Beyond Songs: Analyzing User Sentiment Through Music Playlists and Multimodal Data

YIPEI CHEN, School of Management and Economics, UESTC, China
HUA YUAN*, School of Management and Economics, University of Electronic Science and Technology of China (UESTC), China
BAOJUN MA, School of Business and Management, Shanghai International Studies University, China
LIMIN WANG, School of Management and Economics, UESTC, China
YU QIAN, School of Management and Economics, UESTC, China

The automatic recognition of user sentiments through their music listening behavior is an important research task in cognitive studies. Whereas prior studies were conducted to identify the sentiment conveyed (or evoked) by a song that a user listens to at a particular time, we argue that a more effective method would be to identify the user's induced sentiment based on the comprehensive list of songs they have listened to (e.g., the sequence of music being played). However, recognizing the sentiment information induced by a playlist using machine learning techniques is much more challenging than identifying the sentiment induced by a single song, as it is difficult to obtain accurately labeled training samples for playlists. In this study, we developed the List-Song Relationship Factorization (LSRF) model with the objective of efficiently identifying sentiments induced by playlists. This model employs two side information constraints: the sentiment similarity between songs, based on multimodal information, and the co-occurrence of songs in playlists. These constraints enable the simultaneous co-clustering of songs and playlists. The experimental results demonstrate that the proposed model efficiently and consistently identifies sentiment information evoked by either playlists or individual songs.

CCS Concepts: • **Computing methodologies → Non-negative matrix factorization**; • **Information systems → Multimedia information systems**.

Additional Key Words and Phrases: Music sentiment, playlist, matrix factorization, multimodality data, co-clustering

1 INTRODUCTION

The emotional experience associated with music plays an important role in motivating creators and listeners to engage in musical activities [46, 82]. There are two main aspects of emotion in music: the expression of emotion through composition and performance (i.e., music-expressed emotion [MEE] [76]), and the induction or regulation of emotional states through listening (i.e., music-induced emotion [MIE] [78]). MEE is associated with the music creator and is embedded in the song's data (e.g., audio [96] and lyrics [81]), whereas MIE is linked to the listener

[*] Corresponding author.

Authors' addresses: Yipei Chen, School of Management and Economics, UESTC, Chengdu, China, 1021067299@qq.com; Hua Yuan, yuanhua@uestc.edu.cn, School of Management and Economics, University of Electronic Science and Technology of China (UESTC), Chengdu, China, 611731; Baojun Ma, School of Business and Management, Shanghai International Studies University, Shanghai, China, mabaojun@shisu.edu.cn; Limin Wang, School of Management and Economics, UESTC, Chengdu, China; Yu Qian, School of Management and Economics, UESTC, Chengdu, China, qiany@uestc.edu.cn.

ACM 1551-6865/2024/12-ART
https://doi.org/10.1145/3708346

and can be observed in listener's emotional response data (e.g., self-reports [95], expressive behaviours [19] and physiological responses [40, 50]).

The identification of emotions triggered by music (i.e., MIEs) has been demonstrated to have numerous practical applications in fields such as marketing [6], sport [74], and emotional therapy [17, 55]. Consequently, a surge in research on the recognition of MIEs has been observed in the literature [89]. Prior studies identified MIEs by measuring listeners' emotional responses[48]. However, recent research has also revealed some interesting phenomena in listeners' emotional responses that have been overlooked in prior studies. First, the same listener may exhibit different responses at different times [40] and in different situations [47]. Second, different listeners may respond differently to the same music clip [78, 96]. Thirdly, emotional responses to musical pieces are typically not long-lasting [67]. Consequently, even if researchers accurately identify listeners' MIEs through their emotional responses at a particular point in time, they may observe these MIEs to be in a state of flux, making it challenging to practically apply identification results. For instance, if a listener's emotional response to a piece of music is observed to be "depressed and moody" at a particular time, it is challenging to determine whether this indicates that the listener is experiencing some form of mental distress due to the transient nature of MIEs.

The preceding discussion illustrates that although existing methodologies can be used to detect immediate emotional reactions to music, these reactions may have limited utility in practical applications such as music therapy and situational music-playing scenarios. These findings suggest that while current methods can capture short-term emotional responses to music, they might not be sufficient for effective intervention or interaction in real-world settings. One study suggests that it is more meaningful to identify the listeners' relatively stable and sustained emotions or sentiments [63], evoked by their cumulative listening experience, than by focusing solely on the effects of a single song. For instance, a significant correlation has been demonstrated between the amount of music a patient listens to and changes in their treatment outcomes [14]. Notably, an individual's cumulative listening experience is determined by the songs they listen to over a period of time, typically arranged in the form of a playlist [49].

Playlists can be generated in one of three ways: by a listener's preferences (i.e., a user-generated playlist [30]), secondly, by a system/platform recommendation [12], or by a random generation [5]. Generally, a playlist is a sequence of songs that individuals may listen to collectively [49]. Individuals tend to prefer playlists that are relevant to their daily activities [3, 30, 58], suggesting that the organization of playlists may be based upon underlying logic or themes [5]. It can therefore be assumed that the sentiments evoked by playlists to which users continually listen are more likely to be consistent with their true sentiments. Accordingly, sentiment-based playlists represent a promising use case for music organizations [69].

However, the application of machine learning to identify sentiments associated with playlists presents technical challenges. Firstly, although there is no direct evidence of this, several studies suggest that although a playlist's overall sentiment is shaped by the content of its songs, it cannot be reduced to a simple sum of emotions for each individual song [28, 66]. Secondly, machine learning methods require well-labeled samples to train models, which are typically obtained through expert annotation [10]. However, user-generated playlists are personal, relatively unstructured, and variable in length, making manual annotation time-consuming [2]. Thirdly, it is difficult to ensure the quality and consistency of expert annotation results given songs of different styles. These challenges indicate that the expert annotation of mood for a collection of music is less effective than that for a single piece of music. Finally, the increasing number of user-generated playlists has led to sparsity problems owing to the varying popularity of songs and their frequency of inclusion in playlists [23]. Therefore, the following research question can be posed:

- Given the limited number of expert-annotated samples available, how can we efficiently and accurately identify the sentiments induced by songs and playlists?

One effective approach to address this issue is to utilize the significant correlation between specific underlying interrelated factors, such as songs and playlists [77]. To achieve this, we present the novel List-Song Relationship Factorization (LSRF) model, designed to accurately identify sentiments in both songs and playlists. This study makes three major contributions to the literature:

- We present the novel LSRF model for the sentiment recognition of songs and playlists. The model leverages a vast collection of user-generated playlists to establish a series of "song-list" associations. This methodology enables the precise recognition of playlist sentiments even in situations where only a limited number of playlist sentiment tag samples are available;
- We designed a methodology to vectorize song information through the fusion of multimodal data – specifically audio and text – thereby enabling machine learning algorithms to leverage more accurate information about sentiment similarity between songs;
- Through extensive data experimentation, we found that audio data play a more significant role in recognizing playlist sentiments than textual data such as lyrics and synopses. Moreover, we obtained experimental evidence that sentiment for a given playlist is not simply a sum of the sentiments of individual songs on the playlist.

The remainder of this paper is organized as follows. Section 2 discusses the related work. Section 3 describes the proposed method for recognizing song and playlist sentiments. Section 4 describes the data preprocessing steps, and Section 5 presents experimental results. Finally, Section 6 concludes the paper.

2 RELATED WORK

2.1 Music Data Processing

Our work falls within the domain of music emotion recognition (MER) [29], which entails the use of computational methods to process musical data and establish mapping relations between musical features and the emotional space.

2.1.1 Music data processing. Musical content primarily consists of audio data [64, 96] and textual data (i.e., lyrics) [8, 46, 81], with the occasional inclusion of video content [94]. The processing of *music data* in MER frequently requires the utilization of automated techniques to facilitate the preselection of important features [65, 89]. Earlier studies primarily used statistical methods to achieve this [81]. For example, the TF-IDF model was employed to identify sentiment features within lyrics [81].

Recently, ensemble learning methods have attracted significant attention from researchers owing to their ability to handle more complex training data, allowing them to consider the role of the majority of given data and autonomously learn the weights of different features. However, the aforementioned methods incur information loss in the process of discarding noisy data, either through feature selection from the training set or by assigning weights to features. This process involves transforming the original music data into a suitable representation space and then performing emotion calculations, resulting in satisfactory predictive performance. For instance, [64] examined standard audio features from established frameworks and developed an approach to categorize music into eight categories. In particular, the emergence of deep learning methods has provided superior options for processing music data [51].

2.1.2 Multimodality data fusion. In recent years, researchers have come to recognize the significance of integrating data from multiple modalities to predict musical emotions [38, 98]. The use of multimodal data has enabled the acquisition of large-scale heterogeneous multimedia data for sentiment analysis [79]. One common approach involves integrating audio and textual information from songs to identify the emotions conveyed by the songs [72].

The incorporation of multimodal data into the MER framework expands the scope of information available for the recognition of music emotions [92], including MEEs and MIEs. In the field of MER research, the fusion of audio-text data represents the most extensively discussed topic [79, 83, 98]. The fusion of audio and video data has likewise emerged as a topic of interest [94, 98]. Studies have also emerged on the fusion of music data with physiological data, such as electroencephalography (EEG) [75] and electrodermal activity signals [92].

However, the use of multimodal data introduces significant computational complexities. To address this issue, deep learning approaches such as deep representations [61, 79], are now commonly employed in MER research.

2.2 Music Emotion (Sentiment) Recognition

2.2.1 Emotion model. Emotions associated with music are typical measured using two types of representational models: categorical emotion states (CES) [35] and dimensional emotion space (DES) [68]. The categorical approach describes emotions using a limited number of innate and universal categories such as happiness, sadness, anger, and fear. The dimensional model considers all affective terms arising from independent neurophysiological systems: valence (negative to positive) and arousal (calm to exciting) [38, 68].

Prior research has indicated that the category model is somewhat ambiguous [85], whereas the DES is more consistently effective in evaluating MIE [78]. However, from the perspective of listeners' MIE, category-emotion information offers the advantage of being easily understandable by non-specialists [97]. Moreover, categorical information is closely linked to human behaviors, particularly manual labeling behaviors [71]. It is important to note that the metrics of dimensional models in data-driven emotion research are frequently inaccurately obtained from listeners' self-reports [44]. Consequently, some studies regarded positive and negative as separate categories in contrast to the parametric approach of the valence-arousal space [15, 52].

It is worth examining how MIE data can be obtained from listeners' emotional responses, and how these responses can be mapped to an appropriate emotion model. In general, emotional responses [2] can be quantified through self-report [95], expressive behaviour [19] and physiological responses (i.e., EEG [50]). In the case of music, it can be argued that [2]: 1) expressive behaviour is not a common phenomenon; 2) physiological signals provide a more objective means of observing certain listener behaviors and important physiological data. However, some emotions may be felt more than acted upon, and these emotions might not have obvious behavioral, expressive, or physiological manifestations [96]. Consequently, self-reporting, which requires listeners to report their emotions while listening to a song, is the most widespread - and arguably most informative - measure, as it provides insight into the cognitive aspects of emotions that are otherwise inaccessible [2, 95].

2.2.2 Emotion(Sentiment) Recognition Method. Prior studies on music emotions have focused on detecting two key types of information: MEEs and MIEs [46]. The task of automatically identifying MEEs is related to song data (i.e., audio [64, 96], lyrics [8, 46, 81], and video content [94]), whereas that of identifying MIEs is related to the listeners' emotional response data (i.e., listeners' self-reports [95], behavioural changes [19], and physiological signals [50]).

Generally, the objective of MEE detection is to establish a mapping from song data *SD* to an emotion measurement space *ES*, whereas that of MIE detection is to establish a mapping from emotional response data *RD* to *ES* [29]. Machine learning plays a pivotal role in the recognition of music emotions, particularly given large enough datasets *SD* and *RD* [29, 87, 90]. Table 1 presents a selection of studies that employed machine learning techniques to identify the MIE and MEE.

Note that, the emotional resonance of music allows the listener to connect with the emotions conveyed by the music, thereby serving as a significant indicator of the listener's emotional state [45]. As a result, the processes of identifying the MEE and MIE are inherently interconnected, with the song being listened to acting as the bridge between them.

Table 1. Sample studies on MIE and MEE recognition

Tasks	Data type	Technique
MEE and MIE	Audio	PCA + Random Forest (RF) [87]
	Audio	Gaussian Mixture Model [80]
MEE	Audio	CNN+SVM [36]; CNN + RNN [22]
	Lyrics	Decision Tree (DT) [88], SVM [8]
	Lyrics + Audio	CNN [94]
MIE	Audio	PCA+SVM [91], AdaBoost [91]
	Physiological signal	SVM, NB, KNN, DT [40]
	EEG	PCA [50], CNN+LSTM [43]

The potential business and medical applications of playlists have fueled a range of playlist-related MER studies. Such applications include playlist-based song recommendation [30, 49, 77], the scenario-based generation of music playback sequences [58], and bot-generated (random) playlists on platforms [3, 5]. To the best of our knowledge, the problem of identifying the MIE of a playlist has not yet been addressed. Accordingly, this study was conducted to investigate listeners' emotional responses to music playlists, i.e., music(playlist)-induced emotion.

2.3 Nonnegative Matrix Factorization

Non-negative matrix factorization (NMF) is a matrix operation originally employed as a relatively novel paradigm for dimensionality reduction [86]. The operational features of NMF can be characterized as follows: first, non-negativity constraints are used to obtain a partial representation that improves the problem's interpretability [84]; secondly, by adding some necessary side information [37], NMF can be used to normalize the convergence of matrix factorization results to a particular space.

Although most prior NMF studies focused on the two-factor approach, the three-factor approach has been subjected to a comprehensive and systematic analysis. The results of the study [20] demonstrate that the orthogonal tri-factor NMF (ONMTF) is capable of clustering both the rows and columns of the input data matrix. In the literature, ONMTF is widely used to determine class membership in a diverse range of clustering applications, including recommendations [73], text clustering [11], and sentiment classification [57].

3 METHODOLOGY

3.1 Problem statement

The automatic recognition of user sentiments through their music listening behavior is a significant research topic in cognitive studies. As discussed earlier, we believe it is more effective to identify users' sentiment based on their generated playlist of songs (a series of music being played) rather than the specific song to which they are currently listening.

Assuming that the collection of playlists constitutes a 'playlist space' denoted by $\mathbb{L} = \{l_1, ..., l_m\}$, where there are m distinct lists containing a total of n unique songs, the set of songs forms a 'song space' denoted by $\mathbb{S} = \{s_1, ..., s_n\}$. Furthermore, it is possible to construct a sentiment space with c dimensions, denoted as $\mathbb{E} = \{e_1, ...e_c\}$, where the value of c is determined by the number of sentiment clusters in the song space. We adopted an approach similar to that in previous studies[53, 90] by setting $c = 3$. Thus, our sentiment space was defined as follows:

$$\mathbb{E} = \{e_1, e_2, e_3\} = \{Positive, Negative, Neutral\}. \tag{1}$$

Letting s_{ij} denote the j-th song in the i-th playlist l_i, we can define the structure of a playlist $l_i \in \mathbb{L}$ as follows:

$$l_i = (s_{i1}, ..., s_{ij}, ...), \tag{2}$$

where $s_{ij} \in \mathbb{S}$, and the relationship $l_i \subset \mathbb{S}$ is easily obtained. Based on the co-occurrence relationship of songs in l_i, we may construct the following “list-song” relationship matrix:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \tag{3}$$

where a_{ij} denotes the frequency at which the j-th song appears in playlist l_i. Note that Equation (3) does not assume that all playlists contain the same number of songs. If the j-th song in $\mathbb{S}$ is not in the i-th playlist, then we can simply set $a_{ij} = 0$.

Furthermore, the MIE sentiment inspired by a song $s \in l$ is denoted as $\boldsymbol{e}_s$, whereas that inspired by an entire playlist l is denoted as $\boldsymbol{e}_l$. Both sentiments exist within the space $\mathbb{E}$. Accordingly, the research problem can be expressed as follows:

- How can the prior sentiment information in the *song space* $\mathbb{S}$ and “list-song” association $\boldsymbol{A}$ be used to accurately identify the sentiment of song playlists in $\mathbb{L}$?

In the following subsection, we describe the proposed LSRF model designed to identify sentiments from playlists and individual songs.

3.2 Research framework

ONMTF [20] is commonly used to explore information within data structures [41, 93], as described in Equation (4):

$$\min_{\boldsymbol{U},\boldsymbol{H},\boldsymbol{V}\geq 0} O = \parallel \boldsymbol{A} - \boldsymbol{U}\boldsymbol{H}\boldsymbol{V}^T \parallel_F^2, \ s.t., \quad \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}, \ \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}, \tag{4}$$

where $\parallel \cdot \parallel$ denotes the Frobenuis norm of a matrix, $\boldsymbol{V} \in \mathbb{R}_+^{c\times n}$ is the sentiment matrix of songs, $\boldsymbol{H} \in \mathbb{R}_+^{c\times c}$ is the “list-song” topic relationship matrix, $\boldsymbol{U} \in \mathbb{R}_+^{m\times c}$ is the sentiment matrix of playlists, and $\boldsymbol{I}$ is the identity matrix. This equation describes the decomposition of the co-occurrence matrix $\boldsymbol{A}$ into three matrices $\boldsymbol{U}$, $\boldsymbol{H}$, and $\boldsymbol{V}$, while minimizing the information difference.

The proposed LSRF model is illustrated in Figure 1. The core ONMTF process performs the sentiment-based clustering of both songs and playlists on a list-song matrix $\boldsymbol{A}$. To guarantee that the outcome of the matrix decomposition - namely the matrices $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{H}$ - falls within the sentiment space delineated by Equation (1), LSRF imposes four constraints: the blue path in the upper half of Figure 1 depicts the process of adding the expert annotations as constraints to the song clustering; the gray path shows how similarities between songs, as determined by their multimodal data representations, are added to the song clustering; the blue path in the bottom half of Figure 1 shows the process of imposing fewer expert annotations on the playlist clustering; and the gray path shows the addition of audio and text similarity between playlists to the playlist clustering.

In the training phase, the LSRF accepts the list-song relationships - i.e. the matrix $\boldsymbol{A}$ - as input. The outputs are the sentiment representations of the sample lists and the songs in space $\mathbb{E}$, that is, matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. In the testing phase, matrix $\boldsymbol{U}$ can be used as a classifier to infer sentiment information about unknown playlists by musical data similarity [59] or by training an additional sentiment classifier [86]. Similarly, matrix $\boldsymbol{V}$ can be used as a classifier to infer sentiment information regarding sentiment-unknown songs.

Essentially, the model is trained on a large volume of expert-labeled MIE information from songs (Subsection 3.3) and a small volume of expert-labeled MIE information from playlists (Subsection 3.4). The LSRF incorporates two essential pieces of additional information: the MIE associations between pairs of similar songs (Subsection 3.5), and pairs of similar playlists (Subsection 3.6). Finally, the song-playlist matrix is decomposed to compute the MIE association between songs and playlists (Subsection 3.7).

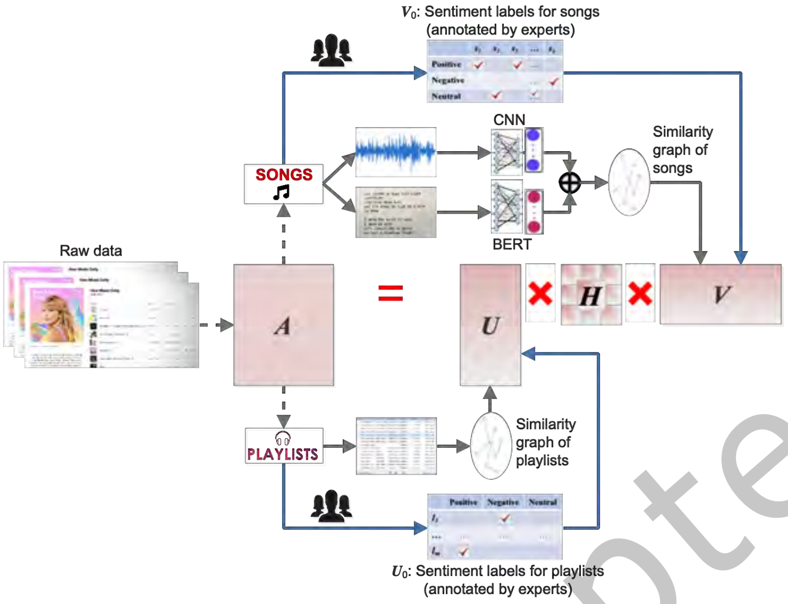


Fig. 1. Framework of the LSRF model.

3.3 Modeling the sentiment annotations of songs

The sentiment annotation results of songs are modeled to ensure that the final sentiment labels assigned to the songs by the LSRF model are maximally consistent with the expert annotations. This can be achieved by minimising the following loss functions:

$$\| \boldsymbol{V} - \boldsymbol{V}_0 \|_F^2, \tag{5}$$

where $\boldsymbol{V} \in \mathbb{R}_+^{n \times c}$ is the song cluster matrix and $\boldsymbol{V}_0 \in \mathbb{R}_+^{n \times c}$ represents the initial sentimental information matrix of the *song space*. According to the definition of Equation (1), $\boldsymbol{V}_0(i, *) = (1, 0, 0)$ indicates that song i has a positive sentiment; $\boldsymbol{V}_0(i, *) = (0, 1, 0)$ indicates that song i has a negative sentiment; while $\boldsymbol{V}_0(i, *) = (0, 0, 1)$ indicates that song i has a neutral sentiment; and $\boldsymbol{V}_0(i, *) = (0, 0, 0)$ indicates that song i has an unknown sentiment.

To mitigate the noisy effects introduced by unknown elements of $\boldsymbol{V}_0$, we introduce the diagonal indicator matrix $\boldsymbol{G}^v \in \{0, 1\}^{n \times n}$ to identify whether the sentiment information of a song has been expertly annotated with a true label. Specifically, $\boldsymbol{G}^v(i, i) = 1$ indicates that song i has been expertly annotated with a true sentiment label, whereas $\boldsymbol{G}^v(i, i) = 0$ indicates otherwise. Using this information, the loss function for all songs can be expressed as follows:

$$\| \boldsymbol{G}^v(\boldsymbol{V} - \boldsymbol{V}_0) \|_F^2 \,. \tag{6}$$

3.4 Modeling the sentiment annotations of playlists

Although we acknowledge that the manual annotation of playlists is more challenging than that of songs, the LSRF model represents a semi-supervised machine learning method that utilizes a limited number of sentiment-marked playlist samples to guide the clustering process, thereby enhancing its efficacy.

In the learning process, the LSRF aims to closely match the initially annotated values of expert-labeled lists with the learned sentiment information, as in the case of modeling song-sentiment annotations. This process can be equated to minimizing the following loss function:

$$\| \boldsymbol{G}^u(\boldsymbol{U} - \boldsymbol{U}_0) \|_F^2 \,. \tag{7}$$

where $\boldsymbol{G}^u \in \{0,1\}^{m \times m}$ is a diagonal indicator matrix indicating whether the playlist has been annotated by experts, and $\boldsymbol{U}_0 \in \mathbb{R}_+^{m \times c}$ represents the initial sentiment information matrix of the playlists.

3.5 Modeling the sentiment relevance of songs

3.5.1 Multimodal data representation of song. When identifying the sentiments conveyed by a song, the LSRF primarily analyzes its audio and textual components [94], with the latter mainly consisting of lyrics (if available) and the song's profile text. To effectively combine the two modalities for sentiment recognition, it is necessary to transform them into a computationally useful format, specifically, a vectorized representation of the song.

Given a song $s_i \in \mathbb{S}$, we employed convolutional neural networks (CNNs) to vectorize the audio data $s_i^{(a)}$ of s_i owing to the proven success of such networks in audio classification tasks [34]. The audio data $s_i^{(a)}$ for song s_i is fed into a CNN architecture, resulting in a d-dimensional vector $\boldsymbol{s}_i^{(a)} \in \mathbb{R}^d$. Given two songs, s_i and s_j, we can calculate their audio modal similarity as follows [70]:

$$sim\big(s_i^{(a)}, s_j^{(a)}\big) = \frac{\boldsymbol{s}_i^{(a)} \cdot \boldsymbol{s}_j^{(a)}}{\| \boldsymbol{s}_i^{(a)} \| \| \boldsymbol{s}_j^{(a)} \|}. \tag{8}$$

Additionally, we treat the text associated with s_i as a document $s_i^{(t)}$ and utilize the BERT [18] method to vectorize it, resulting in a vector $\boldsymbol{s}_i^{(t)} \in \mathbb{R}^d$. The textual similarity $sim\big(s_i^{(t)}, s_j^{(t)}\big)$ can likewise be calculated using Equation (8). Finally, different weights are assigned to each modal similarity to derive the weighted cosine similarity between the two songs [56] as follows:

$$sim(s_i, s_j) = \alpha sim\big(s_i^{(t)}, s_j^{(t)}\big) + (1-\alpha) sim\big(s_i^{(a)}, s_j^{(a)}\big), \tag{9}$$

where $\alpha \in [0,1]$ that represents the importance of the text modal similarity. This scheme partially compensates for the deficiency that results from the inability to synchronize the optimization of CNN/BERT and NMF in the training phase.

3.5.2 Manifold regulation of sentiment relevant songs. Previous research on spectral graph theory [13] and manifold learning theory has demonstrated that a local geometric structure can be accurately represented by a *nearest neighbor graph* based on the distribution of data points [7]. In light of this, we constructed a song-song relationship graph $\mathcal{G}^v$ to depict the sentimental correlation between songs.

We considered the top-k songs that were similar to s_i and also sentimentally relevant to s_i. To simplify the edges in graph $\mathcal{G}^v$, we retained only the connections between each song and its top-k similar songs. Consequently, the adjacency matrix of the graph $\mathcal{G}^v$ can be expressed as follows:

$$\boldsymbol{W}^v(i,j) = \begin{cases} 1, & \text{if } s_j \in \mathcal{N}^k(s_i); \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

where $\mathcal{N}^k(s_i)$ represents the set of k-nearest neighbors for song s_i. We can infer from the previous discussion that if two nodes are close to each other in the graph $\mathcal{G}^v$, their corresponding sentiment labels will also be similar. This inference can be expressed mathematically by minimizing the following loss function:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \| \boldsymbol{V}(i,*) - \boldsymbol{V}(j,*) \|_F^2 \, \boldsymbol{W}^v(i,j) = Tr(\boldsymbol{V}^T(\boldsymbol{D}^v - \boldsymbol{W}^v)\boldsymbol{V}), \tag{11}$$

where $Tr(.)$ is the trace of a matrix, $\boldsymbol{D}^v - \boldsymbol{W}^v$ is the Laplacian matrix of the graph $\mathcal{G}^v$ [13, 27], $\boldsymbol{D}^v \in \mathbb{R}^{n \times n}$, and $\boldsymbol{D}^v(i,i) = \sum_{i=1}^{n} \boldsymbol{W}^v(i,j)$. If two songs are close on graph $\mathcal{G}^v$ but have different sentiment labels, Equation (11) will penalize this discrepancy.

3.6 Modeling the sentiment relevance of playlists

3.6.1 Data representation of playlists. When a user enjoys a particular playlist, the songs comprising the playlist work together to influence their sentiment. Assuming that the audio representation vector for song s_{ij} is denoted as $\boldsymbol{s}_{ij}^{(a)}$, the audio features of l_i can be expressed as follows:

$$\boldsymbol{l}_i^{(a)} = \frac{\sum_j a_{ij} \boldsymbol{s}_{ij}^{(a)}}{\sum_j a_{ij}}. \tag{12}$$

Accordingly, the audio similarity of playlists l_i and l_j can be measured as follows:

$$sim(l_i^{(a)}, l_j^{(a)}) = \frac{\boldsymbol{l}_i^{(a)} \cdot \boldsymbol{l}_j^{(a)}}{\parallel \boldsymbol{l}_i^{(a)} \parallel \parallel \boldsymbol{l}_j^{(a)} \parallel}. \tag{13}$$

Subsequently, the text data of each song in the playlist have been vectorized, enabling the calculation of textual similarity $sim(l_i^{(t)}, l_j^{(t)})$ for the entire playlist using the weighting provided by Equation (12). The multimodal-based similarity between two playlists is then determined through a calculation similar to Equation (9). Furthermore, previous research has demonstrated that the melodies of songs have a greater impact on eliciting sentiments than the lyrics [1]. Hence, in the computation of sentiment similarity between two playlists, it is essential to pay considerable attention to the influence of the audio characteristics of each playlist.

3.6.2 Manifold regulation of sentiment relevant playlists. We constructed a list-list relationship graph, denoted as $\mathcal{G}^u$, to represent the sentiment correlation between playlists. To simplify the edges in $\mathcal{G}^u$, we retained only the connections of each playlist with its top-k similar neighbors. Consequently, the adjacency matrix of $\mathcal{G}^u$ can be expressed as follows:

$$\boldsymbol{W}^u(i,j) = \begin{cases} 1, & \text{if } l_j \in \mathcal{N}^k(l_i); \\ 0, & \text{otherwise,} \end{cases} \tag{14}$$

where $\mathcal{N}^k(l_i)$ represents the set of the k-nearest neighbors of l_i. The rationale behind this approach is that nodes close to each other in a graph tend to have similar sentiment labels. This can be formalized by minimizing the following loss function:

$$\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \parallel \boldsymbol{U}(i,*) - \boldsymbol{U}(j,*) \parallel_F^2 \boldsymbol{W}^u(i,j) = Tr(\boldsymbol{U}^T(\boldsymbol{D}^u - \boldsymbol{W}^u)\boldsymbol{U}), \tag{15}$$

where $Tr(.)$ is the trace of a matrix, $\boldsymbol{D}^u - \boldsymbol{W}^u$ is the Laplacian matrix of graph $\mathcal{G}^u$, $\boldsymbol{D}^u \in \mathbb{R}^{m \times m}$, and $\boldsymbol{D}^u(i,i) = \sum_{i=1}^{m} \boldsymbol{W}^u(i,j)$.

3.7 Exploring sentiments in playlists

3.7.1 Objective function. Building upon the previous analysis, we integrated expert-labeled prior information, geometric regularization, and the original NMF objective function to construct a mathematical model for investigating the sentiments of complex playlists. The model can be formulated by the following optimization problem:

$$\begin{aligned}\min_{\boldsymbol{U},\boldsymbol{H},\boldsymbol{V}\geq 0} O =& \| \boldsymbol{A} - \boldsymbol{U}\boldsymbol{H}\boldsymbol{V}^T \|_F^2 + \lambda_1^u \| \boldsymbol{G}^u(\boldsymbol{U} - \boldsymbol{U}_0) \|_F^2 + \lambda_2^u Tr(\boldsymbol{U}^T(\boldsymbol{D}^u - \boldsymbol{W}^u)\boldsymbol{U}),\\ &+ \lambda_1^v \| \boldsymbol{G}^v(\boldsymbol{V} - \boldsymbol{V}_0) \|_F^2 + \lambda_2^v Tr(\boldsymbol{V}^T(\boldsymbol{D}^v - \boldsymbol{W}^v)\boldsymbol{V}),\\ &s.t., \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I},\ \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I},\end{aligned} \tag{16}$$

where λ_1^u, λ_2^u, λ_1^v, and λ_2^v are positive regularization parameters that control the contribution of expert-labeled playlist sentiment data, sentiment correlation between playlists, expert-labeled song sentiment data, and sentiment correlation between songs, respectively.

3.7.2 Optimization algorithm. Because no closed-form solution exists for model optimization (16), we utilized a local-optimization-based approach that has been previously employed to address similar problems [20, 27, 41].

We first derive the updating rule for $\boldsymbol{H}$ as follows:

$$\boldsymbol{H}(i,j) \leftarrow \boldsymbol{H}(i,j)\sqrt{\frac{[\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V}](i,j)}{[\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{H}\boldsymbol{V}^T\boldsymbol{V}](i,j)}}. \tag{17}$$

To simplify the expression, we define $\boldsymbol{\Gamma}_U = \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{H}^T - \boldsymbol{H}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{H}^T - \boldsymbol{\lambda_1^u}\boldsymbol{U}^T\boldsymbol{G^u}(\boldsymbol{U} - \boldsymbol{U_0}) - \lambda_2^u\boldsymbol{U}^T(\boldsymbol{D}^u - \boldsymbol{W}^u)\boldsymbol{U}$, $\boldsymbol{\Gamma}_U^+(i,j) = \frac{|\boldsymbol{\Gamma}_U(i,j)|+\boldsymbol{\Gamma}_U(i,j)}{2}$ and $\boldsymbol{\Gamma}_U^-(i,j) = \frac{|\boldsymbol{\Gamma}_U(i,j)|-\boldsymbol{\Gamma}_U(i,j)}{2}$, then the updating rule for $\boldsymbol{U}$ is defined by Equation (18).

$$\boldsymbol{U}(i,j) \leftarrow \boldsymbol{U}(i,j)\sqrt{\frac{[\boldsymbol{A}\boldsymbol{V}\boldsymbol{H}^T + \lambda_1^u\boldsymbol{G}^u\boldsymbol{U}_0 + \lambda_2^u\boldsymbol{W}^u\boldsymbol{U} + \boldsymbol{U}\boldsymbol{\Gamma}_U^-](i,j)}{[\boldsymbol{U}\boldsymbol{H}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{H}^T + \lambda_1^u\boldsymbol{G}^u\boldsymbol{U} + \lambda_2^u\boldsymbol{D}^u\boldsymbol{U} + \boldsymbol{U}\boldsymbol{\Gamma}_U^+](i,j)}}. \tag{18}$$

Finally, the updating rule for $\boldsymbol{V}$ is given by the following Equation (19),

$$\boldsymbol{V}(i,j) \leftarrow \boldsymbol{V}(j,j)\sqrt{\frac{[\boldsymbol{A}^T\boldsymbol{U}\boldsymbol{H} + \lambda_1^v\boldsymbol{G}^v\boldsymbol{V}_0 + \lambda_2^v\boldsymbol{W}^v\boldsymbol{V} + \boldsymbol{V}\boldsymbol{\Gamma}_V^-](i,j)}{[\boldsymbol{V}\boldsymbol{H}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{H} + \lambda_1^v\boldsymbol{G}^v\boldsymbol{V} + \lambda_2^v\boldsymbol{D}^v\boldsymbol{V} + \boldsymbol{V}\boldsymbol{\Gamma}_V^+](i,j)}}, \tag{19}$$

where $\boldsymbol{\Gamma}_V^+$ is defined as $\boldsymbol{\Gamma}_V^+(i,j) = \frac{|\boldsymbol{\Gamma}_V(i,j)|+\boldsymbol{\Gamma}_V(i,j)}{2}$, $\boldsymbol{\Gamma}_V^-(i,j) = \frac{|\boldsymbol{\Gamma}_V(i,j)|-\boldsymbol{\Gamma}_V(i,j)}{2}$, and $\boldsymbol{\Gamma}_V$ is defined as $\boldsymbol{V}^T\boldsymbol{A}^T\boldsymbol{U}\boldsymbol{H} - \boldsymbol{H}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{H} - \lambda_1^v\boldsymbol{V}^T\boldsymbol{G}^v(\boldsymbol{V} - \boldsymbol{V}_0) - \lambda_2^v\boldsymbol{V}^T(\boldsymbol{D}^v - \boldsymbol{W}^v)\boldsymbol{V}$.

Accordingly, we present the Algorithm 1 as a way to optimize the model described by Equation (16). Initially, the algorithm accepts the list-song and external information matrices as input. It then constructs the a priori knowledge relationship matrix and initializes a set of matrices of $\boldsymbol{U}$, $\boldsymbol{H}$, and $\boldsymbol{V}$. Two of these matrices are fixed while iterating over the third matrix using Equations (17)-(19) until convergence is reached. Matrix $\boldsymbol{U}$, which is obtained through factorization, represents the sentiment clustering (recognition) outcome for playlists, whereas matrix $\boldsymbol{V}$ represents the sentiment clustering outcome for songs.

4 THE DATASET

4.1 Data source

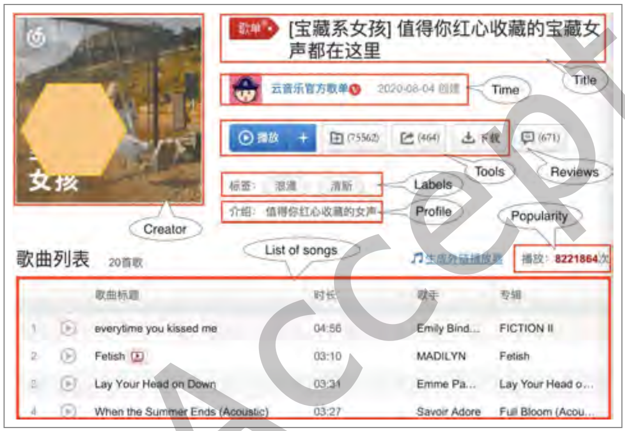
All data utilized in this study were acquired from the digital music social platform, 'NetEase Cloud Music' (http://music.163.com/). Figure 2 illustrates a representative instance of a user-generated playlist on the platform.

We collected 13,565 playlists encompassing 1,621,869 songs. The number of songs on each playlist varied from 1 to 1,000, with an average of 119 songs per playlist. Table 2 presents a summary of the data. Notably, the majority of playlists encompassed multiple genres [54], which posed a challenge for experts in charge of annotating the playlists with sentiment labels.

The songs comprising our dataset are represented by two primary forms of data: audio data in the MP3 file format, and textual data in the form of lyrics and/or profile files. Table 3 presents statistical information regarding the two data modalities.

Algorithm 1 Recognizing Sentiments of Playlists and Songs Based on Matrix Factorization.

```
Input: $\{A, U_0, V_0, W^v; \lambda^u, \lambda_1^v, \lambda_2^v\}$;
Output: Matrix $U$;
Construct matrices $G^u, G^v$;
Construct Laplacian matrix $D^v - W^v$;
Initialize $U, V, H \geq 0$;
while Not convergent do
    Update $H(i, j)$ according to relation (17);
    Update $U(i, j)$ according to relation (18);
    Update $V(i, j)$ according to relation (19);
end while
return $Norm(U)$.
```



歌单 [宝藏系女孩] 值得你红心收藏的宝藏女声都在这里

云音乐官方歌单 2020-08-04 创建

播放 + (75562) (464) 下载 (671)

标签： 浪漫 清新

介绍： 值得你红心收藏的女声

歌曲列表 20首歌

播放： 8221864次

	歌曲标题	时长	歌手	专辑
1	everytime you kissed me	04:56	Emily Bind...	FICTION II
2	Fetish	03:10	MADILYN	Fetish
3	Lay Your Head on Down	03:31	Emme Pa...	Lay Your Head o...
4	When the Summer Ends (Acoustic)	03:27	Savoir Adore	Full Bloom (Acou...

Fig. 2. Example of a playlist in "NetEase Cloud music."

Table 2. Summary of crawled data

Item	Statistics
Total number of playlists	13,565
Total number of songs	1,621,869
Maximum number of songs in the list	1,000
Minimum number of songs in the list	1
Average number of songs in the list	119.56
std	167.23

4.2 Data pre-processing

To identify the sentiment information of songs and playlists from the collected unstructured multimodal music data, three preprocessing stages were required. First, we invited experts to annotate sentiment labels for the sample data. Subsequently, we transformed the audio data into vectors. Finally, we transformed the textual data into vectors.

Table 3. Characteristics n of song data

Audio data	Statistics	Text data	Statistics
Total number of audio data	5,393	Total number of text data	5,290
Maximum audio length (second)	3,600.38	Maximum text length (sentence)	525
Minimum audio length (second)	15.36	Minimum text length (sentence)	1
Average audio length (second)	236.10	Average text length (sentence)	58.40
Standard deviation	145.39	Standard deviation	34.63

4.2.1 Sentiment Label Annotation. The LSRF model was trained in a semi-supervised manner. We asked three experts, one of whom had a qualified background in music education, to manually annotate the sentiment conveyed in a selected number of songs and playlists [90]. This meticulous process was performed to ensure the acquisition of a representative data sample. Each expert individually labeled a group of songs. Because we specifically focused on MIEs, the following labeling guidelines were set: "*after listening to a song (or a list of songs), please report 'positive' if you feel good (joy, happiness, relaxation, excitement, or a similar feeling) or 'negative' if you feel bad (sad, anxious, tired, angry, or a similar feeling); otherwise, please report 'neutral'.*" These guidelines were set because the classification of music MIEs into positive, negative, and neutral categories is intuitive and easy to understand. In addition, this classification scheme relates to the valence dimension of the Russell's circumplex model of affect [68], which is commonly used in MIR research.

Importantly, the challenging nature of annotating the sentiment labels of playlists [39] presents a significant barrier to obtaining sufficient machine learning training samples. We adopted the following strategy to label the sentiment information of sample playlists. First, the experts listened to all songs in the training set independently. Songs in the same playlist were listened to continuously with an interval of less than 45 seconds, whereas intervals of more than 1 min were set between playlists. Secondly, if all three experts agreed on the mood conveyed by a playlist, then the label of the playlist was ascertained and the playlist itself considered 'easy-to-identify.' If the three experts disagreed, the label was determined through discussion and the playlist was considered 'hard-to-identify.' Playlists that could not be evaluated were discarded.

Notably, the playlists within our dataset exhibited varying degrees of quality and required suitable filtration. Because the number of playlist plays can be an indicator of popularity and quality, we used a threshold of 10,000 plays, yielding 1,964 higher-quality playlists. These playlists were subsequently evaluated by the experts, who listened to each song on each list and assigned positive (822), negative (566), or neutral (576) labels based on their overall impressions. In total, these playlists contained 7,909 distinct songs.

Because different sections of a song may evoke disparate emotional responses, the emotional impact of a song cannot be reduced to the sum of its parts, and the focus of this study was on identifying more enduring and prolonged emotions, it was deemed sufficient to simply request the experts to provide their overall sentiment ratings of each song without delving excessively into localized sentiments, such as those of individual clips.

4.2.2 Audio Data Processing. The LSRF generates a d-dimensional vector that characterises the audio information of the music. This process consists of two stages. First, audio features are extracted from the original MP3 files. Second, these features are vectorized.

We used Python LibROSA [62] to extract audio features for data representation. First, we acquired the MP3 audio file of song $s_i \in \mathbb{S}$ and sampled it at a rate of 8kHz to obtain the discrete data sequence of s_i. Subsequently, a frame-splitting operation was conducted [28] on the audio file of the song s_i, with each frame window set at a length of 4096, resulting in a duration of approximately 500 ms per frame. Following this, features were extracted from the framed data of each song, encompassing *root mean square energy*, *zero crossing rate*, *spectral centroid*, *spectral flatness*, *acoustic spectrum attenuation*, *mel-frequency cepstral coefficient*, *chromaticity STFT*, *mel spectrogram*, and more. Notably, 128 *mel-filters* were utilized to produce the *mel-frequency cepstral coefficient* and *mel-spectrogram*. Ultimately, these audio features, excluding the *mel-spectrogram*, were concatenated to form a 57-dimensional *audio-features* vector for each song, with a sample rate of 8 kHz, a frame length of 4096, and a step size of 2048. This resulting 57-dimensional vector serves as a baseline for future comparative experiments.

The *mel-spectrogram* generated by LibROSA is fed directly into a CNN architecture to extract the key features and represent them in a new vector [69], which is utilized in LSRF. The CNN architecture consists of four convolutional layers and two fully connected layers. Each convolutional layer is followed by a Batch Normalisation layer, a Rectified Linear Unit layer, and an average pooling layer. The number of filters in the four convolutional layers are 32, 32, 64, and 128, respectively. The training process utilizes Adam optimization with a batch size of 32. The output of the second fully connected layer is a 200-dimensional representation of the audio data. Notably, the CNN architecture is not trainable with LSRF, as the gradient descent optimization of deep neural networks is not computed in the same way as the multiplicative weight update method used to optimize the LSRF model [20].

4.2.3 Textual Data Processing. The textual data of all the songs in the dataset were employed to train a BERT model, which was subsequently used to generate a d-dimensional textual vector for each song. This process involved two stages: the division of lyrics into sentences, and the vectorization of these sentences.

First, the lyrics of each song were divided into sentences, and the Chinese BERT pre-training model of the Xunfei Joint Laboratory of HAITI [16] was deployed to directly generate sentence vector representations. Subsequently, the average vectors of all sentences for a given song's lyrics were taken as the song's lyrical vector representation. To ensure the consistency of expert annotation, any multilingual lyrics were translated to Chinese.

5 EXPERIMENTAL RESULTS

5.1 Experimental setup

For experimental purposes, both expert-annotated information and similarities between playlists were designated as the sentiment prior for the playlist side. Similarly, both expert-annotated data and similarities between songs were designated as the sentiment prior for the song side. In particular, we empirically set $k = 5$ for Equation (10) and $k = 15$ for Equation (14). Throughout our experiments, the LSRF model employed four parameters set as $\lambda_1^u = \lambda_2^u = \lambda_1^v = \lambda_2^v = 1$. Thus, all sentiment signals were combined with equal weight.

When recognizing the sentiment of a song, the audio features extracted from the song were directly input into each comparison model for training. However, in the case of multimodal data, the data from each modality were vectorized separately and then concatenated as inputs for training. To achieve this, a preprocessing step was required to ensure that the data from each modality were standardized. This approach enables the integration of multiple modalities, such as audio and lyrics, to improve the accuracy of sentiment recognition.

When identifying the sentiments of a playlist, the vectors generated by the data for each song in the list were combined as inputs and used as training data. Although vector concatenation and weighting (averaging) are both commonly used to combine vectors, vector concatenation may face challenges stemming from the potentially large differences in the final vector dimensions caused by the varying number of songs in each playlist. Therefore, we opted to use vector averaging for all songs.

During the experiment, five-fold cross-validation was applied to obtain the results. Specifically, 80% of the data were allocated for training and validation, with the remaining 20% was reserved for testing purposes.

5.2 Baseline methods and evaluation metrics

To validate the performance of the proposed music sentiment recognition model, we conducted a comparative analysis using three categories of baseline models: classical, ensemble learning, and deep learning. Specifically, we selected classical machine learning models including the decision tree (DT), logistic regression (LR), K-means clustering, and support vector machine (SVM) [64, 89]; ensemble methods including the random forests (RF) [21, 42] and AdaBoost

[91]; and deep learning methods such as CNN, long short-term memory (LSTM), and bidirectional LSTM (Bi-LSTM) [22] as comparison models. A summary of the models and their structures is presented in Table 4.

Table 4. Settings for baseline models

Model	Input data (Identifying playlist sentiment)	Input data (Identifying song sentiment)	Model structure (source)
DT	Vectorized playlist data	Vectorized song data	Scikit-learn, [33]
K-means	Vectorized playlist data	Vectorized song data	Scikit learn, [31]
LR	Vectorized playlist data	Vectorized song data	Scikit-learn, [24]
SVM	Vectorized playlist data	Vectorized song data	Scikit-learn, linear kernal, [9, 64, 89]
RF	Vectorized playlist data	Vectorized song data	Scikit-learn, [21, 42]
Adaboost	Vectorized playlist data	Vectorized song data	Scikit-learn, [32, 91]
CNN	Matrix[♯]	Matrix[†]	4*CNN+2*FC[‡]+ softmax
LSTM	Matrix[♯]	Matrix[†]	LSTM+2*FC[‡]+ softmax, [25, 26]
Bi-LSTM	Matrix[♯]	Matrix[†]	BiLSTM+2*FC[‡]+ softmax, [25, 26]

[♯] The matrix comprises all vectors associated with all songs on the playlist;

[†] Each musical composition is divided into clips of 30 seconds in duration [38, 64]. These clips are then vectorised;

[‡] FC = fully connected layer. This is the standard operation of deep classifiers [4].

We assessed each model's performance utilizing the metrics commonly employed in multi-classification tasks [60], namely accuracy (F1-micro), precision, recall, and F1-macro of the models.

5.3 Performance evaluation

5.3.1 Performance comparison: feature-extraction-based data representation. Table 5 presents the performance results of each model in identifying sentiments in songs and playlists. By comparing the results of the various methods, we made the following observations:

- The LSRF method consistently outperformed the baselines for both song and playlist sentiment recognition tasks based on data representation obtained through audio feature extraction.
- The LSRF method demonstrated superior performance in recognizing sentiments in songs, indicating that the constraints of graph relations established from song features can effectively facilitate the clustering of similar songs.
- The LSRF method produces superior results in the task of sentiment recognition in playlists, indicating that the NMF-based model captures certain sentiment-related correlation information, such as co-occurrence between songs.

5.3.2 Performance comparison: multimodality-vector-based data representation. Table 6 summarizes the performance of each model in recognizing sentiments in playlists using a multimodal vector-based data representation. Upon examination of these results, it is evident that:

- The use of deep learning models, such as CNN and BERT, to represent multimodal music data as vectors results in substantial efficiency gains for all algorithms. Specifically, traditional machine learning algorithms and ensemble methods exhibited significant improvements. This underscores the effectiveness of deep learning models in the preprocessing of multimedia data.
- The LSRF clearly outperformed the baselines in nearly all aspects. In particular, it excelled in recognizing the sentiments of playlists, outperforming all other algorithms. In the song sentiment experiment, although the RF method achieved the highest accuracy, the LSRF was close behind with an insignificant difference in performance.

Table 5. Performance of models trained with audio feature vectors

	Song sentiment recognition				Playlist sentiment recognition			
Model	Accuracy	Precision	Recall	F1-macro	Accuracy	Precision	Recall	F1-macro
DT	0.4979	0.4987	0.4605	0.4571	0.6801	0.6813	0.6750	0.6677
K-means	0.5173	0.5018	0.5102	0.4997	0.6694	0.6782	0.6618	0.6573
LR	0.5392	0.5357	0.5105	0.5148	0.7066	0.7030	0.6928	0.6921
SVM	0.5435	0.5561	0.5121	0.5156	0.7107	0.7166	0.6986	0.6957
Adaboost	0.4903	0.4779	0.4651	0.4664	0.6796	0.6731	0.6682	0.6658
RF	0.5392	0.5357	0.5105	0.5148	0.7219	0.7108	0.7097	0.7083
CNN	0.5475	0.5245	0.5340	0.5263	0.6831	0.6681	0.6728	0.6660
LSTM	0.5650	0.5362	0.5453	0.5365	0.7015	0.6880	0.6920	0.6870
Bi-LSTM	0.5692	0.5415	0.5673	0.5471	0.7144	0.7015	0.7051	0.7001
LSRF	**0.6208**	**0.6241**	**0.6241**	**0.6103**	**0.7270**	**0.7234**	**0.7155**	**0.7127**

Table 6. Performance of models trained with multimodal vectors

	Song sentiment recognition				Playlist sentiment recognition			
Model	Accuracy	Precision	Recall	F1-macro	Accuracy	Precision	Recall	F1-macro
DT	0.5795	0.5492	0.5415	0.5423	0.6908	0.6830	0.6757	0.6758
K-means	0.5622	0.5556	0.5595	0.5505	0.6255	0.6256	0.6244	0.6188
LR	0.5969	0.5750	0.5772	0.5733	0.7337	0.7230	0.7217	0.7217
SVM	0.5669	0.5511	0.5503	0.5460	0.7214	0.7099	0.7115	0.7099
Adaboost	0.6079	0.5817	0.5731	0.5732	0.6969	0.6813	0.6818	0.6807
RF	**0.6512**	0.6320	0.6045	0.6049	0.7337	0.7236	0.7215	0.7214
CNN	0.5568	0.5452	0.5441	0.5417	0.6769	0.6733	0.6706	0.6658
LSTM	0.5630	0.5594	0.5577	0.5562	0.7323	0.7218	0.7206	0.7200
Bi-LSTM	0.5688	0.5735	0.5643	0.5606	0.7251	0.7103	0.7121	0.7101
LSRF	0.6423	**0.6352**	**0.6378**	**0.6275**	**0.7378**	**0.7287**	**0.7288**	**0.7253**

5.4 Discussion

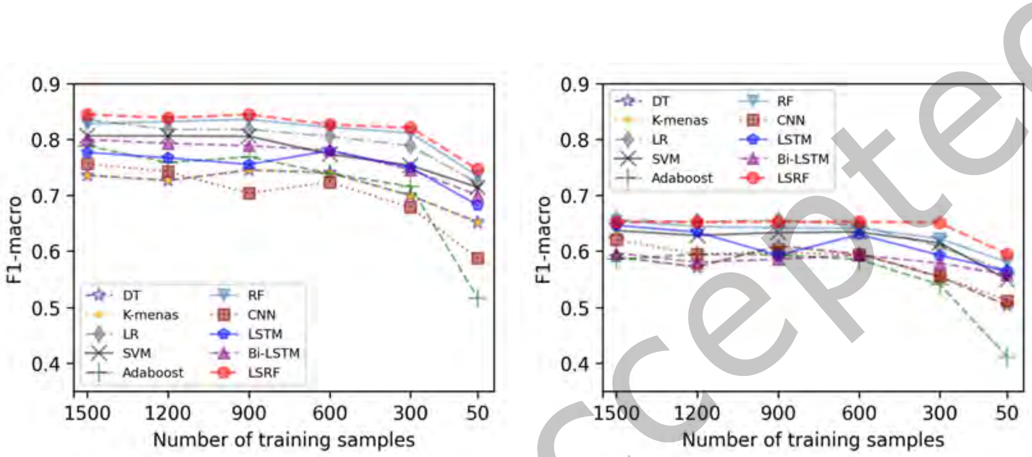
We conducted several experiments to investigate the impact of various factors in the effectiveness of the LSRF model. The primary focus was on expert annotation, value of α, and side information. The relationship between songs and playlist sentiments was also examined.

5.4.1 Influence of expert annotation. Additional experiments were conducted to investigate the effectiveness of models trained using fewer playlist samples. In these experiments, we categorized expert-labeled sample playlists into two distinct groups: easy-to-identify and hard-to-identify.

The experimental design was set as follows. First, we selected two test datasets of easy- and hard-to-identify playlists, respectively, each comprising 200 playlists, to ensure comparative consistency. Next, we randomly selected samples from expert-labeled playlists with capacities of 1500, 1200, 900, 600, 300, and 50 as the training set. Each round of experiments was conducted five times, with five datasets randomly selected to train the models at each sample capacity. Finally, the average experimental results were reported for each model.

It is worth noting that the input data for each model in the experiment were the multimodal representation vectors of songs. The experimental results shown in Figure 3 indicate the following:

- By comparing the results shown in Figure 3(a) and 3(b), it is clear that the proposed LSRF model outperforms all baselines. Additionally, the performance of each method improved as the number of expert-labeled training samples increased. Nevertheless, it is noteworthy that all methods exhibited considerably lower performance on the test data for hard-to-identify playlists;
- In the sample set of easy-to-identify playlists (see Figure 3(a)), the LSRF method exhibited the most outstanding performance, closely followed by the RF method, which performed exceptionally well as the sample size increased. Notably, the performance of the LSRF remained consistently stable even as the sample size decreased, with the least variation in the F1-macro values.
- In the sample set of hard-to-identify playlists (see figure 3(b)), the LSRF outperformed all other algorithms while exhibiting consistent stability.



(a) Easy-to-identify playlists

(b) Hard-to-identify playlists

Fig. 3. Performance of models on small playlist samples.

5.4.2 Influence of the value of α. Equation (9) suggests that the LSRF must balance the contributions of both textual and audio data when recognizing music sentiment. We therefore conducted experiments to investigate the effect of α in Equation (9) on LSRF performance, with results presented in Table 7.

Table 7. Effect of α on LSRF performance

α	0	0.1	0.2	0.3	0.4	0.5
Accuracy	0.7270	0.7342	0.7342	**0.7362**	0.7332	0.7306
P_{Macro}	0.7166	0.7222	0.7224	**0.7246**	0.7210	0.7184
R_{Macro}	0.7178	0.7235	0.7250	**0.7268**	0.7232	0.7198
F1-macro	0.7136	0.7204	0.7214	**0.7234**	0.7201	0.7169

We can observe that the performance of the LSRF improved as the proportion of lyrics in the concatenation vector increased, until it reaches a peak at $\alpha = 0.3$. Beyond this point, performance began to decrease. This suggests that incorporating lyrical information is beneficial for improving the accuracy of the LSRF in sentiment recognition.

Nonetheless, audio data play a more significant role in identifying the sentiment of a playlist. Therefore, we set α to the optimal value of 0.3.

5.4.3 Effects of the side information. We conducted further experiments to compare the LSRF model's performance in recognizing the sentiments of songs and playlists after removing one, two, three, or four types of side information. To achieve this, we removed a priori sentimental information from the LSRF framework by setting the values of the corresponding parameters λ_1^u, λ_1^v, λ_2^u, and λ_2^v to 0.

Table 8. Efficiency of Sentiment Recognition in Songs with Various Types of Side Information

Row	Strategy	λ_1^u	λ_1^v	λ_2^u	λ_2^v	F1-macro (loss)
1	Default	1	1	1	1	0.6275 (-)
2	Knock out one side information	0	1	1	1	0.6107(-2.67%)
3		1	0	1	1	0.6225(-0.79%)
4		1	1	0	1	0.6027(-3.95%)
5		1	1	1	0	0.6222(-0.84%)
6	Knock out two side information	0	0	1	1	0.5922(-5.63%)
7		0	1	0	1	0.6066(-3.33%)
8		0	1	1	0	0.6115(-2.55%)
9		1	0	0	1	0.6169(-1.69%)
10		1	0	1	0	0.6231(-0.70%)
11		1	1	0	0	0.5947(-5.23%)
12	Knock out three side information	0	0	0	1	0.5758(-8.25%)
13		0	0	1	0	0.5925(-5.58%)
14		0	1	0	0	0.6041(-3.73%)
15		1	0	0	0	0.6184(-1.45%)
16	Knock out four side information	0	0	0	0	0.5732 (-8.65%)

The results presented in Table 8 correspond to the effects of the four types of prior information (constraints) on the effectiveness of song-based sentiment recognition. The third column lists parameter settings and the last column lists the F1-macro values. The values in brackets indicates the percentage decrease in LSRF performance at the current parameter setting compared to the scenario in which all constraints are present. From the table, we can make the following observations:

- The performance of the LSRF model in recognizing song sentiments decreased when any a priori information was removed.
- When determining the sentimental content of a song, the correlation between playlists (Row 4) and expert annotations of playlists (Row 2) had a significant impact on the performance of LSRF.
- Combining the expert annotation of songs with their playlist annotation (Row 6) or combining the sentiment relevance of songs with the sentiment relationship of playlists (Row 11) has a significant impact on model performance.

Table 9 presents the impacts of the four types of prior information (constraints) on the effectiveness of playlist sentiment recognition. First, all four types provide essential contributions for identifying playlist sentiments (Rows 12-16). Second, as the LSRF represents a typical semi-supervised learning method, the expert playlist annotations are crucial in ensuring satisfactory performance (Rows 6 and 8).

Interestingly, the correlation between playlists (Rows 4 and 15) and the results of expert annotation for song sentiment (Rows 3 and 9) had a relatively minor impact on model performance in identifying playlist sentiment. This suggests that the sentiment content of a playlist is generally determined by the combination of all constituent songs, rather than a simple presentation of the individual sentiment of each song. Therefore, the sentiment information of a playlist cannot be determined solely based on the sentiment content of its individual songs.

Table 9. Efficiency of Sentiment Recognition in Playlists with Various Types of Side Information

Row	Strategy	λ_1^u	λ_1^v	λ_2^u	λ_2^v	F1-macro (loss)
1	Default	1	1	1	1	0.7253 (-)
2	Knock out one side information	0	1	1	1	0.7023(-3.17%)
3		1	0	1	1	0.7135(-1.62%)
4		1	1	0	1	0.7188(-0.89%)
5		1	1	1	0	0.7122(-1.80%)
6	Knock out two side information	0	0	1	1	0.6837(-5.73%)
7		0	1	0	1	0.7062(-2.62%)
8		0	1	1	0	0.6902(-4.83%)
9		1	0	0	1	0.7174(-1.08%)
10		1	0	1	0	0.7110(-1.97%)
11		1	1	0	0	0.7174(-1.09%)
12	Knock out three side information	0	0	0	1	0.6825(-5.89%)
13		0	0	1	0	0.6603(-8.96%)
14		0	1	0	0	0.6836(-5.75%)
15		1	0	0	0	0.7048(-2.82%)
16	Knock out four side information	0	0	0	0	0.6544(-9.77%)

5.4.4 Summed-song-sentiment vs. playlist-sentiment. When presented with a list of songs, listeners have the option to express their MIE in two distinct manners: they can either assign a specific MIE score to each individual song after listening to it, with these individual scores contributing to a cumulative *summed-song-sentiment*, or they can provide an overall MIE rating for the entire playlist (i.e., *playlist-sentiment*) after listening to all the songs in the list.

The following experiments were conducted to investigate the differences between the *summed-song-sentiment* and *playlist-sentiment*. Because the RF achieved highest accuracy in recognizing song sentiments among the baselines, it was selected alongside the LSRF model. First, each of the models identified the sentiment of each song. Second, the model identified the sentiment of the playlist. Finally, the *summed-song-sentiment* of each playlist and the identified overall sentiment of each playlist - i.e., *playlist-sentiment* - were compared with the expert annotations. The results are listed in Table 10.

Table 10. Summed-song-sentiment vs. playlist-sentiment

	Accuracy	Precision	Recall	F1-macro
RF (summed-song-sentiment)	0.6634	0.6929	0.6318	0.6256
RF (playlist-sentiment)	0.7337	0.7236	0.7215	**0.7214**
LSRF (summed-song-sentiment)	0.7347	0.7427	0.7290	0.7248
LSRF (playlist-sentiment)	0.7378	0.7287	0.7288	**0.7253**

The results in terms of F1-macro indicate that both RF and LSRF are more effective in recognizing the sentiments of playlists as a whole than using the aggregated sentiments of individual songs in the list. Furthermore, it is evident that the enhancement in the outcomes of comparative experiments conducted using the RF algorithm was more pronounced than that of the LSRF. This can be attributed to the fact that in the calculation of *summed-song-sentiment*, RF utilizes only the vectorized data of individual songs for sentiment prediction, whereas the LSRF algorithm employs the data of individual songs within the $\boldsymbol{V}$-matrix. It is clear that the $\boldsymbol{V}$-matrix acquires some of the “song-list” correlation information during the matrix decomposition process. This analysis demonstrates that a sentiment expressed in a playlist is not simply a sum of the sentiments expressed in the individual songs.

6 CONCLUSION

In this study, we developed the LSRF model to address the challenge of identifying the sentiments of songs as well as playlists. To achieve this, a significant number of songs were initially annotated by domain experts as training samples. The multimodal audio and textual information were then utilized to compute the “song-to-song” and the “playlist-to-playlist” similarities. Finally, a non-negative matrix factorization-based model was developed, emphasizing the co-occurrence of songs in playlists, as well as enabling the co-clustering of songs and playlists.

The experimental results demonstrate that the LSRF can simultaneously recognize sentiment information in both songs and playlists. Moreover, by utilizing the relationship between songs and playlists, as well as incorporating relevant side information constraints, the LSRF achieved remarkably efficient and consistent performance in recognizing playlist sentiments, even given a limited number of annotated samples.

ACKNOWLEDGMENTS

This work was partly supported by the National Natural Science Foundation of China (91846105, 71671027, 72172092, 71772017 and 72293571), the Natural Science Foundation of Sichuan Province (2024NSFSC0273), Innovative Research Team of Shanghai International Studies University (2020114044), and the Fundamental Research Funds for the Central Universities (2019114032).

REFERENCES

[1] S. Omar Ali and Zehra F. Peynircioğlu. 2006. Songs and emotions: are lyrics and melodies equal partners? *Psychology of Music* 34, 4 (2006), 511–534.

[2] Anna Aljanaki, Frans Wiering, and Remco C. Veltkamp. 2016. Studying emotion induced by music through a crowdsourcing game. *Information Processing & Management* 52, 1 (2016), 115–128.

[3] Marcos Alves de Almeida, Carolina Coimbra Vieira, Pedro Olmo Stancioli Vaz De Melo, and Renato Martins Assunção. 2019. Random Playlists Smoothly Commuting Between Styles. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 4, Article 104 (2019), 20 pages.

[4] S.H. Shabbeer Basha, Shiv Ram Dubey, Viswanath Pulabaigari, and Snehasis Mukherjee. 2020. Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing* 378 (2020), 112–119.

[5] Geoffray Bonnin and Dietmar Jannach. 2014. Automated Generation of Music Playlists: Survey and Experiments. *ACM Comput. Surv.* 47, 2, Article 26 (nov 2014), 35 pages.

[6] Gordon C Bruner. 1990. Music, mood, and marketing. *Journal of marketing* 54, 4 (1990), 94–104.

[7] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S. Huang. 2011. Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 8 (2011), 1548–1560.

[8] Erion Çano and Maurizio Morisio. 2017. MoodyLyrics: A Sentiment Annotated Lyrics Dataset. In *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI ’17)*. ACM, 118–124.

[9] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 1–27.

[10] Vybhav Chaturvedi, Arman Beer Kaur, Vedansh Varshney, Anupam Garg, Gurpal Singh Chhabra, and Munish Kumar. 2022. Music mood and human emotion recognition based on physiological signals: a systematic review. *Multimedia Systems* 28, 1 (2022), 21–44.

[11] Yufu Chen, Zhiqi Lei, Yanghui Rao, Haoran Xie, Fu Lee Wang, Jian Yin, and Qing Li. 2023. Parallel Non-Negative Matrix Tri-Factorization for Text Data Co-Clustering. *IEEE Transactions on Knowledge and Data Engineering* 35, 5 (2023), 5132–5146.

[12] Zhiyong Cheng, Jialie Shen, Lei Zhu, Mohan S Kankanhalli, and Liqiang Nie. 2017. Exploiting Music Play Sequence for Music Recommendation.. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (Melbourne, Australia), Vol. 17. AAAI Press, 3654–3660.

[13] Fan RK Chung and Fan Chung Graham. 1997. *Spectral graph theory*. Vol. 92. American Mathematical Soc.

[14] Michael Clark, Gloria Isaacks-Downton, Nancy Wells, Sheryl Redlin-Frazier, Carol Eck, Joseph T. Hepworth, and Bapsi Chakravarthy. 2006. Use of Preferred Music to Reduce Emotional Distress and Symptom Activity During Radiation Therapy. *Journal of Music Therapy* 43, 3 (2006), 247–265.

[15] John R Crawford and Julie D Henry. 2004. The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British journal of clinical psychology* 43, 3 (2004), 245–265.

[16] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-Training With Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3504–3514.

[17] Martina de Witte, Ana da Silva Pinho, Geert-Jan Stams, Xavier Moonen, Arjan E.R. Bos, and Susan van Hooren. 2022. Music therapy for stress reduction: a systematic review and meta-analysis. *Health Psychology Review* 16, 1 (2022), 134–159.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, Minneapolis, Minnesota, 4171–4186.

[19] Ulf Dimberg, Monika Thunberg, and Sara Grunedal. 2002. Facial reactions to emotional stimuli: Automatically controlled emotional responses. *Cognition & Emotion* 16, 4 (2002), 449–471.

[20] Chris Ding, Tao Li, Wei Peng, and Haesun Park. 2006. Orthogonal Nonnegative Matrix T-Factorizations for Clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Philadelphia, PA, USA) *(KDD '06)*. ACM, 126–135.

[21] J.A. Domínguez-Jiménez, K.C. Campo-Landines, J.C. Martínez-Santos, E.J. Delahoz, and S.H. Contreras-Ortiz. 2020. A machine learning model for emotion recognition from physiological signals. *Biomedical Signal Processing and Control* 55 (2020), 101646.

[22] Yizhuo Dong, Xinyu Yang, Xi Zhao, and Juan Li. 2019. Bidirectional Convolutional Recurrent Sparse Network (BCRSN): An Efficient Model for Music Emotion Recognition. *IEEE Transactions on Multimedia* 21, 12 (2019), 3150–3163.

[23] Rui Duan, Cuiqing Jiang, and Hemant K. Jain. 2022. Combining review-based collaborative filtering and matrix factorization: A solution to rating's sparsity problem. *Decision Support Systems* 156 (2022), 113748.

[24] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *the Journal of machine Learning research* 9 (2008), 1871–1874.

[25] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks* 18, 5-6 (2005), 602–610.

[26] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems* 28, 10 (2016), 2222–2232.

[27] Quanquan Gu and Jie Zhou. 2009. Co-Clustering on Manifolds. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Paris, France) *(KDD '09)*. ACM, 359–368.

[28] Byeong-jun Han, Seungmin Rho, Sanghoon Jun, and Eenjun Hwang. 2010. Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications* 47 (2010), 433–460.

[29] Donghong Han, Yanru Kong, Jiayi Han, and Guoren Wang. 2022. A survey of music emotion recognition. *Frontiers of Computer Science* 16, 6 (2022), 166335.

[30] Emilia Parada-Cabaleiro Harald Victor Schweiger and Markus Schedl. 2021. Does Track Sequence in User-generated Playlists Matter?. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR'21)*. 618–625.

[31] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.

[32] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. 2009. Multi-class adaboost. *Statistics and its Interface* 2, 3 (2009), 349–360.

[33] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

[34] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[35] Kate Hevner. 1936. Experimental studies of the elements of expression in music. *The American journal of psychology* 48, 2 (1936), 246–268.

[36] M. Shamim Hossain and Ghulam Muhammad. 2019. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion* 49 (2019), 69–78.

[37] Changwei Hu, Piyush Rai, and Lawrence Carin. 2016. Non-negative Matrix Factorization for Discrete Data with Hierarchical Side-Information. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, Arthur Gretton and Christian C. Robert (Eds.), Vol. 51. 1124–1132.

[38] Xiao Hu, Kahyun Choi, and J. Stephen Downie. 2017. A framework for evaluating multimodal music mood classification. *Journal of the Association for Information Science and Technology* 68, 2 (2017), 273–285.

[39] Xiao Hu and Noriko Kando. 2017. Task complexity and difficulty in music information retrieval. *Journal of the Association for Information Science and Technology* 68, 7 (2017), 1711–1723.

[40] Xiao Hu, Fanjie Li, and Tzi-Dong Jeremy Ng. 2018. On the Relationships between Music-induced Emotion and Physiological Signals. In *19th International Society for Music Information Retrieval Conference*. 362–369.

[41] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. 2013. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*. 607–618.

[42] Theodoros Iliou and Christos-Nikolaos Anagnostopoulos. 2009. Comparison of Different Classifiers for Emotion Recognition. In *2009 13th Panhellenic Conference on Informatics*. 102–106.

[43] Abhishek Iyer, Srimit Sritik Das, Reva Teotia, Shishir Maheshwari, and Rishi Raj Sharma. 2023. CNN and LSTM based ensemble learning for human emotion recognition using EEG recordings. *Multimedia Tools and Applications* 82, 4 (2023), 4883–4896.

[44] Patrik N. Juslin. 2013. What does music express? Basic emotions and beyond. *Frontiers in Psychology* 4 (2013), 1–14.

[45] Patrik N. Juslin, László Harmat, and Tuomas Eerola. 2014. What makes music emotionally significant? Exploring the underlying mechanisms. *Psychology of Music* 42, 4 (2014), 599–623.

[46] Patrik N. Juslin and Petri Laukka. 2004. Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening. *Journal of New Music Research* 33, 3 (2004), 217–238.

[47] Patrik N. Juslin, Simon Liljeström, Daniel Västfjäll, Gonçalo Barradas, and Ana Silva. 2008. An experience sampling study of emotional reactions to music: Listener, music, and situation. *Emotion* 8, 5 (2008), 668–683.

[48] Patrik N. Juslin and Daniel Västfjäll. 2008. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences* 31, 5 (2008), 559–575.

[49] Iman Kamehkhosh, Geoffray Bonnin, and Dietmar Jannach. 2020. Effects of recommendations on the playlist creation behavior of users. *User Modeling and User-Adapted Interaction* 30, 2 (2020), 285–322.

[50] Hamid Khabiri, Mohammad Naseh Talebi, Mehdi Fakhimi Kamran, Shadi Akbari, Farzaneh Zarrin, and Fatemeh Mohandesi. 2024. Music-induced emotion recognition based on feature reduction using PCA from EEG signals. *Frontiers in Biomedical Technologies* 11, 1 (2024), 59–68.

[51] Seon Tae Kim and Joo Hee Oh. 2021. Music intelligence: Granular data and prediction of top ten hit songs. *Decision Support Systems* 145 (2021), 113535.

[52] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. 2010. Music emotion recognition: A state of the art review. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, Vol. 86. 937–952.

[53] Stefan Koelsch, Thomas Fritz, D Yves v. Cramon, Karsten Müller, and Angela D Friederici. 2006. Investigating emotion with music: an fMRI study. *Human brain mapping* 27, 3 (2006), 239–250.

[54] Deborah Lee, Lyn Robinson, and David Bawden. 2021. Orthogonality, dependency, and music: An exploration of the relationships between music facets. *Journal of the Association for Information Science and Technology* 72, 5 (2021), 570–582.

[55] Alexander W. Legge. 2015. On the Neural Mechanisms of Music Therapy in Mental Health Care: Literature Review and Clinical Implications. *Music Therapy Perspectives* 33, 2 (2015), 128–141.

[56] Dan Li, Tong Xu, Peilun Zhou, Weidong He, Yanbin Hao, Yi Zheng, and Enhong Chen. 2021. Social Context-Aware Person Search in Videos via Multi-Modal Cues. *ACM Trans. Inf. Syst.* 40, 3, Article 52 (2021), 25 pages.

[57] Tao Li, Yi Zhang, and Vikas Sindhwani. 2009. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 244–252.

[58] Elad Liebman, Maytal Saar-Tsechansky, and Peter Stone. 2019. The right music at the right time: Adaptive personalized playlists based on sequence modeling. *MIS quarterly* 43, 3 (2019), 765–786.

[59] Chuan Ma, Yingwei Zhang, and Chun-Yi Su. 2023. Graph-Based Multicentroid Nonnegative Matrix Factorization. *IEEE Transactions on Neural Networks and Learning Systems* (2023), 1–12. https://doi.org/10.1109/TNNLS.2023.3332360

[60] Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. 2012. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition* 45, 9 (2012), 3084–3104.

[61] Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. 2023. Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 3, Article 74 (2023), 34 pages.

[62] Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*. SciPy, 18–24.

[63] Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Transactions on Affective Computing* 5, 2 (2014), 101–111.

[64] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. 2020. Novel Audio Features for Music Emotion Recognition. *IEEE Transactions on Affective Computing* 11, 4 (2020), 614–626.

[65] Renato Panda, Ricardo Manuel Malheiro, and Rui Pedro Paiva. 2023. Audio Features for Music Emotion Recognition: a Survey. *IEEE Transactions on Affective Computing* 14, 1 (2023), 68–88.

[66] Maxime Résibois, Elise K Kalokerinos, Gregory Verleysen, Peter Kuppens, Iven Van Mechelen, Philippe Fossati, and Philippe Verduyn. 2018. The relation between rumination and temporal features of emotion intensity. *Cognition and Emotion* 32, 2 (2018), 259–274.

[67] Fabiana Silva Ribeiro, Flávia Heloísa Santos, Pedro Barbas Albuquerque, and Patrícia Oliveira-Silva. 2019. Emotional Induction Through Music: Measuring Cardiac and Electrodermal Responses of Emotional States and Their Persistence. *Frontiers in Psychology* 10, Article 451 (2019), 13 pages.

[68] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.

[69] Mladen Russo, Luka Kraljević, Maja Stella, and Marjan Sikora. 2020. Cochleogram-based approach for detecting perceived emotions in music. *Information Processing & Management* 57, 5 (2020), 102270.

[70] Markus Schedl, Tim Pohle, Peter Knees, and Gerhard Widmer. 2011. Exploring the Music Similarity Space on the Web. *ACM Trans. Inf. Syst.* 29, 3, Article 14 (2011), 24 pages.

[71] Yading Song, Simon Dixon, and Marcus T Pearce. 2012. Evaluation of musical features for emotion classification.. In *ISMIR*. 523–528.

[72] Dan Su, Pascale Fung, and Nicolas Auguin. 2013. Multimodal music emotion classification using AdaBoost with decision stumps. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 3447–3451.

[73] Anu Taneja and Anuja Arora. 2018. Cross domain recommendation using multidimensional tensor factorization. *Expert Systems with Applications* 92 (2018), 304–316.

[74] Peter C Terry, Costas I Karageorghis, Michelle L Curran, Olwenn V Martin, and Renée L Parsons-Smith. 2020. Effects of music in exercise and sport: A meta-analytic review. *Psychological bulletin* 146, 2 (2020), 91.

[75] Nattapong Thammasan, Ken-ichi Fukui, and Masayuki Numao. 2017. Multimodal Fusion of EEG and Musical Features in Music-Emotion Recognition. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, California, USA) *(AAAI'17)*. AAAI Press, 4991–4992.

[76] CG Tsai. 2013. *The cognitive psychology of Music.* National Taiwan University Press, Taipei.

[77] Andreu Vall, Matthias Dorfer, Markus Schedl, and Gerhard Widmer. 2018. A Hybrid Approach to Music Playlist Continuation Based on Playlist-Song Membership. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (Pau, France) *(SAC '18)*. 1374–1382.

[78] Jonna K. Vuoskoski and Tuomas Eerola. 2011. Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. *Musicae Scientiae* 15, 2 (2011), 159–173.

[79] Jingyao Wang, Luntian Mou, Lei Ma, Tiejun Huang, and Wen Gao. 2023. AMSA: Adaptive Multimodal Learning for Sentiment Analysis. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 3s, Article 135 (2023), 21 pages.

[80] Ju-Chiang Wang, Yi-Hsuan Yang, Hsin-Min Wang, and Shyh-Kang Jeng. 2015. Modeling the Affective Content of Music with a Gaussian Mixture Model. *IEEE Transactions on Affective Computing* 6, 1 (2015), 56–68.

[81] Xing Wang, Xiaoou Chen, Deshun Yang, and Yuqian Wu. 2011. Music Emotion Classification of Chinese Songs based on Lyrics Using TF*IDF and Rhyme. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*. ISMIR, Miami, United States, 765–770.

[82] Xinxi Wang, David Rosenblum, and Ye Wang. 2012. Context-Aware Mobile Music Recommendation for Daily Activities. In *Proceedings of the 20th ACM International Conference on Multimedia (MM '12)*. ACM, 99–108.

[83] Yang Wang. 2021. Survey on Deep Multi-modal Data Analytics: Collaboration, Rivalry, and Fusion. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 1s, Article 10 (2021), 25 pages.

[84] Yu-Xiong Wang and Yu-Jin Zhang. 2013. Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Transactions on Knowledge and Data Engineering* 25, 6 (2013), 1336–1353.

[85] Felix Weninger, Florian Eyben, and Björn Schuller. 2014. On-line continuous-time music mood regression with deep recurrent neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5412–5416.

[86] Wenhui Wu, Sam Kwong, Junhui Hou, Yuheng Jia, and Horace Ho Shing Ip. 2019. Simultaneous Dimensionality Reduction and Classification via Dual Embedding Regularized Nonnegative Matrix Factorization. *IEEE Transactions on Image Processing* 28, 8 (2019), 3836–3847.

[87] Liang Xu, Xin Wen, Jiaming Shi, Shutong Li, Yuhan Xiao, Qun Wan, and Xiuying Qian. 2021. Effects of individual factors on perceived emotion and felt emotion of music: Based on machine learning methods. *Psychology of Music* 49, 5 (2021), 1069–1087.

[88] Dan Yang and Won-Sook Lee. 2009. Music Emotion Identification from Lyrics. In *2009 11th IEEE International Symposium on Multimedia*. 624–629.

[89] Xinyu Yang, Yizhuo Dong, and Juan Li. 2018. Review of data features-based music emotion recognition methods. *Multimedia Systems* 24, 4 (2018), 365–389.

[90] Yi-Hsuan Yang and Homer H. Chen. 2012. Machine Recognition of Music Emotion: A Review. *ACM Trans. Intell. Syst. Technol.* 3, 3, Article 40 (2012), 30 pages.

[91] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and H. H. Chen. 2008. A Regression Approach to Music Emotion Recognition. *Trans. Audio, Speech and Lang. Proc.* 16, 2 (2008), 448–457.

[92] Guanghao Yin, Shouqian Sun, Dian Yu, Dejian Li, and Kejun Zhang. 2022. A Multimodal Framework for Large-Scale Emotion Recognition by Fusing Music and Electrodermal Activity Signals. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 3, Article 78 (2022), 23 pages.

[93] Jiho Yoo and Seungjin Choi. 2010. Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on Stiefel manifolds. *Information Processing & Management* 46, 5 (2010), 559–570.

[94] Yi Yu, Suhua Tang, Francisco Raposo, and Lei Chen. 2019. Deep Cross-Modal Correlation Learning for Audio and Lyrics in Music Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1, Article 20 (2019), 16 pages.

[95] Marcel Zentner and Tuomas Eerola. 2010. Self-Report Measures and Models. In *Handbook of Music and Emotion: Theory, Research,Applications.* Oxford University Press, 187–221.

[96] M. Zentner, D. Grandjean, and K. R. Scherer. 2008. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion* 8, 4 (2008), 494–521.

[97] Sicheng Zhao, Yaxian Li, Xingxu Yao, Weizhi Nie, Pengfei Xu, Jufeng Yang, and Kurt Keutzer. 2020. Emotion-Based End-to-End Matching Between Image and Music in Valence-Arousal Space. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20).* New York, NY, USA, 2945–2954.

[98] Sicheng Zhao, Shangfei Wang, Mohammad Soleymani, Dhiraj Joshi, and Qiang Ji. 2019. Affective Computing for Large-Scale Heterogeneous Multimedia Data: A Survey. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 3s, Article 93 (2019), 32 pages.

Received 18 December 2023; revised 15 November 2024; accepted 18 November 2024