterns with specific quantifiable music characteristics that depict those music preferences. To the best of our knowledge, Spear et.al. [1] is the only study that described the music preferences of children using quantifiable characteristics and compared the LEs of children at different ages. However, this study only analyzes a limited number of LEs corresponding to children and explores the songs only based on their audio characteristics. Spear et.al. [1] utilizes the LastFM-1b dataset, which includes 1 billion LEs, but children are underrepresented, as only 3416 of the users are children. Furthermore, focusing solely on the audio characteristics of songs restricts the scope of the study to the sound characteristics of the music. However, the lyrics also contain important information about the song. Lyrics enable visual imagery and contagion [23]. The words of the song capture the theme and the sentiment that is expressed by the artist, as well as the audience it targets. To the best of our knowledge, there is no analysis on the lyric characteristics of the songs children prefer.

2.2.2. Music Recommender Systems

Extensive research has been conducted in the field of RSs to provide relevant recommendations. The majority of the studies involve content-based, collaborative filtering or hybrid methods. Barragans-Martinez et.al. performed user-based and item-based collaborative filtering for TV program recommendation [24]. However, collaborative filtering is vulnerable to the sparsity of a dataset and the gray sheep problem: people with unique preferences and tastes make it difficult to develop relevant recommendations [25]. Sanchez-Moreno et.al. [26] has combined collaborative filtering with user attributes that express how unusual are the preferences of a user and it has shown to perform well in the music domain.

Collaborative filtering methods perform poorly with items lacking historical data and content-based recommenders are less optimized for lacking similarity [27], therefore robustness is often achieved with hybrid RSs. The hybrid RS suggested by Val et.al. [28] outperforms collaborative filtering methods in automated playlist continuation. Wang et.al. combine their hybrid recommendation model with a deep learning model that converts the handcrafted audio attributes to features that are optimized for an RS [29]. However, the state-of-the-art music RSs typically suggest items without accounting for the distinct characteristics of different age groups. Notably, children exhibit behaviors and preferences that differ significantly from those of adults [18]. Consequently, a detailed examination of children's unique traits is essential to develop RS that effectively address their specific needs.

Part II

Methodology and Findings

3
Exploring Music Preferences of Children

In this chapter, we explore the listening behaviors of children, analyzing how musical preferences shift as they move through different ages. By analyzing a large dataset of listening events enriched with various descriptive features of the songs, we create profiles that characterize each age group's unique musical landscape. These profiles offer insights into how song characteristics resonate differently with children as they grow, revealing age-related shifts in musical engagement. The chapter is divided into two sections: the first outlines the methodology and details of the experimental setup, and the second one delves into the results and the key takeaways of this empirical exploration.

3.1. Methodology and Experimental Setup

In this section, we detail the methodology employed in our experiments. Specifically, we describe the dataset utilized, including the audio and lyric features enriched to enhance its relevance. Furthermore, we discuss the rationale for selecting these particular characteristics and outline the experimental approach designed to investigate children's music preferences.

3.1.1. Dataset

To explore music listening behavior and song characteristics among young listeners, we use LastFM-2b [30] dataset. This dataset is a comprehensive collection of 2 billion Listening Events (LEs) that were collected from February 2005 until March 2020 by 120.000 users and 50 million distinct tracks. We chose a minimum threshold of 5 interactions per user and track to ensure a meaningful level of engagement. The dataset also includes the age of the users, based on which, we only keep the LEs that correspond to children between the ages 6-17. Table 3.1 provides an overview of the number of users, songs and distinct LEs per educational level that are available in the filtered dataset.

To investigate the music characteristics of songs, we enrich the dataset with track features. The lyrics are extracted using Genius API [31] and then the sentiment, readability and vocabulary characteristics of the lyrics are computed. The audio features of the songs are extracted using the Spotify API [7]. The original format of the LastFM-2b dataset is <user-id, artist-id, track-id, album-id, timestamp>, but our enriched dataset also includes all track features that correspond to the track-id of the LE. In some cases, the song lyrics or the audio features were not available, or the readability could not be computed because of a text of less than 100 words, therefore the distinct LEs for which all song features are available are considerably smaller.

	Education Level		
	GS	MS	HS
Users	55	277	3,018
Songs	141,760	202.325	946,449
Distinct LEs	184,162	353,263	4,938,132

Table 3.1: LastFM-2b overview per educational level

3.1.2. Features

To effectively capture and understand users' music preferences, it is essential to describe songs through quantifiable metrics that highlight both the similarities and unique characteristics of each piece. These metrics can be divided into audio features and lyrical features, each providing distinct insights into the song's impact and appeal.

Audio Features

Analyzing audio features provides a comprehensive way to capture and interpret the sound of music. Audio features allow to capture the sound characteristics of a song, such as the tempo, whether or not it contains musical instruments or its appropriateness for dancing. Each feature offers a unique lens, contributing to a holistic understanding of the track's structure, mood, and setting it is appropriate for. The audio characteristics offer an objective way to explore music listening preferences, compared to conventional categories of music genres that are open to subjective interpretation [32]. The audio attributes that will be analyzed are the ones provided by Spotify API [7]. They cover a wide range of valuable information from the presence and the kind of instruments to the volume and the recording setting. Specifically, the following features will be explored:

- key is an integer from 0 to 11, each representing a different musical key starting from C note to B note. This feature is fundamental for understanding the harmonic foundation of a track, as the key sets the tonal center and influences the mood and emotional response elicited by the music.
- acousticness (range [0,1]) describes whether the track is acoustic. Higher values correspond to live non-synthesized instruments predominate, whereas electronic or heavily produced tracks typically have lower values. This feature is useful for differentiating between different production styles.
- danceability (range [0,1]) describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.
- energy (range [0,1]) represents a perceptual measure of intensity and activity.
- instrumentalness (range [0,1]) expresses the likelihood of the song containing no vocals. The closer the instrumentalness value is to 1.0, the greater the likelihood the track contains no vocal content.
- loudness (range [-60,0]) is counted in dB and is averaged across a full track, expressing how loud a track is.
- speechiness (range [0,1]) detects the presence of spoken words in a track. The more exclusively speech-like the recording, the closer to 1.0 the attribute value.
- liveness (range ([0,1]) expresses the probability of the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
- mode (range([0,1]) indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.

Sentiment Analysis

Apart from the audio characteristics of a song, valuable information is also present in the sentiment of the lyrics. Children's understanding of emotion changes while they grow up [33]. Younger children might gravitate towards simple, positive sentiments that evoke happiness and safety, while older children could appreciate more complex emotional narratives, as children develop their emotional understanding over time [34]. Therefore, sentiment analysis of song lyrics is utilized to determine the emotional tone conveyed in a song. The sentiment of a song has also been a valuable predictor of the mood of the listener [35] and thus allows the investigation of the child's feelings while listening to a song. Apart from the polarity, the lyrics are analyzed based on the 8 basic emotions: sadness, anger, fear, trust, disgust, surprise, joy and anticipation[36]. The NRC Word-Emotion Association Lexicon and the NRC hashtag sentiment lexicon are utilized for this purpose [11]. The lexicons provide a score for the sentiments that correspond to a word and the sentiment of the song is calculated as the average of the emotions of all words in the song. Then we capture the distribution of each sentiment for each educational level, by including the sentiment values that correspond to the songs of the distinct LEs. **Readability**

The readability of the lyrics is also assessed to understand how complex vocabulary influences children's preferences. Older children outperform younger ones in reading skills [37] and therefore might prefer more sophisticated language in the songs their prefer. In other activities like research participation, it is important that the information children receive information according to their capacity of understanding [38]. However, to the best of our knowledge, the alignment of their readability skills to the complexity of the music content they prefer has not been investigated. The readability is analyzed using the Flesch-Kincaid Grade Level [13], a straightforward method that expresses the readability of a text based on the number of words, number of sentences, and number of syllables. The following formula is used :

$$Flesch - Kincaid = 0.39\frac{words}{sentences} + 11.8\frac{syllabes}{words} - 15.59 \tag{3.1}$$

Prevalent Words All the features we have examined so far depict a song in terms of sound, sentiment and readability, but do not consider the exact topics discussed in a song. By examining the most prevalent words in the songs, we aim to identify the dominant themes and subjects addressed in these lyrics. This examination will provide insights into the topics and narratives that resonate with or potentially influence young listeners. Additionally, by examining the main themes addressed in a song, we can assess whether the lyrical content aligns with age-appropriate standards, evaluating its suitability and relevance for children's developmental stages and sensibilities.

3.1.3. Experimental Setup

To examine similarities and differences in children's music preferences, an appropriate lens for analyzing the data is required. Using age as a lens helps us draw meaningful conclusions about music preferences of children at different stages of their childhood and adolescence [39]. The LastFM-2b dataset includes only 55 users between the ages 6-11, so grouping by exact age number would not provide trustworthy results. Therefore, we follow previous research on children's music preferences [1] and group the users in educational levels: Grade School students (6-11 years old), Middle School students (12-14 years old) and High School students (15-17 years old). By categorizing listeners based on age groups, this study also explores patterns in audio and lyric features that uniquely resonate with children at various developmental stages.

For each of these educational levels, we collect and analyze the distinct LEs corresponding to its users. This means that each user u_i of a certain educational level listening to a song i_j is considered a single sample regardless of the times the song was listened, but multiple users $u_1, u_2, ..., un$ listen to the song i_j, are considered n samples. This design choice was implemented to appropriately weigh songs listened to by multiple users while minimizing the bias introduced by individual users repeatedly listening to the same track. For each educational level, we collect the distribution of distinct LEs made by the users of the corresponding age range and compute all song features for each sample of the distribution. For each of the numeric features, the means per educational level are captured and compared, to make meaningful conclusions about how music listening preferences differ across age groups. Finally, to compare feature distributions across educational levels, we conduct a one-way ANOVA for each numeric feature. Statistically significant results are identified with p $\leq$ 0.05.

For the prevalent words which cannot be expressed as a numeric feature, a different setup is followed. The distinct LEs are grouped by the educational level of the user. The TF-IDF algorithm is employed between the distinct LEs of children at the same educational level, with the song corresponding to the LE being the document. The TF-IDF is chosen because of its usefulness in identifying the most prevalent words in a song, but penalizing words that are popular among all songs in the collection. Then the 5 highest-ranked words of each song are chosen as representative of the song. The words with the highest frequency among the representative words of each song are chosen as representative of the listening events of a certain educational level. As many of the prevalent words are interjections such as 'oh', 'la', 'ah', that do not offer valuable input in capturing the topics discussed in a song, those are removed from the representative words. Then the distribution of prevalent words in each educational level is calculated, where the value of each word is the amount of times it is included in the representative words of a Listening Event divided by the the total number of representative words. A visualization of this procedure is shown in Figure 3.1.

3.2. Results and Analysis

In this section, we present the results of our empirical exploration and explain the key takeaways that derive from our work. The first part analyzes the results on the music preferences of children at different

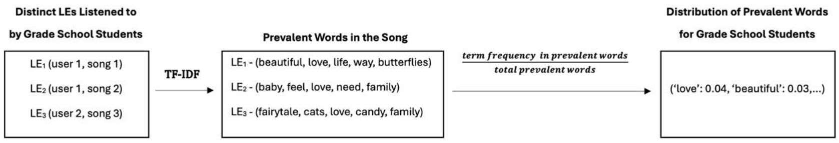

Distinct LEs Listened to by Grade School Students
LE₁ (user 1, song 1)
LE₂ (user 1, song 2)
LE₃ (user 2, song 3)
TF-IDF
Prevalent Words in the Song
LE₁ - (beautiful, love, life, way, butterflies)
LE₂ - (baby, feel, love, need, family)
LE₃ - (fairytale, cats, love, candy, family)
term frequency in prevalent words / total prevalent words
Distribution of Prevalent Words for Grade School Students
('love': 0.04, 'beautiful': 0.03,...)

Figure 3.1: Visualization of the methodology to calculate the distribution of prevalent words in an educational level

ages and the second part explains the conclusions drawn from this analysis and how those inform us for the utilization of the RS.

3.2.1. Results

This section presents a comparative analysis of the audio and lyric features of songs preferred by users of different educational levels: grade, middle, and high school. Each feature category —audio and lyric— is analyzed starting with normality tests to assess the distribution of the data. Subsequently, profiles of the audio and lyric characteristics are developed for each educational group, highlighting key differences and similarities. This approach provides insights into how educational background influences musical preferences, focusing on variations in audio characteristics, lyrical themes, sentiment analysis as well as readability.

Audio Features

To determine whether the music listening traits among children from grade school, middle school, and high school follow a normal distribution, we conducted normality tests using the Shapiro-Wilk test. The test was applied to the distributions of distinct listening events at each educational level.

The results of the Shapiro-Wilk are presented in Table 3.2. The test indicated that the null hypothesis of normality was rejected for all educational levels and variables, suggesting that the distribution of music listening traits in this group significantly deviates from normality. However, even small deviations can cause low significance, when the sample size is very large, as in the case of LastFM LE distributions. The value W is in all cases above 0.8, which indicates that the data is close to being normally distributed. This, combined with the robustness of ANOVA in handling non-normally distributed data when the sample size is sufficiently large, supports the use of one-way ANOVA to compare the distributions of the variables between different educational levels.

Table 3.2: Shapiro-Wilk Normality Test Results of Audio Features

Education Level	Statistic	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness
Grade School	p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	W	0.99	0.93	0.93	0.88	0.60	0.57	0.77	0.68	0.77
Middle School	p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	W	0.99	0.92	0.93	0.85	0.62	0.63	0.71	0.61	0.77
High School	p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	W	1.00	0.91	0.90	0.86	0.66	0.63	0.71	0.62	0.79

To gain an initial understanding of how audio characteristics vary between the distinct LEs experienced by children across different educational levels, we perform a comparative analysis of the mean values of each audio feature within these educational groups, shown in 3.3. By comparing the means of the distributions for each audio characteristic, we aim to identify potential differences or trends in the audio features associated with various educational levels. This approach provides a foundational overview of how the auditory properties of listening events may shift based on educational context. The first noticeable trend observed across most audio features is that the songs listened to by middle school and high school students tend to exhibit similar characteristics, while showing significant differences compared to those listened to by grade school students. Specifically, energy, acousticness and instrumentalness means showcase a difference larger than 0.04, where the range of those variables is [0,1]. We expect this difference to be validated in the pairwise one-way ANOVA comparison between grade and middle school as well as grade and high school.

Table 3.3: Audio Characteristics Means Comparison

Educational Level	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness
Grade School	0.498	0.661	5.264	-8.874	0.654	0.076	0.25	0.228	0.205
Middle School	0.504	0.707	5.288	-7.56	0.616	0.0836	0.194	0.176	0.21
High School	0.495	0.707	5.287	-7.582	0.618	0.083	0.193	0.183	0.21

The one-way ANOVA results are presented in Table 3.4. The analysis reveals that the distribution of songs listened to by grade school students differs significantly from that of both high school and middle school students across all audio features. Among these features, energy, acousticness, and instrumentalness exhibit the highest F-statistics, indicating that the variance between group means is considerably greater than the variance within the groups. Loudness, however, is counted in the range [-60,0], which causes large variance for all groups. When comparing middle school and high school students, significant differences in means are observed for danceability, loudness, mode, speechiness, and instrumentalness. However, the p-values for energy, key, acousticness, and liveness do not suggest a statistically significant difference in means between the distinct LEs of middle school and high school students.

Table 3.4: One-Way ANOVA Results for Audio Features

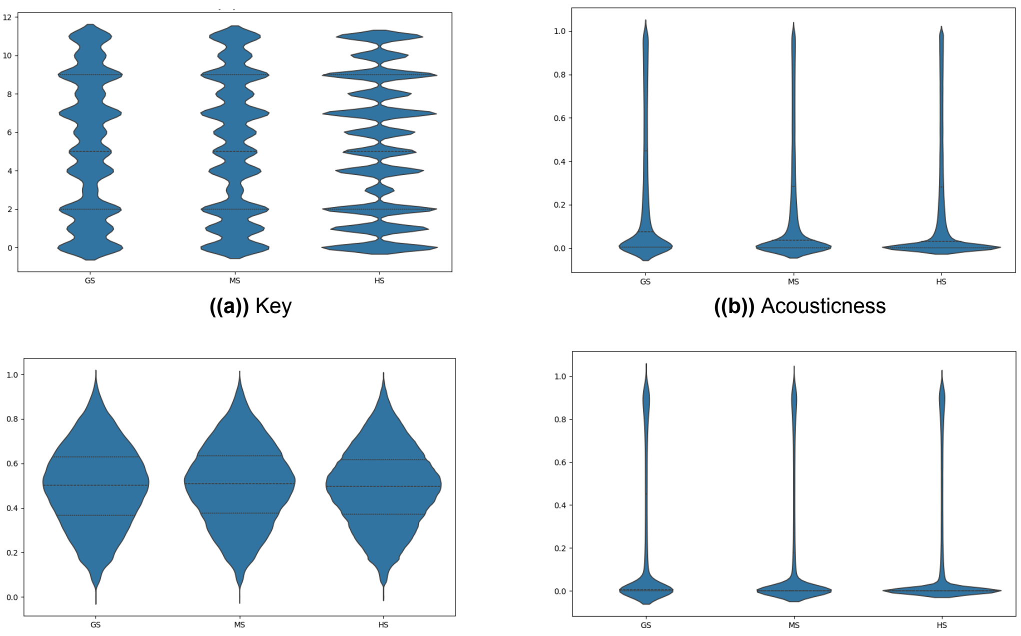
Education Level	Statistic	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	Liveness
Grade and Middle	p-value	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	W	152.65	4274.89	5.04	10321.50	752.15	945.08	4459.58	3246.47	107.82
Grade and High	p-value	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	W	54.30	6480.56	6.44	15623.16	986.76	1139.21	7012.67	3613.65	166.07
Middle and High	p-value	0.00	0.27	0.81	0.00	0.03	0.00	0.22	0.00	0.66
	W	957.33	1.24	0.06	7.93	4.79	17.53	1.50	187.84	0.19

The one-way ANOVA test is a statistical method used to determine whether there are significant differences between the means of the groups. Therefore, it does not fully capture the differences in the distribution of distinct LEs between educational levels, as two distributions might have similar means with their values varying significantly. Violin plots are particularly effective in illustrating these cases. For instance in Figure 3.2, even though instrumentalness means of middle and high school are relatively similar, the plot reveals that middle school students exhibit a broader range of instrumentalness values compared to their high school counterparts. Another example where the mean difference is misleading is the case of the key. Even though the values between 0 and 11 corresponding to the musical keys from C to B are in a specific musical order, the exploration of this feature can be significantly more valuable when visualizing the exact keys that are more often in the distributions. As shown in Figure 3.2, it is evident that high school students show a preference for certain keys, such as 5:F, 6: F♯/G♭, which are less common among other educational levels.

Table 3.4 shows that in all cases grade school students have more significant differences from students of middle school and high school students, than the students of high school and middle school have with each other. It is clear that grade school students have a unique identity as children in an early development stage and this identity is portrayed in their music listening preferences. Below is an overview of the differences observed in each audio feature between children of different educational levels:

- key:HS shows a significantly different distribution than the other groups, where 1: C♯/D♭, 5:F, 6: F♯/G♭ and G♯/A♭ are more common, as shown in Figure 3.2.
- acousticness: GS exhibits the widest variation in preference for acoustic music, with a significant amount of LEs corresponding to high acousticness songs, compared to other educational levels, as shown in Figure 3.2.
- danceability: Figure 3.2 shows that the danceability of songs appears to be very similar across age groups.
- energy: HS group shows a higher frequency of energy values that are near 1, compared to the other groups.

- instrumentalness: GS group exhibits a wider range of values, as HS group show a very narrow range of values near 0.
- loudness: There is a significant difference between grade school students and older students, as unique LEs listened by grade school students refer to songs with mean loudness of -8.871, compared to -7.56 and -7.582 for middle school and high school respectively.
- speechiness: The difference between grade school students and older students is also large, with older students listening to songs with higher speechiness.
- liveness: Grade school students listen to tracks with significantly lower liveness compared to both middle school and high school students.
- mode: Major is more often in the music grade school students, compared to older students.



((a)) Key

((b)) Acousticness

((c)) Danceability

((d)) Instrumentalness

Figure 3.2: Comparison of Distributions of Audio Features Across Educational Levels

Sentiment Analysis

To determine whether the lyric traits among songs listened by children from grade school, middle school, and high school follow a normal distribution, we conducted normality tests using the Shapiro-Wilk test. The results of the Shapiro-Wilk are presented in Table 3.5. Similarly to the audio features, the test indicated that the null hypothesis of normality was rejected for all educational levels and variables, suggesting that the distribution of music listening traits in this group significantly deviates from normality. However, the ANOVA test will still provide trustworthy results, due the high W value in all cases that indicates data close to normality and the large size of the dataset.

Table 3.5: Shapiro-Wilk Normality Test Results of Lyric Features

Education Level	Statistic	Positive	Negative	Anticipation	Disgust	Anger	Fear	Joy	Surprise	Flesch-Kincaid
Grade School	p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	W	0.92	0.96	0.84	0.91	0.92	0.86	0.83	0.89	0.9
Middle School	p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	W	0.93	0.95	0.83	0.9	0.9	0.88	0.87	0.91	0.91
High School	p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	W	0.93	0.95	0.83	0.9	0.91	0.88	0.87	0.91	0.91

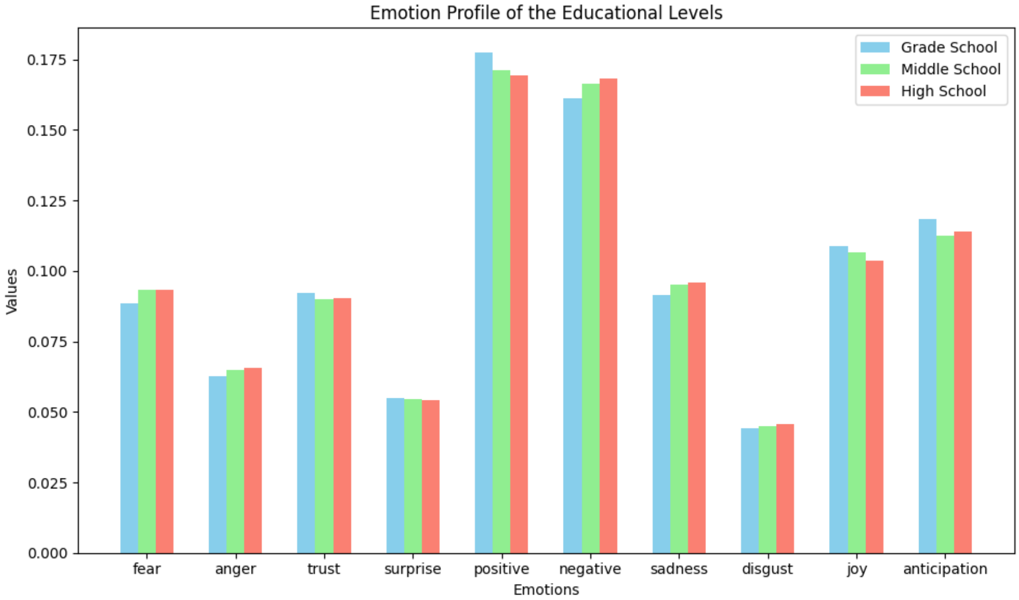


Figure 3.3: Sentiment Average for Listening Events per Educational Level

The one-way ANOVA results for the lyrics are presented in Table 3.6. The analysis reveals that the distribution of songs listened to by grade school students differs significantly from that of both high school and middle school students across all lyric features except for disgust. Among these features,negativity, surprise, anger and fear exhibit the highest F-statistics, indicating that the variance between group means is considerably greater than the variance within the groups. The differences between middle and high school are significant as well in the cases of positive and negative emotion, anger, surprise and readability, but the lower F-statistic in all cases indicates that the musical preferences of children in middle school and high school are relatively close, even though quite different from those of grade school students. Since disgust does not provide any valuable distinction between any educational levels, it will not be employed in the RS. All other sentiment and readability features will be utilized for the recommendation process.

As shown in Figure 3.3, there is a clear difference in the polarity of the sentiment across the different age groups. Specifically, grade school students tend to listen to the most positive songs, middle school students to less positive and high school students to the most negative. Anticipation is also an emotion where the songs of grade school students are associated with higher values than those of the older students. On the other hand, fear, anger and sadness are associated with lower values in the grade school students.

Table 3.6: One-Way ANOVA Results for Lyric Features

Education Levels	Statistic	Positive	Negative	Anticipation	Disgust	Anger	Fear	Joy	Surprise	Flesch-Kincaid
Grade and Middle	p-value	0.00	0.00	0.01	0.88	0.00	0.00	0.00	0.00	0.00
	F-statistic	24.18	128.34	6.91	0.69	42.02	75.18	34.95	133.63	180.44
Grade and High	p-value	0.00	0.00	0.00	0.41	0.00	0.00	0.00	0.00	0.00
	F-statistic	74.58	273.34	16.86	0.69	92.68	89.92	3.02	130.75	103.2
Middle and High	p-value	0.00	0.00	0.2	0.36	0.01	0.13	0.08	0.00	0.00
	F-statistic	13.57	17.26	1.58	0.83	6.21	2.24	3.02	15.7	71.83

Readability

As shown in Figure 3.5, the readability increases as the age of the listener increases. The vast majority of LEs of grade school students are values under 60, compared to middle school students and high school students who have a significant percentage of their distributions in values above 60 or even above 100. Therefore, the readability skills of children overall match the readability level of the lyrics of the songs they listen to.

Prevalent words

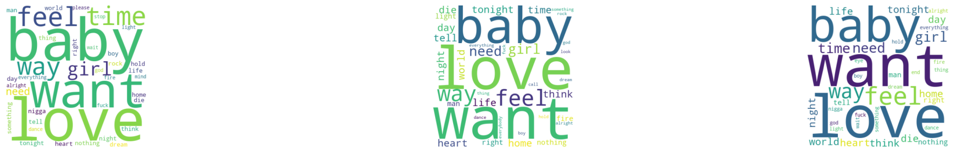
To investigate the content of the lyrics that children of different ages listen to, we identified the most prevalent words for each educational level. As we observe in the word clouds in Figure 3.4, a lot of words are present in the popular words of all collections. Even though the TF-IDF weighting scheme was chosen to penalize very common words present in many songs, words like ‘baby‘ and ‘love‘ are still mentioned in multiple songs across all educational levels. This indicates that love songs are very popular among children of all ages.

There are also some significant differences in the popular themes across age groups, which can be showcased in the prevalent words. The word ‘god‘ is not present in the collection of popular words for grade school students, even though this is the case for both high school and middle school students. This might mean that older students might prefer spiritual and philosophical themes. A similar case is the word ‘die‘, which is significantly more popular across middle school and high school compared to younger students. The subject of death is a difficult and sensitive topic that young children may not be inclined to engage with through songs. As mentioned, younger children prefer more uplifting songs with positive lyrics and therefore avoid emotionally challenging topics. On the other side, the word ‘nigga‘ is among the most popular words in LEs of grade school students, however, is much less popular across middle and high school students. This word is commonly used in hip-hop songs and this indicates that grade school students probably like hip-hop songs, and do not seem to always prefer content that is appropriate for their age.

To identify how similar or different are the prevalent words of the three educational levels, we identified the distributions of the top 100 words for each educational level and calculated the cosine similarity. Table 3.7 illustrates that the distributions of the prevalent words in middle school and high school students are highly similar, while the similarity between the prevalent words of grade school and the ones from middle school or high school remains very high, but comparatively lower. The high similarity in the distributions indicates that the vocabulary used in the songs listened to by children at different ages does not differ significantly, and therefore will not be utilized in the recommendation process.

Comparison	Cosine Similarity
Grade School and Middle School	0.937
Grade School and High School	0.948
Middle School and High School	0.984

Table 3.7: Cosine Similarity Between Prevalent Words of Educational Levels



((a)) Grade School

((b)) Middle School

((c)) High School

Figure 3.4: Wordclouds of the Most Prominent Words in the LEs of Each Educational Level

3.2.2. Discussion

In this section, we present an analysis of the key takeaways of our study’s findings. Specifically, we examine their alignment with existing literature on the music preferences of children and evaluate which characteristics are considered useful for providing relevant recommendations.

Result Interpretation. Our findings confirm prior research on children’s music listening preferences and build upon it by analyzing the lyric-based characteristics of songs. Specifically, we confirm previous findings [1] that instrumentalness and acousticness are significantly higher in the LEs of grade school

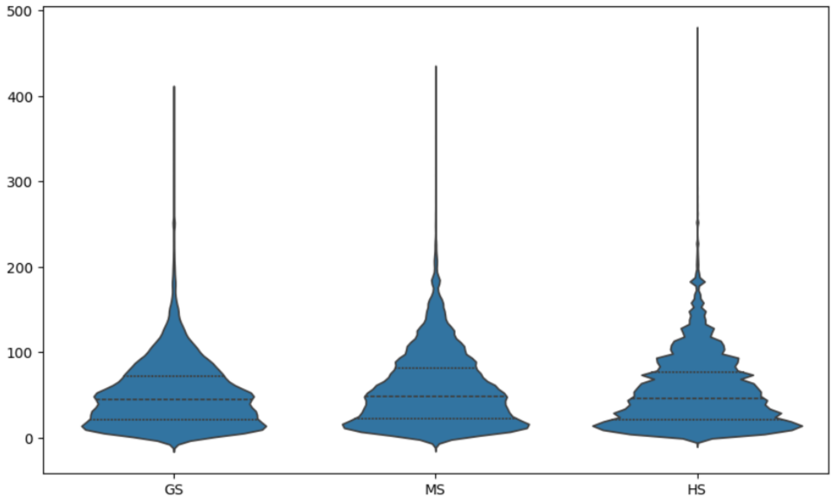


Figure 3.5: Distributions of Flesch-Kincaid Grade per Educational Level

students compared to those of older age groups. We also confirm that energy levels are lower in grade school students compared to older children.

The assumption that grade school students might gravitate towards happier and more uplifting songs [1] is also verified by the sentiment analysis results, which show that the songs listened by grade school students include on average more positive and joyful lyrics compared to those of middle school and high school, which contain fear, sadness and negative emotions. This is reasonable, as children also experience more negative emotions as they grow older [40]. Those emotions are responses to threats or challenges and younger children who have grown up in protected environments are usually less exposed to those. Middle and high school students go through adolescence, where significant cognitive changes occur, often leading to an increasing gap with their parents [41]. In that stage, listening to music helps them calm down their anxiety, anger and sadness [42].

The findings also show that the readability of the lyrics in the LEs of older students is higher compared to the ones of younger ones. This shows that the readability levels of the lyrics matches the readability skills of the children. This is not the case however for the appropriateness, as children of all ages tend to also listen to songs that are not necessarily appropriate for their age. Inappropriate content is available and even recommended to children in popular content platforms, such as Youtube [43]. Watching or listening to inappropriate content is an indication of problematic media use, which is linked to problematic behaviors, sleep problems and lower emotional intelligence [44]. The responsibility to avoid inappropriate content belongs mainly to the parents. Restrictive strategies from the parents and strong parent-child relationships reduce the risk of problematic media use [44].

This study shows that children at different ages have very different music preferences in terms of audio, readability and sentiment. However, current state-of-the art music RS do not cater to their specific needs and preferences. With this study, we have identified the discriminative features that capture the music preferences of children and therefore can enhance the recommendation process for children.

Features that can inform the recommendation process for children. Our empirical exploration showed significant differences between the educational levels in certain music characteristics. The identification of those characteristics informs us for the employment of these features in the RS. The significant differences show that a certain music characteristic is not universally appreciated to the same extent, but help capture the preferences of certain individuals and distinguish them from others. Therefore, we expect those audio and lyric features to be of great value when incorporated in a music RS.

In all audio features, a significant difference is observed for at least one pair of educational levels. The audio preferences of children vary significantly and we expect this to be a valuable predictor in the RS. Regarding sentiment features, a significant difference is observed for all features, except for disgust. By employing these characteristics in the RS, we expect that the recommendations for a certain user will adapt to the sentiment characteristics the user or similar users prefer. Readability also captures a significant difference between all pairs of educational levels with older students listening to songs with more complicated lyrics. Thus, we expect that incorporating readability in the RS will adjust the recommendations to cater for the preferences of children in terms of lyric complexity. Our analysis on the prevalent words showcases that the vocabulary used in the songs does not differ significantly among the LEs of children of different educational levels, therefore no feature regarding the vocabulary will be utilized. All features for which a significant difference is observed in at least one pair of educational levels are incorporated in the music RS, to achieve personalized music recommendations that adapt to the audio, sentiment and readability preferences of a user.

4

Enhancing the Recommendation Process

In this chapter, we integrate the features identified as discriminative into an RS to assess their effectiveness in generating relevant music recommendations. The chapter is organized into two sections: the first outlines the methodology employed, while the second presents the results and discusses the conclusions that can be inferred from these findings.

4.1. Methodology and Experimental Setup

After identifying the features for which the difference between different educational levels or between children and adults is significant, these features are used for an ablation study, where the RS will be trained on different sets of features. The evaluation metrics on each of the set of features, as well as the time to converge is captured for each set of features.

4.1.1. Recommendation Algorithm

To examine the impact of the discriminative features on the recommendation process, we employ Factorization Machines (FMs) [45] as an RS. FMs are supervised learning models that map real-valued features into a low-dimensional latent factor space. When used as an RS, FMs represent user-item interactions as tuples of real-valued feature vectors and numeric target variables. The following second-order model is used:

$$f(x) = w_0 + \sum_{p=1}^{P} w_p x_p + \sum_{p=1}^{P-1} \sum_{q=p+1}^{P} \langle v_p, v_q \rangle x_p x_q, \tag{4.1}$$

where the v_p, v_q are latent factor space embeddings of features p and q, w_p is the weight corresponding to feature p and x_p, x_q: the values of features p and q. The first summation term accounts for the individual contribution of each feature and the second term refers to the interaction terms for each pairwise feature combination. FMs use factorized interaction parameters: feature interaction weights are represented as the inner product of the two features' latent factor space embeddings.

As the LastFM dataset only includes implicit feedback, the Bayesian Personalized Ranking (BPR) loss [46] is employed. BPR attempts to learn the correct rank-ordering of items for each user by maximizing the posterior probability of the model parameters given a data set of observed user-item preferences and a chosen prior distribution. Each user's observed items are assumed to be preferred over the unobserved items, and all pairwise preferences are assumed to be independent. To learn these preferences, one creates training samples comprised of [user (u), observed item (i), unobserved item (j)] tuples and maximize the following log-likelihood function with respect to the model parameters. This is evident in the following formula:

$$\text{max } \theta \sum_{(u,i,j)\in S} \text{ln}\left(\sigma\left(f(u,i|\theta) - f(u,j|\theta)\right)\right) - \lambda\|\theta\|^2,$$

(4.2)

where i is an observed item for the user u, j is an unobserved item for the user u, $f(u, i \mid \theta)$ is the predicted score for user u and item i, parameterized by θ. $f(u, j \mid \theta)$ is the predicted score for user u and item j, parameterized by θ. λ is the regularization coefficient and $\sigma(x)$ is the sigmoid function. The first term is the difference between the predicted utility scores of the user's observed (i) and unobserved (j) items mapped onto [0, 1] using the sigmoid function and the second is a regularization term.

FMs are able to estimate parameters in very sparse datasets and calculate the model equation in linear complexity with respect to the number of features. The FM model is trained to learn latent factors for the song features, enabling it to capture the intricate relationships between users and songs. By leveraging Stochastic Gradient Descent (SGD) for optimization, we ensure scalable training on this large dataset. The trained FM model predicts the likelihood of a user listening to a particular song, facilitating personalized music recommendations based on the features.

We suggest an adaptation of FM for a music RS, where the items are combined with the auxiliary features that describe their audio characteristics, their sentiment scores and their readability. We will call this adaptation FMkids and will evaluate the performance of the RS using the combination of all features as well as the different subsets of features.

4.1.2. Metrics

To assess the effectiveness of our RS, we employ three key metrics: Normalized Discounted Cumulative Gain (NDCG) [47], Mean Reciprocal Rank (MRR) [48] and Hits@k, a widely used evaluation metric in RS research [49, 50, 51]. NDCG is utilized to measure the ranking quality by considering both the relevance of the recommendations and their positions in the ranking, providing a normalized score that indicates the performance of our RS relative to an ideal ranking. Hits@k is chosen to evaluate the system's ability to return relevant items within the top k results, highlighting the RS's recall capability in practical scenarios. Lastly, MRR is used to quantify the accuracy of the RS by focusing on the position of the first relevant recommendation, thus emphasizing the importance of retrieving relevant items as early as possible.

Together, these metrics offer a comprehensive evaluation of our RS. We investigate those metrics in terms of top-10 recommendations, a widely adopted and reliable approach in RS research [52, 53]. However, previous research shows that children do not often go further than the 5^{th} rank when using a search engine [54]. As children have a short attention span [55], we expect that they exhibit similar tendencies when interacting with recommender systems. Therefore, apart from top-10 recommendations, it is crucial to also investigate how the RS performs in terms of top-5 recommendations for each user.

4.1.3. Dataset

For the experiments we conduct on the RS, we employ the user-item interaction dataset from LastFM-2b, focusing on interactions recorded by children aged 6–17. This dataset is enriched with both audio and lyric features for each song. To ensure consistency across models during training and evaluation, only interactions with songs with complete feature data are included. Table 4.1 presents an overview of the dataset, detailing the number of interactions, unique users, and songs in the overall dataset, as well as in the training and test sets.

Dataset	Total Interactions	Unique Users	Unique Songs
Total Dataset	1,830,249	3,339	258,390
Training Set	1,372,770	3,332	231,703
Test Set	457,479	3,290	141,590

Table 4.1: Overview of the LastFM-2b Children's Interactions with Available Audio and Lyric Characteristics

4.1.4. Experiments

To explore how incorporating song features can enhance the recommendation process, we conduct an ablation study using FMs as the RS. Our primary goal is to evaluate the impact of incorporating various subsets of song features on enhancing the performance of the RS. To this end, we conduct a systematic ablation study, where we train and evaluate the RS for each set of features individually as well as combined with other subsets of features. Specifically, the following sets of features are employed: no features (baseline), audio features, sentiment features, lyric-based features:sentiment and readability, all features combined in the model we suggest called FMkids.

All experiments are conducted using a 75%-25% train-test split. We utilize FMs as the base model, chosen for their efficacy in modeling feature interactions, particularly in sparse datasets. The model was trained for 100 epochs with a fixed learning rate of 0.01. Optimization was carried out using Stochastic Gradient Descent (SGD), selected for its computational efficiency and scalability in large-scale data scenarios. These hyperparameters were determined based on preliminary experiments aimed at optimizing convergence and model performance.

The performance of FMs trained with each subset of song features is compared against a baseline model that trains only with user-item interaction data without additional song features. All models are evaluated for both top-5 and top-10 recommendations in terms of NDCG, MRR and Hits. The results of the models with the features are compared to the results of the baseline using a paired t-test in each evaluation metric. The paired t-test can be utilized because the results are directly comparable, as the training and the test set is the same for all models. The difference is considered significant if the p-value is below the threshold of 0.05. The results of the RS will also be captured and compared per educational level, to assess the reliability of recommendations to users of different ages. Since the data for grade school and middle school students is limited, those results are considered preliminary. The models are also compared based on the number of epochs required for convergence.

4.2. Results and Analysis

In this section, we present an analysis of our RS's performance, by training and evaluating the RS on different sets of features alongside user-interaction data to enhance recommendation accuracy. Evaluation metrics, including NDCG, Hit Ratio (Hits), and MRR, were utilized to rigorously assess the model's effectiveness. The subsequent analysis highlights the model's performance across these metrics, offering insights into its strengths and areas for potential refinement. The evaluation is captured for top-5 and top-10 recommendations. We also analyze the performance of the RS in providing recommendations for users of different age groups.

4.2.1. Recommender System Performance in Children

To explore how the audio and lyric features enhance the recommendation process, we analyze how the RS incorporating those features compares to the baseline model. This evaluation focuses on the top-5 and top-10 recommendations within the children population, as well as the top-10 recommendations for each educational level.

As shown in Table 4.2, the incorporation of song features in the FM can improve the performance in all evaluation metrics in the top-10 recommendations. The incorporation of audio features results in a more reliable model in terms of all evaluation metrics in the general population of children. When incorporating only sentiment-based features in the RS, the performance decreases in terms of Hits@10 and stays the same in terms of NDCG@10 MRR@10. This indicates that the user gets at least one recommended item less times than the baseline, but the relevant songs get ranked equally high in the top-10 as the baseline. However, readability has a noticeable effect, as it enhances the performance in all evaluation metrics compared to the baseline, when incorporated in the model alongside sentiment features. None of the observed differences is considered significant on the overall children population, as depicted in the results of the paired t-test in Table 4.3.

Even though the models including audio features and lyric features separately outperform the baseline, the combination of all features in FMkids performs worse than the baseline in terms of all evaluation metrics. This indicates that the pairs between sentiment, readability and audio features do not contribute useful information with predictive power to the RS. Additionally, FMkids and the model with the audio features achieve significant decrease in the loss function from the very early epochs, as shown in 4.1. The use of auxiliary features therefore not only increases the overall performance of the RS in children, but also decreases the number of epochs that the model needs to converge.

Different trends are noticed in the results of top-5 recommendations in Table 4.4, as no model outperforms the baseline in terms of all evaluation metrics. The model with audio features has a high performance

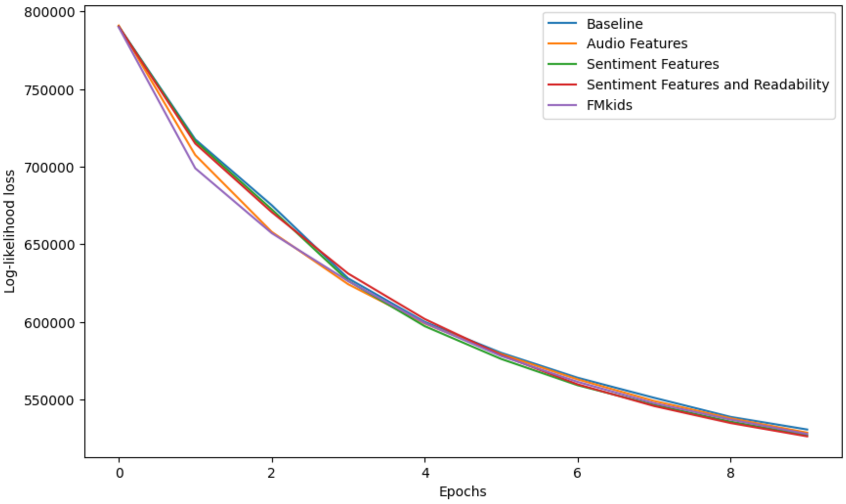


Figure 4.1: Loss Function Over 10 Epochs for all RS models

in the top-5 recommendations as well. This model has an enhanced performance in terms of NDCG@5 and MRR@5 compared to the baseline, but a decreased performance in terms of Hits@5. This indicates that the amount of times at least one relevant item is recommended decreases, but the most relevant items are ranked higher than the baseline. The same stands for the model with the sentiment features. In the top-5 recommendations, the model trained with the sentiment features and the readability is the model with the lowest performance in all evaluation metrics. This showcases that readability does not have an enhancing effect in the top-5 recommendations. While for top-10 recommendations readability enabled the capturing of relevant items more often and a high ranking of relevant items, it has the opposite effect in top-5 recommendations.

Table 4.2: Recommender System Performance for Top-10 Recommendations

Model	**Hits @10**	**NDCG @10**	**MRR @10**
Baseline	0.532	0.29	0.222
Audio features	0.538	0.294	0.223
Sentiment features	0.526	0.29	0.222
Sentiment features and readability	0.534	0.292	0.223
FMkids	0.527	0.287	0.217

4.2.2. Recommender System Results per Educational Level

As we previously discovered, children have different music preferences at different ages. Therefore, it is crucial to investigate the performance of the RS for children of different educational levels. For grade school students, Table 4.5 shows that the model incorporating audio features has the best performance compared to all other models. This indicates that the audio characteristics of a song can be a valuable predictor for an RS for grade school students. The recommender employing the audio features achieves a significantly higher Hits@10 value of 0.604, an NDCG@10 of 0.339 and an MRR@10 of 0.273, which is noticeably higher than the 0.547, 0.316 and 0.26 respectively of the baseline. This can be observed in Figure 4.2, where the median of the model with audio features outperforms the baseline in terms of NDCG@10 for grade school students. However, no other model outperforms the baseline for grade school students. The effect of readability is also clearly noticeable in grade school students, as the model that incorporates sentiment and readability performs better than the one utilizing only sentiment features. For

Table 4.3: Results of T-Test: Comparing the Performance of the Baseline with the Performance of Models Including Audio and Lyric Features on the Top-10 Recommendations

Models	Statistic	Hits@10	NDCG@10	MRR@10
Baseline vs Model with Audio Features	p-value	0.471	0.509	0.829
	t-statistic	−0.722	−0.661	−0.216
Baseline vs Model with Sentiment Features	p-value	0.467	0.906	0.99
	t-statistic	0.727	0.117	0.019
Baseline vs Model with Sentiment Features and Readability	p-value	0.847	0.768	0.928
	t-statistic	−0.192	−0.294	−0.09
Baseline vs FMkids	p-value	0.51	0.452	0.411
	t-statistic	0.658	0.014	0.823

Table 4.4: Recommender System Performance of All Models for the Top-5 Recommendations

Model	Hits @5	NDCG @5	MRR @5
Baseline	0.359	0.231	0.19
Audio features	0.357	0.234	0.195
Sentiment features	0.356	0.234	0.196
Sentiment features and readability	0.352	0.229	0.191
FMkids	0.358	0.231	0.191

middle school students, no model surpasses the performance of the baseline in Hits@10 and NDCG@10, while FMkids perform slightly better than the baseline in ranking the most relevant item. In high school students, both the model with audio features and the model with the lyric-based features outperform the baseline in terms of all evaluation metrics. The impact of readability is notably more pronounced in grade school students compared to middle and high school students. This observation aligns with the fact that children during the ages 6 and 11 have significant advancements in their reading skills.

The differences in the performance of the models for grade school and middle school students are more distinct compared to the high school students. This occurs because grade school and middle school students are underrepresented in the dataset and the majority of children are high school students. The results of grade school and middle school students should be considered preliminary, but provide valuable insight regarding the age groups that can benefit from an RS employing song features.

4.2.3. Discussion

In this section, we analyze the outcomes of our experiments, answer our research question based on them. We discuss the takeaways of our study and analyze the strengths and the limitations of the suggested adaptation of FMs.

The top-10 recommendation results indicate that incorporating audio and lyric features into the RS offers a performance enhancement in the general children population. However, this enhancement lacks statistical significance, suggesting insufficient evidence to support the utility of audio and lyric features for a population of children aged 6 to 17. The model with the audio features and the model with the sentiment and readability features outperform the baseline in terms of top-10 recommendations. However, combining these features does not lead to a performance enhancement. FMs utilize pairwise interactions between features, allowing the RS to capture combinations of audio and lyric characteristics. The combination of lyric and audio features does not yield to an increase in performance, which indicates that no meaningful combination between audio and lyric is captured by the model.

Additionally, the results reveal that readability is an important song characteristic, enhancing noticeably the RS performance in the general population of children and especially in grade school students . Incorporating readability adds a layer of personalization, addressing the cognitive and linguistic abilities of children and thereby improving the system's alignment with their needs.

Table 4.5: Recommender System Results for Each Educational Level

Model	Educational Level	Hits @10	NDCG @10	MRR @10
Baseline	Grade School	0.547	0.316	0.26
	Middle School	0.521	0.262	0.184
	High School	0.533	0.293	0.225
Audio Features	Grade School	0.604	0.339	0.273
	Middle School	0.472	0.239	0.164
	High School	0.543	0.298	0.228
Sentiment Features	Grade School	0.472	0.277	0.245
	Middle School	0.442	0.237	0.176
	High School	0.535	0.295	0.226
Sentiment Features and Readability	Grade School	0.528	0.308	0.25
	Middle School	0.468	0.243	0.176
	High School	0.54	0.296	0.226
FMkids	Grade School	0.491	0.302	0.244
	Middle School	0.472	0.251	0.185
	High School	0.532	0.29	0.22

In all evaluation metrics of top-5 and top-10 recommendations, at least one subset of features surpasses the baseline across NDCG and MRR. This indicates that incorporating song features into an RS can be a valuable contributor in the recommendation process, especially by ranking the relevant songs in a high position. To answer our research question, we find that leveraging song features such as the audio characteristics, sentiment, and readability can enhance the recommendation relevance by aligning recommendations with the developmental and cognitive preferences of children. For example, the inclusion of readability metrics promotes recommendations that match the vocabulary and comprehension levels of younger children, while sentiment analysis aligns recommendations with their emotional needs for positive emotions.

The performance enhancement is not considered significant in any of the experiments. This can be attributed to certain limitations of our study. The numerical values of audio, sentiment and readability characteristics were incorporated in the RS without further processing. However, these values were not originally designed for music recommendation purposes. Previous research has employed a Deep Belief Network to simultaneously transform the audio characteristics of songs and provide relevant recommendations and has shown promising results [29]. We believe that utilizing this algorithm for all discriminative features we have identified may enhance the recommendation process further. Additionally, there are some song characteristics that could be proved crucial but were not investigated so far. Content appropriateness is a crucial metric regarding the content children interact with [56] and including it in the RS may align the recommendations to the age of the child, thus protecting the children from lyrics that are not appropriate for their age.

Evaluating the RS across educational levels highlights the age groups that benefit most from the incorporation of song features, as well as the ones that do not benefit. The incorporation of audio features in the RS enhances the performance for grade school students in all aspects: more relevant items are included in the top-10 recommendation and the relevant items are ranked in higher positions. Grade school students are underrepresented in the dataset and item features can be helpful in capturing user preferences when limited user data is available [57]. As we investigated previously, their music preferences deviate from their older peers and therefore capturing them is crucial to provide reliable recommendations. Therefore, the model that integrates the audio characteristics can be a trustworthy solution for grade school students. The disadvantage of the algorithm is that it performs worse than the baseline in middle school students.

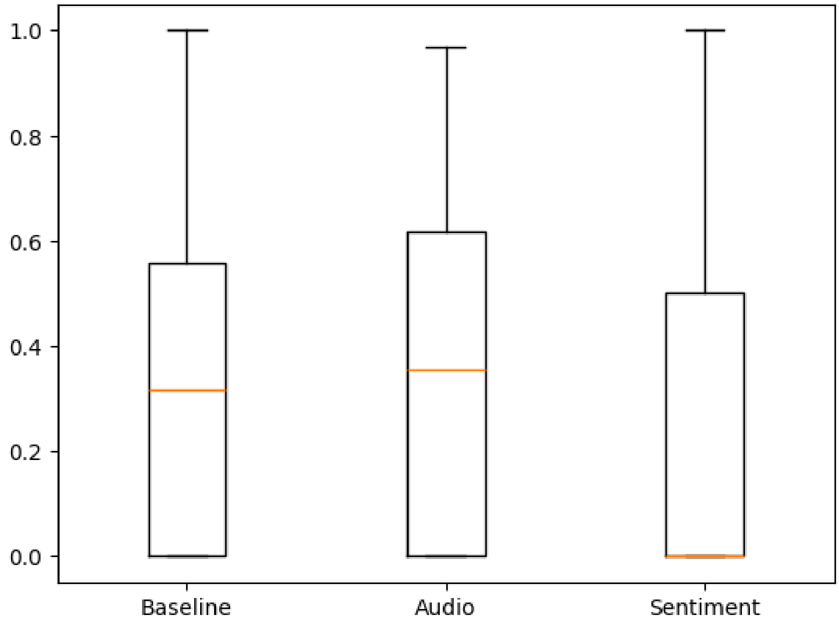


Figure 4.2: NDCG@10 Distribution for Grade School students for the baseline model, the model with the audio features, and the model with the sentiment features

These findings emphasize the importance of tailoring recommendation systems to children's cognitive and emotional needs. Additionally, they inform us about the age groups that can benefit from recommendations based on audio and lyric characteristics of songs. Based on the findings, we are able to draw conclusions about the limitations that hinder the discriminative features from enhancing the performance of the RS in children in a statistically significant way.

5 Ethical Considerations

5.1. Data Management

For this study we utilize the LastFM-2b dataset, the audio features dataset extracted using Spotify API and the lyrics dataset that is extracted using Genius API. This data is publicly available, therefore we do not participate in any data collection. We do not redistribute any of the datasets involved in this study and only publish the code that is necessary for the reproduction of our results.

5.2. Ethics

Research involving children must be conducted carefully to ensure ethical standards are upheld. In this study, the data utilized is publicly available. The data is anonymized and the research does not introduce any additional risks or vulnerabilities. This determination was made in collaboration with the data stewards of TU Delft.

5.2.1. Authorship policy

In the writing of this manuscript, we have utilized ChatGPT to assist with grammar and spelling corrections. This tool was used exclusively for these specific aspects of writing and had no impact on the development of ideas, research methods, or the academic content of this work, which are entirely our own.

Part III

Closure

6 Conclusion

This study presents an in-depth exploration of children's music preferences and the adaptation of an RS to leverage audio and lyric features to enhance recommendation relevance. The findings underscore the importance of understanding the unique cognitive and emotional needs of different age groups, providing a foundation for future research and system improvements. Below, we summarize the key insights from this work and outline potential directions for addressing existing limitations and advancing the field.

6.1. Conclusion

This study investigated the music preferences of children across different educational levels and developed a Recommender System (RS) tailored to their unique needs. By analyzing both audio and lyric features of songs, significant differences were identified in the preferences of grade school, middle school, and high school students. The empirical analysis demonstrated that younger children prefer songs with higher acousticness, instrumentalness, and positivity, while older students exhibit a preference for lyrics that convey negative emotions and anger more often than their younger peers. Similar vocabulary is prevalent in the songs listened by children at different ages, with grade school students slightly differing from middle school and high school students.

The audio and lyric characteristics that allow us to distinguish differences between the children are incorporated in an adaptation of FMs called FMkids. The evaluation of the FMkids incorporating these features revealed mixed results. The inclusion of audio and lyric features separately showed a modest improvement in NDCG@10 and MRR@10 and Hits@10. Regarding top-5 recommendations, the model incorporating audio features and the model with sentiment characteristics provides more relevant recommendations than the baseline in terms of NDCG@5 and MRR@5. The most notable difference is observed in grade school students, where the model trained on audio features noteably enhances the relevance of the recommendations in terms of all evaluation metrics. The findings highlight the importance of incorporating song characteristics into RS design to meet the unique preferences of children. The study also emphasizes the role of readability and sentiment as valuable predictors for music recommendation, especially for younger audiences. Despite limitations, the proposed approach provides a foundation for building RSs that better align with children's cognitive and emotional needs.

Although the enhancement of the recommendation performance is not statistically significant, it is evident that the incorporation of the discriminative audio and song characteristics in the recommendation process is a promising direction for recommender systems in children. Children have different inclinations compared to adults and investigating in-depth their unique preferences is the only way to design recommender systems that can capture their preferences.

6.2. Future Work

Even though LastFM-2b is a large and widely adopted dataset, it underrepresents children below 12 years old, compromising the reliability of the results and leading to potential biases in the RS performance. Future studies should aim to curate a more balanced dataset that adequately represents all educational levels to improve the generalizability of the findings. The findings of our research depicts significant differences in the music tastes of children at different ages, but does not investigate the differences between children and adults. Investigating the distinctions between the music preferences of adults and children provides essential context for our research and facilitates the identification of musical characteristics that are decisive in differentiating between these two groups. These characteristics can also enhance the recommendation process.

Our analysis explored various aspects of song characteristics, but there remain other intriguing angles to consider. For instance, examining the appropriateness of song lyrics could offer valuable insights into whether the appropriateness aligns with the age group of children that listen to a song. Additionally, we examined song characteristics individually, but did not explore how different audio and lyric characteristics correlate with each other. Correlation analysis would yield valuable insights regarding the co-occurrence of certain song characteristics and how this could be integrated in the RS.

The audio and lyric characteristics utilized in this empirical exploration proved valuable for gaining insights into children's music preferences, but they were not originally designed for music recommendation purposes. Prior research has employed deep learning networks to simultaneously transform audio features and generate relevant recommendations [29]. Integrating audio and lyric features into such deep learning frameworks would be a promising future research direction.

References

[1] Lawrence Spear et al. “Baby Shark to Barracuda: Analyzing Children’s Music Listening Behavior”. In: *Proceedings of the 15th ACM Conference on Recommender Systems*. RecSys ’21. Amsterdam, Netherlands: Association for Computing Machinery, 2021, pp. 639–644. DOI: `10.1145/3460231.3478856`. URL: `https://doi.org/10.1145/3460231.3478856`.

[2] Susan Hallam. “The power of music: Its impact on the intellectual, social and personal development of children and young people”. In: *International journal of music education* 28.3 (2010), pp. 269–289.

[3] GOTTFRIED SCHLAUG et al. “Effects of Music Training on the Child’s Brain and Cognitive Development”. In: *Annals of the New York Academy of Sciences* 1060.1 (2005), pp. 219–230. DOI: `https://doi.org/10.1196/annals.1360.015`. eprint: `https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1196/annals.1360.015`. URL: `https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1196/annals.1360.015`.

[4] Jacquelynne S Eccles. “The development of children ages 6 to 14”. In: *The future of children* (1999), pp. 30–44.

[5] Michael D. Ekstrand et al. “All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler et al. Vol. 81. Proceedings of Machine Learning Research. PMLR, 23–24 Feb 2018, pp. 172–186. URL: `https://proceedings.mlr.press/v81/ekstrand18b.html`.

[6] Andrew H Gregory et al. “The development of emotional responses to music in young children”. In: *Motivation and Emotion* 20 (1996), pp. 341–348.

[7] *Spotify API*. `https://developer.spotify.com/documentation/web-api`. Accessed: 2024-05-20.

[8] Maite Taboada et al. “Lexicon-based methods for sentiment analysis”. In: *Computational linguistics* 37.2 (2011), pp. 267–307.

[9] Kian Long Tan et al. “A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research”. In: *Applied Sciences* 13.7 (2023). DOI: `10.3390/app13074550`. URL: `https://www.mdpi.com/2076-3417/13/7/4550`.

[10] Erik Tromp et al. “Rule-based Emotion Detection on Social Media: Putting Tweets on Plutchik’s Wheel”. In: (Dec. 2014).

[11] Saif M. Mohammad et al. “Crowdsourcing a Word-Emotion Association Lexicon”. In: *Computational Intelligence* 29.3 (2013), pp. 436–465.

[12] George R Klare. “Assessing readability”. In: *Reading research quarterly* (1974), pp. 62–102.

[13] Rudolf Flesch. “Flesch-Kincaid readability test”. In: *Retrieved October* 26.3 (2007), p. 2007.

[14] Jeanne S Chall. *Readability revisited: The new Dale-Chall readability formula*. 1995.

[15] George D Spache. “Good reading for poor readers”. In: (1968).

[16] Francesco Ricci et al. “Recommender Systems Handbook”. In: vol. 1-35. Oct. 2010, pp. 1–35. DOI: `10.1007/978-0-387-85820-3_1`.

[17] Rashmi Sinha et al. “Comparing Recommendations Made by Online Systems and Friends”. In: (Sept. 2001).

[18] Arielle Bonneville-Roussy et al. “Music through the ages: Trends in musical engagement and preferences from adolescence through middle adulthood.” In: *Journal of personality and social psychology* 105.4 (2013), p. 703.

[19] Jinghan Gong. “The correlations between music preferences and personality”. In: *2020 5th International Conference on Humanities Science and Society Development (ICHSSD 2020)*. Atlantis Press. 2020, pp. 47–52.

[20] ALEXANDRA LAMONT. “Toddlers’ Musical Preferences”. In: *Annals of the New York Academy of Sciences* 999.1 (2003), pp. 518–519. DOI: https://doi.org/10.1196/annals.1284.063. eprint: https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1196/annals.1284.063. URL: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1196/annals.1284.063.

[21] David J. Hargreaves et al. “135Musical Preference and Taste in Childhood and Adolescence”. In: *The Child as Musician: A handbook of musical development*. Oxford University Press, June 2006. DOI: 10.1093/acprof:oso/9780198530329.003.0007. eprint: https://academic.oup.com/book/0/chapter/142895019/chapter-ag-pdf/45808552/book_2564_section_142895019.ag.pdf. URL: https://doi.org/10.1093/acprof:oso/9780198530329.003.0007.

[22] David J. Hargreaves et al. “303How and why do musical preferences change in childhood and adolescence?” In: *The Child as Musician: A handbook of musical development*. Oxford University Press, Sept. 2015. DOI: 10.1093/acprof:oso/9780198744443.003.0016. eprint: https://academic.oup.com/book/0/chapter/272408599/chapter-ag-pdf/44536987/book_32719_section_272408599.ag.pdf. URL: https://doi.org/10.1093/acprof:oso/9780198744443.003.0016.

[23] G. T. Barradas et al. “When words matter: A cross-cultural perspective on lyrics and their relationship to musical emotions”. In: *Journal of Adolescent Research* 50.2 (2022), pp. 1086–1114.

[24] Ana Belén Barragáns-Martínez et al. “A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition”. In: *Information Sciences* 180.22 (2010), pp. 4290–4311. DOI: https://doi.org/10.1016/j.ins.2010.07.024. URL: https://www.sciencedirect.com/science/article/pii/S0020025510003427.

[25] Hyeyoung Ko et al. “A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields”. In: *Electronics* 11 (Jan. 2022), p. 141. DOI: 10.3390/electronics11010141.

[26] Diego Sánchez-Moreno et al. “A collaborative filtering method for music recommendation using playing coefficients for artists and users”. In: *Expert Systems with Applications* 66 (2016), pp. 234–244. DOI: https://doi.org/10.1016/j.eswa.2016.09.019. URL: https://www.sciencedirect.com/science/article/pii/S0957417416304973.

[27] Brian McFee et al. *Learning content similarity for music recommendation*. 2011. arXiv: 1105.2344 [cs.MM]. URL: https://arxiv.org/abs/1105.2344.

[28] Andreu Vall et al. “Feature-combination hybrid recommender systems for automated music playlist continuation”. In: *User Modeling and User-Adapted Interaction* 29 (2019), pp. 527–572.

[29] Xinxi Wang et al. “Improving Content-based and Hybrid Music Recommendation using Deep Learning”. In: *Proceedings of the 22nd ACM International Conference on Multimedia*. MM ’14. Orlando, Florida, USA: Association for Computing Machinery, 2014, pp. 627–636. DOI: 10.1145/2647868.2654940. URL: https://doi.org/10.1145/2647868.2654940.

[30] Markus Schedl et al. “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*. CHIIR ’22. Regensburg, Germany: Association for Computing Machinery, 2022, pp. 337–341. DOI: 10.1145/3498366.3505791. URL: https://doi.org/10.1145/3498366.3505791.

[31] Genius. *Genius API Documentation*. https://docs.genius.com/. Accessed: 2024-11-06. 2024.

[32] E.S. Van der Valk Bouman et al. “The impact of different music genres on pain tolerance: emphasizing the significance of individual music genre preferences”. In: *Scientific Reports* 14 (2024), p. 21798. DOI: 10.1038/s41598-024-72882-2. URL: https://doi.org/10.1038/s41598-024-72882-2.

[33] Paul L Harris Francisco Pons et al. “Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization”. In: *European Journal of Developmental Psychology* 1.2

(2004), pp. 127–152. DOI: 10.1080/17405620344000022. eprint: https://doi.org/10.1080/17405620344000022. URL: https://doi.org/10.1080/17405620344000022.

[34] Nicole B. Capobianco et al. “Emotional Development: Cultural Influences on Young Children’s Emotional Competence”. In: *Children’s Social Worlds in Cultural Context*. Ed. by Tiia Tulviste et al. Cham: Springer International Publishing, 2019, pp. 55–73. DOI: 10.1007/978-3-030-27033-9_5. URL: https://doi.org/10.1007/978-3-030-27033-9_5.

[35] Stuti Shukla et al. “Review on sentiment analysis on music”. In: *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*. 2017, pp. 777–780. DOI: 10.1109/ICTUS.2017.8286111.

[36] Saif Mohammad et al. “NRC emotion lexicon”. In: (Jan. 2013). DOI: 10.4224/21270984.

[37] Filippos Vlachos et al. “Effect of age and gender on children’s reading performance: The possible neural underpinnings”. In: *Cogent Psychology* 2.1 (2015). Ed. by Peter Walla, p. 1045224. DOI: 10.1080/23311908.2015.1045224. eprint: https://doi.org/10.1080/23311908.2015.1045224. URL: https://doi.org/10.1080/23311908.2015.1045224.

[38] Patricia Grootens-Wiegers et al. “Research information for minors: Suitable formats and readability. A systematic review”. In: *Journal of Paediatrics and Child Health* 51.5 (May 2015). Epub 2014 Nov 2, pp. 505–511. DOI: 10.1111/jpc.12762.

[39] Bruce Ferwerda et al. “Personality traits and music genre preferences: how music taste varies over age groups”. In: *1st Workshop on Temporal Reasoning in Recommender Systems (RecTemp) at the 11th ACM Conference on Recommender Systems, Como, August 31, 2017.* Vol. 1922. CEUR-WS. 2017, pp. 16–20.

[40] Federica Izzo et al. “Children’s and Adolescents’ Happiness and Family Functioning: A Systematic Literature Review”. In: *International Journal of Environmental Research and Public Health* 19.24 (Dec. 2022), p. 16593. DOI: 10.3390/ijerph192416593.

[41] Sally I Powers et al. “Adolescent mental health.” In: *American psychologist* 44.2 (1989), p. 200.

[42] E Fiossi-Kpadonou et al. “Music and emotions of teenagers in Benin”. In: *Journal of Child and Adolescent Behavior (2016)* (2016), pp. 1–7.

[43] Kostantinos Papadamou et al. “Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), pp. 522–533. DOI: 10.1609/icwsm.v14i1.7320.

[44] Valeria Rega et al. “Problematic Media Use among Children up to the Age of 10: A Systematic Literature Review”. In: *International Journal of Environmental Research and Public Health* 20.10 (2023). DOI: 10.3390/ijerph20105854. URL: https://www.mdpi.com/1660-4601/20/10/5854.

[45] Steffen Rendle. “Factorization machines”. In: *2010 IEEE International conference on data mining*. IEEE. 2010, pp. 995–1000.

[46] Steffen Rendle et al. *BPR: Bayesian Personalized Ranking from Implicit Feedback*. 2012. arXiv: 1205.2618 [cs.IR]. URL: https://arxiv.org/abs/1205.2618.

[47] Yining Wang et al. *A Theoretical Analysis of NDCG Type Ranking Measures*. 2013. arXiv: 1304.6480 [cs.LG]. URL: https://arxiv.org/abs/1304.6480.

[48] Ellen M Voorhees et al. “The trec-8 question answering track report.” In: *Trec*. Vol. 99. 1999, pp. 77–82.

[49] Qinyong Wang et al. “Neural Memory Streaming Recommender Networks with Adversarial Training”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD ’18. London, United Kingdom: Association for Computing Machinery, 2018, pp. 2467–2475. DOI: 10.1145/3219819.3220004. URL: https://doi.org/10.1145/3219819.3220004.

[50] Tong Chen et al. “AIR: Attentional Intention-Aware Recommender Systems”. In: *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. 2019, pp. 304–315. DOI: 10.1109/ICDE.2019.00035.

[51] Hongzhi Yin et al. "SPTF: A Scalable Probabilistic Tensor Factorization Model for Semantic-Aware Behavior Prediction". In: *2017 IEEE International Conference on Data Mining (ICDM)*. 2017, pp. 585–594. DOI: `10.1109/ICDM.2017.68`.

[52] Vito Walter Anelli et al. "Top-N Recommendation Algorithms: A Quest for the State-of-the-Art". In: *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP '22. ACM, July 2022, pp. 121–131. DOI: `10.1145/3503252.3531292`. URL: `http://dx.doi.org/10.1145/3503252.3531292`.

[53] Paolo Cremonesi et al. "Performance of recommender algorithms on top-n recommendation tasks". In: *Proceedings of the fourth ACM conference on Recommender systems*. 2010, pp. 39–46.

[54] Jacek Gwizdka et al. "Analysis of Children's Queries and Click Behavior on Ranked Results and Their Thought Processes in Google Search". In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. CHIIR '17. Oslo, Norway: Association for Computing Machinery, 2017, pp. 377–380. DOI: `10.1145/3020165.3022157`. URL: `https://doi.org/10.1145/3020165.3022157`.

[55] Kenneth E Moyer et al. "The concept of attention spans in children". In: *The Elementary School Journal* 54.8 (1954), pp. 464–466.

[56] Ying Chen et al. "Children's exposure to mobile in-app advertising: an analysis of content appropriateness". In: *2013 International Conference on Social Computing*. IEEE. 2013, pp. 196–203.

[57] Ningxia Wang et al. "The impacts of item features and user characteristics on users' perceived serendipity of recommendations". In: *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 2020, pp. 266–274.