Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 4939048, 12 pages
https://doi.org/10.1155/2023/4939048

Research Article

Real-Time Human-Music Emotional Interaction Based on Deep Learning and Multimodal Sentiment Analysis

Tianyue Jiang,[1,2] **Sanhong Deng**,[1,2] **Peng Wu**,[3] **and Haibi Jiang**[4]

[1]*School of Information Management, Nanjing University, Nanjing 210023, China*
[2]*International Joint Informatics Laboratory, Nanjing University, Jiangsu 210023, China*
[3]*School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China*
[4]*School of Materials Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*

Correspondence should be addressed to Sanhong Deng; sanhong@nju.edu.cn

Received 15 September 2022; Revised 16 October 2022; Accepted 29 October 2022; Published 14 April 2023

Academic Editor: Li You

Music, as an integral component of culture, holds a prominent position and is widely accessible. There has been growing interest in studying sentiment represented by music and its emotional effects on its audiences, however, much of the existing literature is subjective and overlooks the impact of music on the real-time expression of emotion. In this article, two labeled datasets for music sentiment classification and multimodal sentiment classification were developed. Deep learning is used to classify music sentiment, while decision-level fusion is used to classify the multimodal sentiment of real-time listeners. We combine sentiment analysis with a conventional online music playback system and propose an innovative human-music emotional interaction system based on multimodal sentiment analysis and deep learning. It has been demonstrated through individual observation and questionnaire studies that the interaction between human and musical sentiments has a positive impact on the negative emotions of listeners.

1. Introduction

The growth of music streaming media platforms has accelerated along with advancements in computer technology and Internet applications, and user bases have grown significantly. In 2020, the global recorded music revenue totalled $21.6 billion, of which the streaming media revenue accounted for 62.1% [1]. The number of downstream end users in China's digital music industry has increased from 580 million to 980 million over the past five years at a compound annual growth rate of 9.3% [2]. As a typical representative of culture and art, it is evident that music has become an integral part of daily life. We believe that music can not only express the sentiments of creators but also have a significant impact on the emotions of listeners, notably considering the increasing number of monthly active users of NetEase Cloud Music, the time tendency, comment sentiment trend of users [3], and the younger age of users. Therefore, it has become a new area of study to determine how to affect human sentiments by quantifying the fuzzy emotional colour of music into musical emotional characteristics and classifying them.

The greatest obstacle in the study of the emotional interaction between music and listeners is the difficulty of quantifying the sentiment of both music and listeners over time and demonstrating the process of listener sentiment change caused by music playing through specific indicators. While numerous prior studies have demonstrated that music can regulate and control people's negative emotions to a certain extent [4], most of them employ the fuzzy comprehensive evaluation method to carry out literal research. On the one hand, there is a lack of effective quantitative indicators for music and users' sentiments, and on the other, it is challenging to obtain effective statistics on emotional changes among users. Even though studies in the medical and physiological fields have shown that changes in listeners' sentiments can be observed through their physiological indicators, these studies have been conducted under controlled laboratory conditions with small sample sizes, and the music has not been adapted to reflect the real-time emotional status of listeners.

To overcome this challenge, we divide the emotional interaction between listeners and music into three sequential stages: emotional classification, emotional interaction matching, and emotional interaction utility. First, it is necessary to classify the sentiment of the music itself and the listener's real-time sentiment. Whereupon, based on a set of rules, we need to match the music to the sentiment of the listener, allowing the listener to choose the duration and mode of audio-visual interaction with the music. Finally, the effectiveness of this interaction is assessed utilizing the physiological indicators and other quantifiable parameters of the listener.

Data have been collected from social media platforms such as Weibo and established a bimodal sentiment dataset employing oral and fragmented language as well as self-portraits containing faces to extract the real-time multimodal data information of the research object. The next step is to classify them and perform multimodal fusion utilizing a deep learning approach. Concurrently, polyphonic classical music is designated to establish a dataset, and the sentiment is classified according to a three-dimensional sentiment model to generate a music sentiment database. Real-time sentiment and musical sentiment are compared and matched in accordance with certain rules, and the interaction between multimodal real-time sentiment and the musical sentiment is explored using physiological indicators and other large-scale statistical methods, with a particular emphasis on the regulatory effect of music sentiment on real-time negative sentiment.

The main contributions of this paper are divided into three parts:

(1) A small music sentiment database with three-dimensional sentiment labels and bipolar sentiment labels is established using the deep learning method in combination with the P-A-D sentiment three-dimensional classification model, and a deep learning model suited for classifying classical music was obtained

(2) A bimodal dataset from Weibo is collected and labelled, and the weight parameters of different modes are determined based on the classification accuracy when using late fusion for multimodal sentiment classification

(3) A human-music emotional interaction system based on deep learning is developed in order to study the interaction between music sentiments and user sentiments. The interaction effectiveness has been evaluated through individual physiological observation and questionnaire surveys

The remainder of this paper is structured as follows: Section 2 provides a concise literature review on multimodal sentiment classification, music sentiment classification, and their interaction effects. The multimodal sentiment analysis method is proposed in the Section 4. Section 5 of the paper evaluates these concepts further with experimental results. The final section provides a summary of the paper.

2. Related Work

To study the interaction between music and real-time sentiment, it is necessary to systematically identify both the music and the real-time sentiment of listeners, and then apply specific rules. In this section, we will discuss the experimental methods of music sentiment classification and multimodal sentiment classification developed in recent years, as well as the role of music in regulating sentiment.

2.1. Music Sentiment Classification. Classification of musical sentiment is an interdisciplinary field of study that primarily involves psychology, auditory physiology, musicology, signal and information processing, pattern recognition, etc. In contrast to language, style, and genre, the emotional understanding and definition of music by listeners are highly subjective, making it difficult to classify music according to emotion.

The conventional method of manual annotation requires listening to music in its entirety, which is time-consuming. It is not suitable for emotional classification, which requires a large number of musical works, and it cannot fulfil the real-time classification requirements. The current methods for classifying music sentiment are based on deep learning techniques that are derived from traditional manual annotation and can function using extracted music features.

Nag et al. proposed a music segment sentiment classification method based on deep learning without the usages of Indian Classical Music audio clips as training data [5], thereby enhancing the training speed and classification accuracy of the model. Hizlisoy et al. proposed an approach for music emotion recognition based on a deep neural network and employing a new dataset of 124 Turkish traditional music excerpts [6]. The proposed system with 10-fold cross-validation achieves an overall accuracy of 99.19%.

Since 2017, there has been a major rise in research on music sentiment classification, as evidenced by the literature measurement of all music sentiment classification-related literature on CNKI (China National Knowledge Infrastructure) from 2014 to 2020, as shown in Figure 1. There has also been a significant rise in the amount of literature utilizing multimodal methods for music sentiment classification. A significant sentiment of research on multimodal music sentiment classification had been published as of 2020, demonstrating that deep learning and multimodal methods have entered the mainstream and are commonly present.

2.2. Multimodal Sentiment Classification. American researcher Ekman discovered that humans have six basic emotions that can be combined to generate other complex emotions, such as depression, tension, and anxiety [7]. Since the proposal of this classification method, it has gradually become one of the most prevalent standards in the current sentiment classification. Additionally, the sentiment classification standard can be diminished to three labels: positive, negative, and neutral.

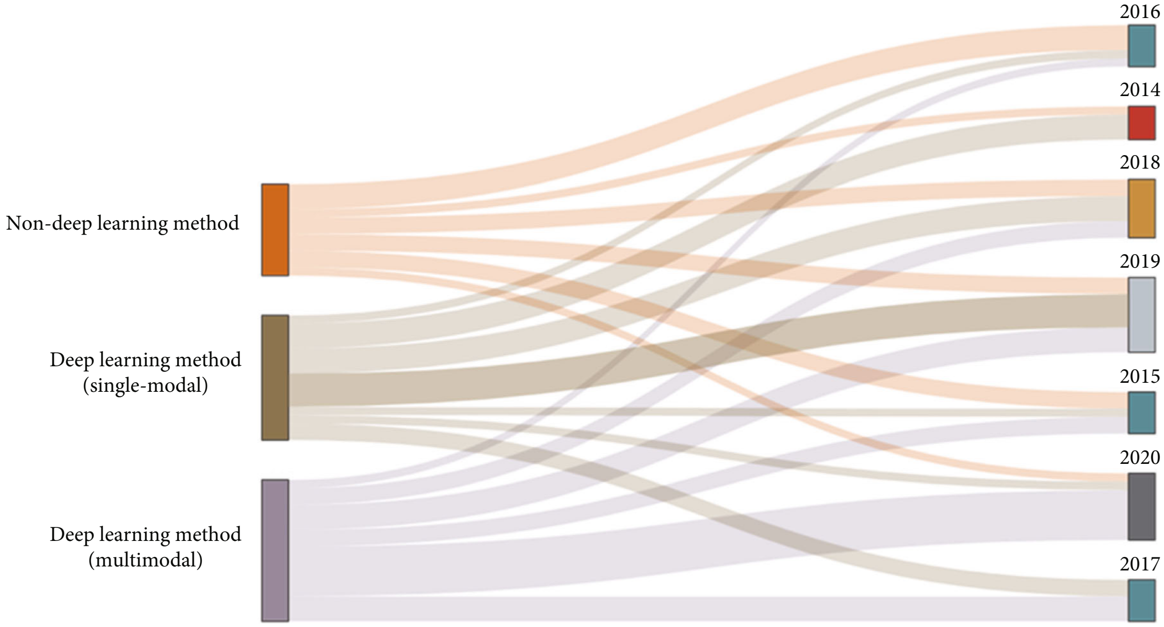


Figure 1: Music sentiment classification bibliometric Sankey chart. CNKI provided the entire body of the statistical literature, which covers the years 2014 through 2020. With the passage of time, it is evident that deep learning methods dominate a greater proportion of music sentiment classification, and there are an increasing number of studies employing multiple modes.

Hablani leveraged the VGG19 model of transfer learning to recognize the JAFFEE dataset in terms of single-modal sentiment recognition, and the accuracy of the results is 94.44% [8]. Zhang et al. utilized the multilevel multi-input-output model of the bidirectional recurrent neural network, which took the semantic and lexical information of emotional expression into consideration. Han et al. proposed a pretraining multitask learning model based on Bidirectional Gated Recurrent Units, and the experimental results demonstrated the model's high computational efficiency and success in the customer comment dataset [9]. Experimental results on SentiDrugs demonstrated that compared to other advanced architectures, this method enhances the effectiveness of drug review sentiment classification [10].

In terms of emotional classification of multimodal information, Tseng et al. fused sound and text features utilizing the decision-level fusion method and discovered that the decision-level fusion model can estimate the behavior rating more accurately than a single acoustic or lexical system [11]. Huang et al. used vector regression to incorporate the estimates of various feature sets during the continuous multimodal sentiment prediction by recurrent neural network procedure. This model achieved promising results in the AVEC 2017 development and testing set [12].

The number of articles on multimodal sentiment classification in the ACM (Association for Computing Machinery) Digital Library from 2015 to 2019 revealed that the proportion of multimodal sentiment classification articles employing deep learning methods increased from less than 20% in 2015 to more than 82% in 2019. From 2017 to 2021, the proportion of multimodal sentiment classification articles employing deep learning methods increased almost continuously from 33.3% in 2017 to 79.4% in 2021, as determined by a measurement of relevant literature on Science Direct, as shown in Figure 2.

In the project of sentiment recognition for the research object, the most convenient mode for real-time extraction is the object's current facial expression and real-time voice sentiment content. To express the real-time sentiment of the research object, this study chooses two modes: visual and textual.

2.3. Adjustment of Music to Sentiment. Gebhardt et al. elaborated on the relationship between music sentiment and sentiment regulation, and music therapy confirmed their findings [13]. In the majority of cases, other studies on the influence of music and the level of individual physiological changes revealed a close relationship between the two. Music with negative sentiments would prompt listeners to develop negative sentiments, whereas music with positive sentiments would induce positive sentiments.

Furuya et al. developed a method for classifying the emotional sentiment of music based on the lyrics and have applied it to music therapy. The target classification accuracy test of the sentiment category representing sadness [14] is used to demonstrate the effectiveness of the classification method. Ko et al. extracted positive and negative music patterns by analysing the amplitude, frequency, and rhythm of music using data mining and reverse engineering. They divided music into positive and negative patterns, created therapeutic music, applied the research findings to music therapy, and achieved convincing results [15].

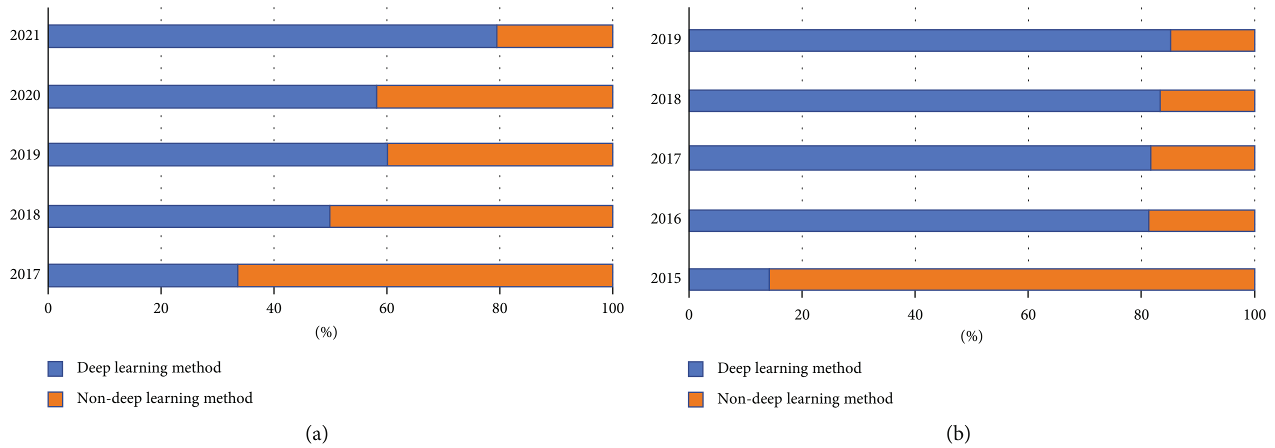


Figure 2: The proportion of multimodal sentiment classification literature measurement in recent five years. Calculated numbers of research about multimodal sentiment analysis from ACM digital library, from 2015-2019 (a) and that from Science Direct, from 2017-2021(b).

3. Sentiment Classification and Interaction

The deep learning approach and models used in the sentiment classification experiment will be covered in this section. We will first present a model for classifying the sentiments of music, followed by descriptions of models for classifying the textual and visual sentiments, and a multimodal sentiment fusion method needed by listeners' sentiment classification.

3.1. Music Sentiment Classification Model. Prior to Mehrabian and Russell's 1974 proposal of the P-A-D three-dimensional sentiment model, it was difficult to quantitatively classify the sentiment of music, images, and other data, or it could only be classified utilizing simple binary values. In the P-A-D model of sentiment, "P" represents the degree of pleasure that can reflect the positive and negative characteristics of any given emotional state. While "A" represents the level of activation that can reflect the neurophysiological activation level of each individual, "D" represents the level of dominance that can reflect the individual's control over the situation and others [16]. In the P-A-D sentiment model, all sentiments can be represented by different combinations of dimensions.

We use the P-A-D sentiment model as a benchmark to manually label music collected for this study, and Librosa was used to extract 14 features, such as maximum tone, tone standard deviation, average MFCC(Mel Frequency Cepstral Coefficients), and so on [17]. To classify music clips, a CNN (Convolutional Neural Network) model with two convolution layers and two fully-connected layers is developed.

The convolutional neural network is a special multilayer perceptron or feedforward neural network. It is comprised of a large number of neurons organized in a particular manner to respond to overlapping areas in the visual field, and it possesses the characteristics of local connection and weight sharing. It consists primarily of convolution layers, pooling layers, a fully connected layer, and other fundamental units.

The convolution kernel side length of each convolution layer in our custom-built CNN model is 5 and the activation function is ReLU. The dropout parameter is set to 0.5 during a full connection. The structure of the network is shown in Figure 3.

The processed 194×1 dimension music feature vector was input into the first convolutional layer, and the matrix with the size of 190×5 was obtained after the convolution with the size of the convolution kernel, and the channel count of 5 was obtained; whereupon, the second convolutional layer is entered. Following convolution with a kernel size of 5 and channel number of 5, the matrix with a size of 186×5 is obtained. The ReLU activation function is utilized between the convolution layers so that the output of the network is no longer a linear combination of the inputs and the network can be sparse to avoid gradient disappearance.

Whereupon, the matrix with a size of 186×5 is input into the fully connected layer, which is converted into a one-dimensional vector of 903×1 by Flatten function and entered into the hidden layer. Through the hidden layer, the full connection from 903×1 dimensional vector to 1024 nodes and from 1024 nodes to 3 nodes is completed. The Dropout parameter is set to 0.5 and the ReLU activation function is used to connect the two fully connected layers in this process.

3.2. Textual Sentiment Classification Model. In this study, a two-way gated cyclic unit network was utilized for text sentiment recognition. GRU (Gate Recurrent Unit Network) model refers to a model that maintains the LSTM (Long Short-Term Memory) effect [18] but features the advantages of a simpler structure, fewer parameters, and better convergence. The specific structure of the model is shown in Figure 4.

In order to construct the text sentiment classification model, we need to first preprocess the text data in the dataset using Jieba for word segmentation and sentence decomposition into a collection of single words. After removing stop words, the words are entered into word2vec to obtain the word vector. The 400-node Bi-GRU model is fully connected

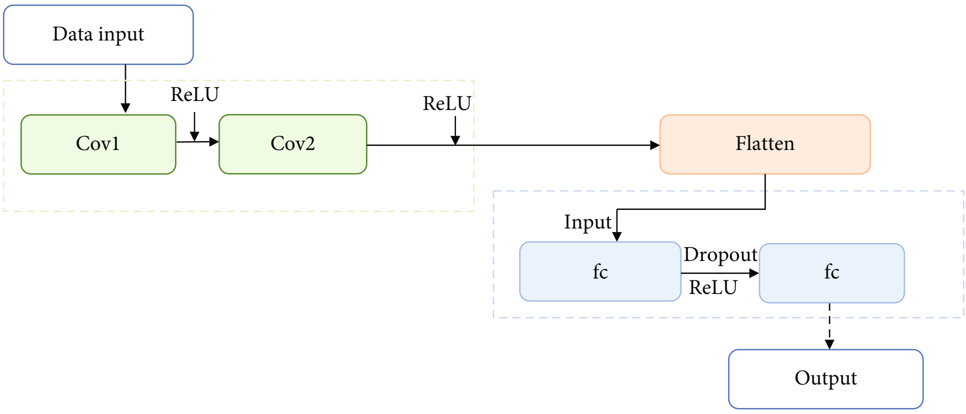


Figure 3: Model of music sentiment classification. A convolutional neural network model with two convolution layers and two fully connected layers for classifying sentiment labels on any musical clip.

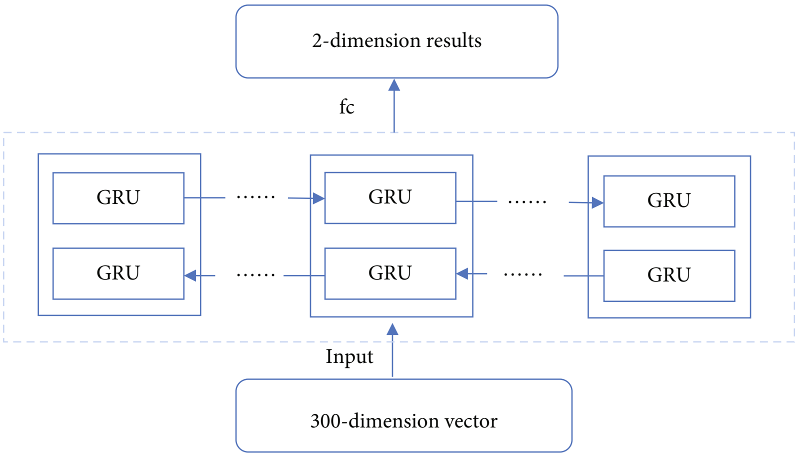


Figure 4: Bigated Recurrent Unit Networks Model. With the input of a 300-dimension vector, 400 nodes in the model are able to transfer the vector into 2-dimension results.

to the word vector to obtain the sentiment classification result, which is recorded as t_0.

The algorithm of the GRU model is shown in the following formulas, (1) to (5); where σ represents the sigmoid function. The Sigmoid function is commonly used as the activation function of neural networks to map variables between 0 and 1 due to its single increment and inverse function single increment properties. $\odot$ represents the tanh function. Since its graph is approximately linear in the vicinity of domain 0 and is derivable in the entire domain, this function is extensively used in the field of deep learning as the activation function of neurons in neural networks. h_t represents the hidden state passed down by the current node, x_t represents the input of the node, and y_t represents the output of the node.

Besides, z_u represents the update gate while z_r represents the reset gate. The updated gate is used to regulate the extent to which the state information from the previous moment is brought into the current state. The greater the value of the update gate, the greater the amount of previous state information that is brought into the current state. The reset gate controls how much information from the previous state is written to the current candidate set $h_t{}'$.

$$z_u = \sigma\left(w_u \times \left[h_t - 1, x_t \right] + b_u \right), \tag{1}$$

$$z_r = \sigma\left(w_i \times \left[h_t - 1, x_t \right] + b_r \right), \tag{2}$$

$$h_t{}' = \sigma\left(w \times \left[h_t - 1 \odot z_r, x_t \right] + b \right), \tag{3}$$

$$h_t = z_u \odot c_t - 1 + (1 - z_u) \odot z, \tag{4}$$

$$y_t = \sigma\left(w^{'} h_t \right). \tag{5}$$

3.3. Visual Sentiment Classification Model. For visual sentiment classification, we select the VGG19 (Visual Geometry Group 19) model, which is a CNN model for image feature extraction and classification proposed in 2014 by Visual Geometry Group Oxford. It demonstrates that the performance of the model is somewhat positively correlated with network depth [19]. VGGNet builds a CNN model capable of extracting image features by repeatedly superimposing

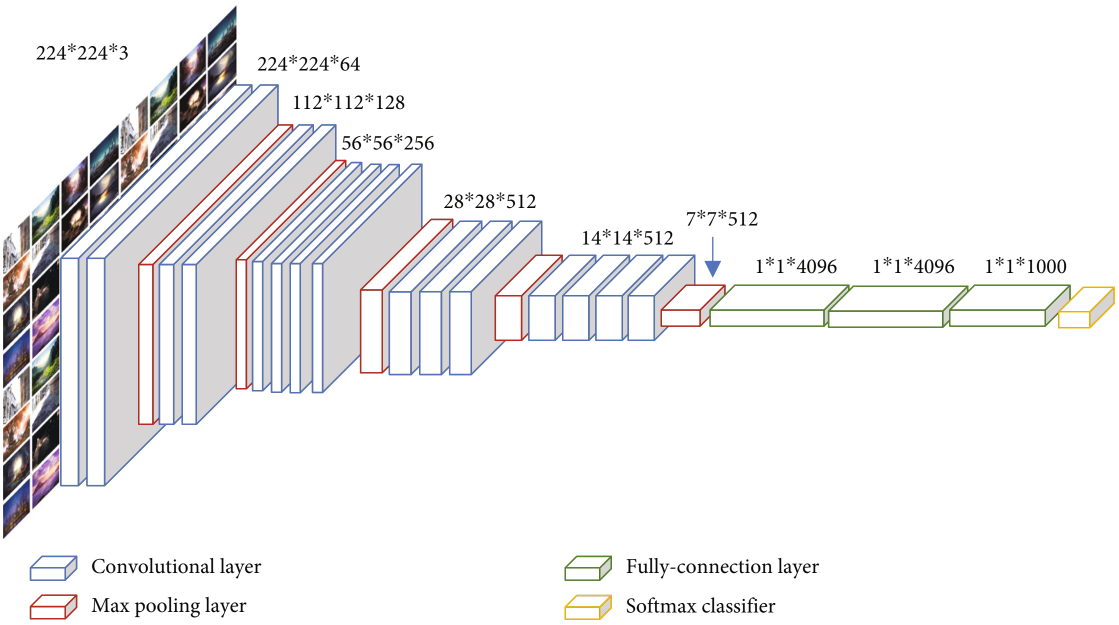


Figure 5: VGG19 model hierarchy diagram. It encompasses 19 hidden layers, including 16 convolution layers and 3 fully-connected layers.

the small convolution kernel with a side length of 3 and the maximum pool layer with a side length of 2.

As illustrated in the figure, the input layer is followed by two convolutional layers and then a maximum pooling layer, which is responsible for halving the dimension of the image. These three layers are considered a unit, repeated five times, and then three fully connected layers are connected to extract the feature vector of the input picture data. The extracted feature vectors are fed into the Softmax classifier, allowing for the classification of pictures. Instead of using a larger convolution kernel to decrease the number of convolution layers, VGG Net prefers to reduce the side length of the convolution kernel, notwithstanding the fact that this increases the number of convolution layers, as shown in Figure 5.

The original face images in different emotional states, with a size of 48×48, are cropped to 44×44 and decolored. Random clipping is used in the training set and central clipping is used in the test set. The images are then inverted to achieve data enhancement, forming a dataset as v_0'. The output of each convolution layer is $v_n' = \mathrm{relu}(v_{n-1}')$, $n \in R_{16}$, and the final output v_{16}' is the eigenvector of $1 \times 1 \times 512$. Input the vector into the full connection layer to obtain normalized features v_i:

$$v_i = W_v v_i + b_v, v_i \in R^{N\times d}. \tag{6}$$

The dlib library is used to preprocess images, including face recognition and face extraction, in the process of practical application as the proportion of face size in the whole image input by listeners is significantly smaller than that in the image dataset.

3.4. Multimodal Sentiment Fusion Method. The classification of multimodal sentiment fusion based on deep learning is primarily comprised of two fundamental technologies: feature-level fusion (early fusion) and decision-level fusion (late fusion). In addition, it includes a number of alternative techniques, including intermediate fusion and hybrid multimodal fusion. Feature level fusion is accomplished by connecting the feature vectors of all modes to form a single, long feature vector. While the correlation between different modal information can be considered, obtaining large-scale datasets with a one-to-one correspondence between modes remains difficult when self-built datasets have volume and label distribution issue, as shown in Figure 6.

Researchers use classifiers to label the features extracted from each mode in the decision-level fusion process, and then apply rules and weight matching to obtain the final sentiment classification. Therefore, the most obvious distinction between late fusion and early fusion is whether the fusion object is a feature prior to single-mode classification or the result of single-mode classification.

The overall performance of the model can be enhanced by selecting the network model with the highest accuracy for feature extraction and classification of each mode during the decision-level fusion process [12]. Moreover, when the information of each mode is classified, the errors caused by different classifiers will not interfere with one another.

The following Figure 7 depicts the percentage distribution of multimodal emotion fusion literature in the ACM Digital Library. It is evident that despite a slight decline in their overall numbers over the past five years, decision-level and feature-level fusion methods have consistently retained the dominant position in multimodal emotion fusion classification methods. While the total number of middle-level fusion and hybrid fusion methods is less than

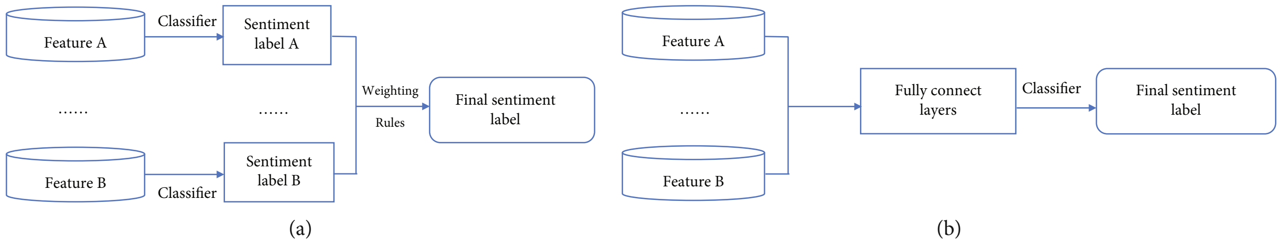


Figure 6: Schematic diagram of the process of feature-level fusion (a) and decision-level fusion (b) methods used in multimodal fusion. The feature level fusion concatenates the feature vectors of all modes and then inputs them into the classifier, whereas the decision level fusion inputs the feature vectors of different modes directly into the classifier and aggregates the feature scores according to different weights.

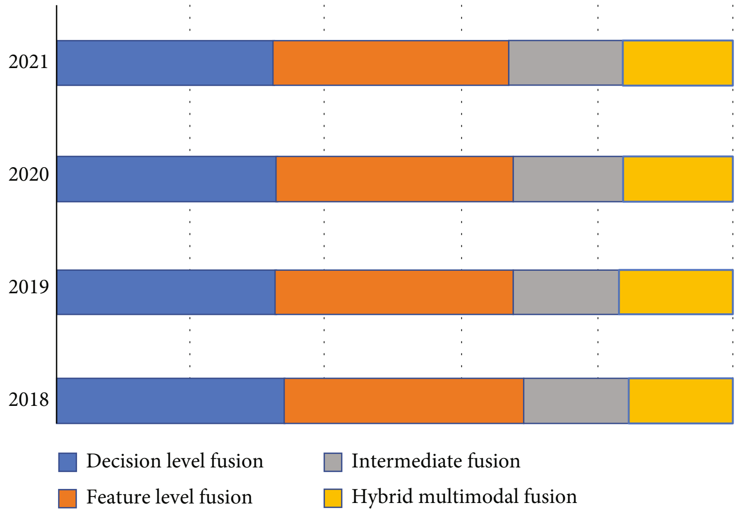


Figure 7: Multimodal sentiment fusion method Bibliometrics, data from ACM Digital Library from 2018 to 2021. The two basic methods, feature level fusion, and decision level fusion, reimain the mainstream methods with unique advantages.

that of the other two methods, the number of articles has increased slightly over time, which bodes well for the classification of multi-modal emotion fusion.

Due to the immediacy of data input, the bimodal data in this study are closely related; therefore, decision-level fusion is selected as the multimodal sentiment fusion method. Assuming that the classification result of textual feature is t_0, and the classification result of visual feature is v_0, the decision level fusion method is r. The algorithm used in decision level fusing is shown in the following.

$$\text{Label}_{\text{fusion}} = t_0 \times r_1 + v_0 \times r_2 \in R^C, \tag{7}$$

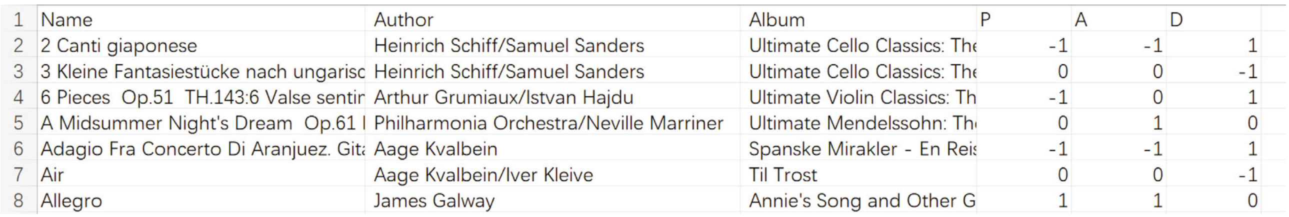
where r_1 is the weight of the textual sentiment classification label, r_2 is the weight of the visual sentiment classification label and represents the number of sentiment categories. This method can be used to categorize the emotional dispositions of listeners. Regarding the determination of weight, we refer to Yan et al's research [20] and determine the weight coefficients occupied by different modes in decision-level fusion based on the characteristics of user classification and the accuracy results.

4. Experimental Results

4.1. Datasets. Four data sets were used in this study to classify the real-time sentiment of listeners and the sentiment of music, as well as to determine the weight of the sentiment classification results of different modes in the fusion of the two-mode sentiment of listeners.

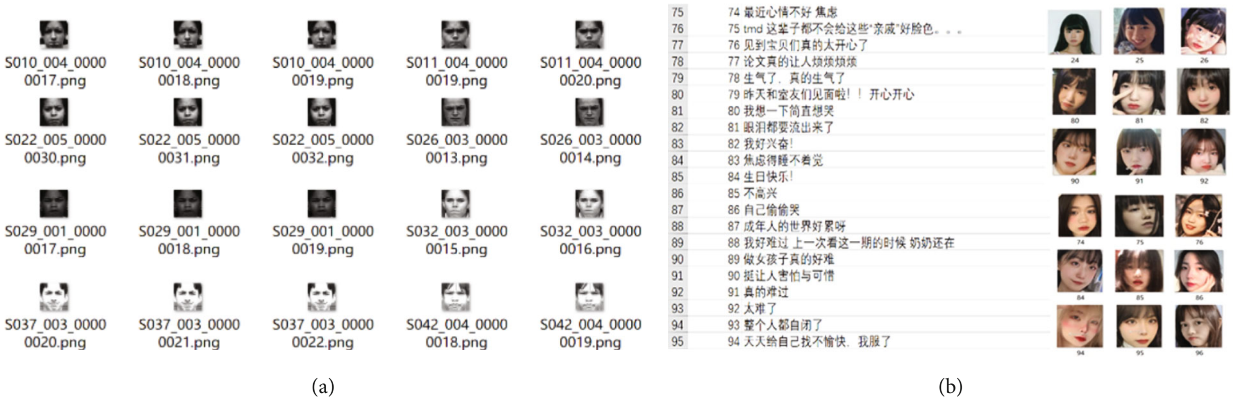
4.1.1. JCMD Dataset (Jiang's Classical Music Dataset). Since the majority of the current large data sets consist of songs with lyrics or pure piano music, we select 250 pieces of polyphonic classical music for normalization to create a dataset of music. We trimmed the music and selected 10 to 15 second phrases close to the beginning of each piece that exhibit an obvious emotional tendency, and construct a small music segment dataset with P-A-D labels for the classification of music sentiment, shown in Figure 8.

4.1.2. CK+ Dataset (the Extended Cohn-Kanade Dataset). We use the open CK+ dataset when training the VGG model. This dataset containing 593 video sequences of 123 subjects with expression tagging [21] was collected and managed by the lab, as shown in Figure 9(a).



1	Name	Author	Album	P	A	D
2	2 Canti giaponese	Heinrich Schiff/Samuel Sanders	Ultimate Cello Classics: The	-1	-1	1
3	3 Kleine Fantasiestücke nach ungarisc	Heinrich Schiff/Samuel Sanders	Ultimate Cello Classics: The	0	0	-1
4	6 Pieces Op.51 TH.143:6 Valse sentir	Arthur Grumiaux/Istvan Hajdu	Ultimate Violin Classics: Th	-1	0	1
5	A Midsummer Night's Dream Op.61 I	Philharmonia Orchestra/Neville Marriner	Ultimate Mendelssohn: Th	0	1	0
6	Adagio Fra Concerto Di Aranjuez. Gita	Aage Kvalbein	Spanske Mirakler - En Reis	-1	-1	1
7	Air	Aage Kvalbein/Iver Kleive	Til Trost	0	0	-1
8	Allegro	James Galway	Annie's Song and Other G	1	1	0

Figure 8: Samples of JCMD Dataset (Jiang's classical music dataset). 250 normalized clips of polyphonic classical music are collected and labelled. Three discrete labels -1, 0, and 1 are used to label different dimensions in the PAD 3D emotion model, which is then used to train and verify the model.



75	74	最近心情不好 焦虑
76	75	tmd 这辈子都不会给这些"亲戚"好脸色。。。
77	76	见到宝贝们真的太开心了
78	77	论文真的让人烦烦烦烦
79	78	生气了．真的生气了
80	79	昨天和室友们见面啦！！ 开心开心
81	80	我想一下简直想哭
82	81	眼泪都要流出来了
83	82	我好兴奋!
84	83	焦虑得睡不着觉
85	84	生日快乐!
86	85	不高兴
87	86	自己偷偷哭
88	87	成年人的世界好累呀
89	88	我好难过 上一次看这一期的时候 奶奶还在
90	89	做女孩子真的好难
91	90	挺让人害怕与可惜
92	91	真的难过
93	92	太难了
94	93	整个人都自闭了
95	94	天天给自己找不愉快．我服了

Figure 9: The Extended Cohn-Kanade Dataset (a) and Bimodal Dataset (b). The CK+ dataset images have been preprocessed with face extraction, colour removal, and equal scale cutting. The weight occupied by the bimodal in decision-level fusion is determined by individually matching the emotion text with the color face image in the bimodal emotion dataset that I created.

4.1.3. Weibo_Senti_100k Dataset. The dataset used for text sentiment classification training and testing in this study is Weibo_Senti_100k, including 119988 emotional forwarding or original clips in the microblog, with positive, negative, and neutral emotional labels [22].

4.1.4. Bimodal Dataset. For the purpose of determining the weight in the late fusion method, a Weibo-based self-built bimodal dataset is utilized to validate the model. At the time of release, 1,000 face images and associated sentiment words are individually matched up, as shown in Figure 9(b).

4.2. Results on Sentiment Classification. The CNN model developed in the previous section is used to train the JCMD dataset, and the validation dataset is extracted to validate the classification accuracy of the model in the P-A-D dimensions. The results are shown in Table 1.

Since listener sentiment will ultimately be classified into positive and negative categories, the following rules from Table 2 are applied when converting P-A-D three-dimensional sentiment tags into positive-negative ones:

Currently, the database contains 123 positive songs and 127 negative songs. Based on the classification results generated by the model, any music can be classified and added to the database. In the single modal sentiment classification of listeners, SGD (Stochastic Gradient Descent) is used to optimize the VGG model, and the epoch-accuracy curve during operation is shown in the following Figure 10:

Table 1: Accuracy of music sentiment classification.

Dimension	Accuracy
P (pleasure)	0.6522
A (activation)	0.6087
D (dominance)	0.5217

Table 2: : Rules for the corresponding P-A-D with two sentiment dimensions.

Music sentiment	P	A	D
Positive	1	0, 1	-1, 1
Negative	-1, 0	-1, 0	0, 1

Finally, text sentiment classification is accurate to a level of 74.3%, while visual sentiment classification can achieve accuracy levels of nearly 100%.

The bimodal sentiment dataset outlined in Section 4.1 is utilized to determine the sentiment weight of each and every modal classification model. Comparing the final sentiment labels calculated by the weight rule with the manual annotation results reveals that the multimodal sentiment classification has the highest accuracy of 87.27%, as shown in Table 3. The rule assigns a weight of 0.3 to visual sentiment and 0.7 to textual sentiment.

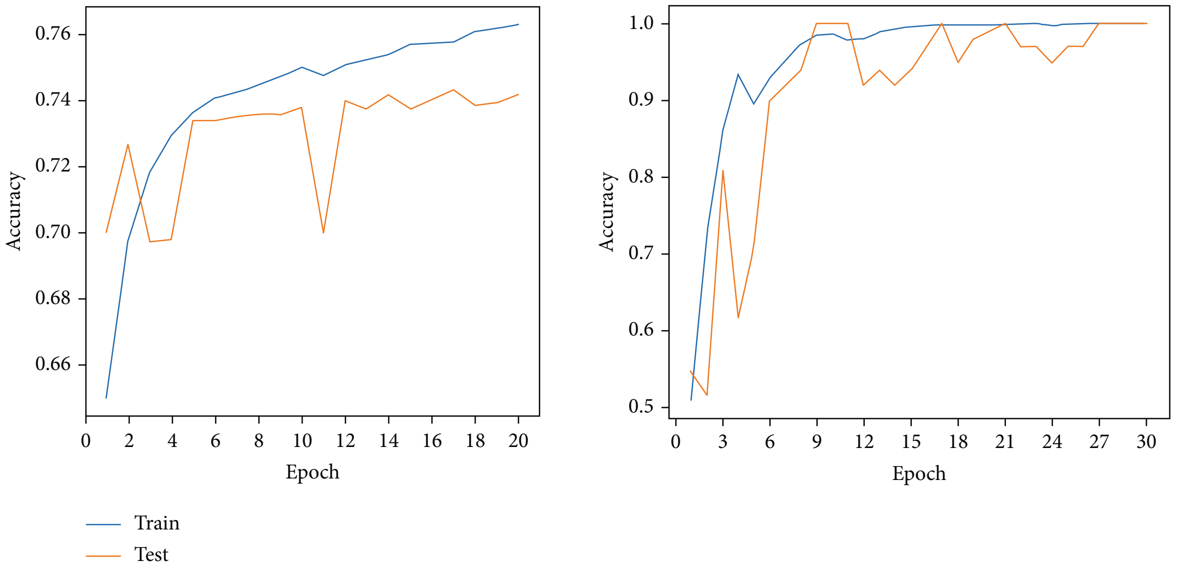


Figure 10: Epoch-Accuracy chart of textual and visual sentiment.

Table 3: Accuracy of corresponding to different weights.

Weight of text	Weight of image	Accuracy
0.1	0.9	0.8179
0.3	0.7	0.8421
0.5	0.5	0.8539
0.7	0.3	0.8727
0.9	0.1	0.8635

4.3. Implementation. The interaction system for music sentiment is comprised of deep learning models, a music sentiment database, and web pages. The real-time emotional performance of the listener is converted into data using the camera and voice-to-text API, and the music sentiment is matched and interacted with based on the two tags. Through individual tracking methods and group survey methods, whether music has a regulatory effect on negative sentiments can be observed.

(1) Individual observation

In their research, Harris et al. confirmed that negative sentiments, particularly anger, have a strong positive correlation with heart rate [23]. As a result, this study will select a participant as its sample, recording, and analysing the heart rate fluctuations of the observation object. The rationality coefficient ε is calculated by the following:

$$\varepsilon = \frac{\beta - m}{m}, \tag{8}$$

where β represents the heart rate of the object when angry, and m represents the daily average heart rate of the volunteer. The volunteer is suitable for the experiment when ε is greater than 0.1.

If α indicates the heart rate level of the subject after using the music interaction system, the meaning of β is the same as that in formula (8). The variation coefficient of this empirical analysis μ is calculated by the following:

$$\mu = \left| \frac{\alpha - \beta}{\beta} \right|. \tag{9}$$

The volunteer of the study is female, 54 years old with an ε of 0.32, indicating that she is suitable for implementation. The heart rate of her in different states is shown in Figure 11:

Using the daily average heart rate (mean) as a baseline, it can be seen that the individual's heart rate increases significantly during anger, with the rate of increase ranging between 20% and 50%. After interacting with music sentiment, the heart rate decreases by 15% to 20%, although it remains above the daily mean heart rate.

Based on the tracking of the observed object, it is evidenced that the music sentiment interaction system has a relatively obvious regulation effect on negative sentiments (especially anger).

(2) Group survey

Since it is difficult to track and investigate a large-scale sample size in a strictly controlled laboratory setting, a questionnaire is used to determine whether music can effectively regulate negative sentiments. The questionnaire designed for this study contains 12 questions with seven measurement levels, more conducive to later modelling and analysis.

This survey received a total of 271 completed questionnaires, of which 250 were adjudged valid. The questionnaire is distributed at random to as many universities, enterprises, and institutions as possible, targeting individuals of all ages and genders, as shown in Figure 12.

The data revealed that women and young people are more likely to use the system for emotional interaction with

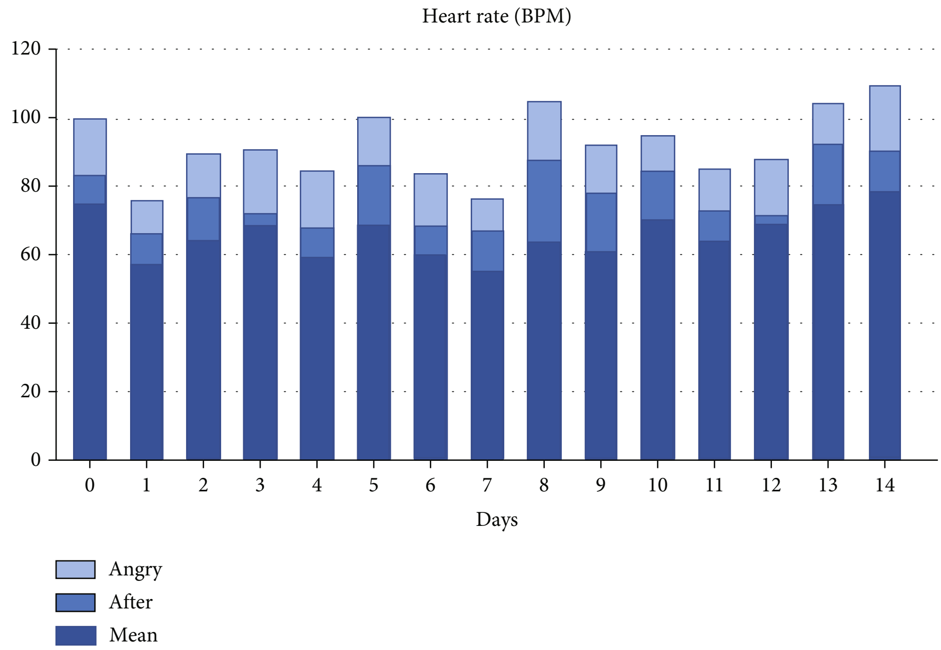


Figure 11: Investigation of individual heart rate.

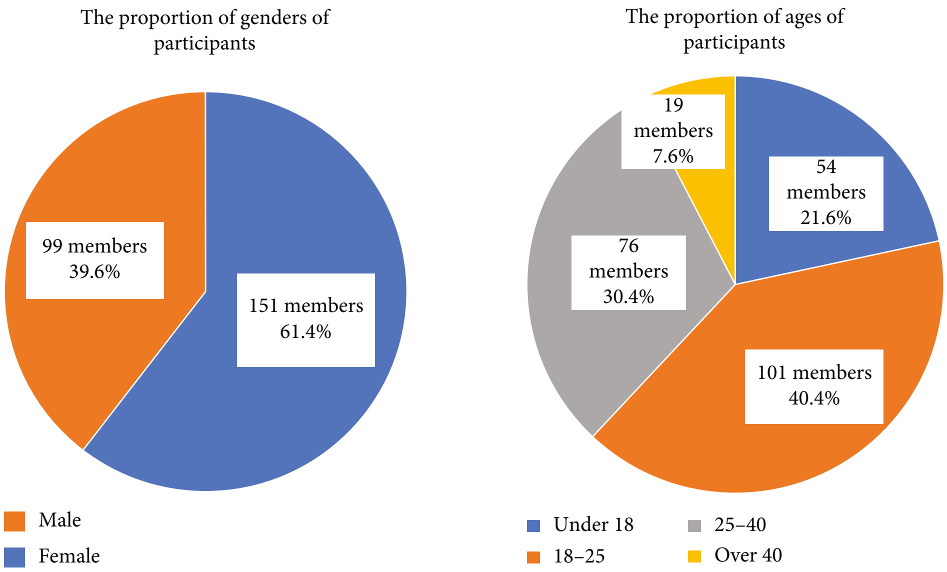


Figure 12: Descriptive statistics of gender and age of questionnaire participants. Descriptive statistics of gender and age of questionnaire participants. The participants were both male and female and spanned a wide range of age groups.

music and emotional regulation. The user rating of the system is 4.54 (the full score is 7).

The Cronbach alpha coefficient of the questionnaire is 0.789, indicating that it merits further investigation. ANOVA was used to determine if there were statistically significant differences in the rate of heart rate reduction from anger to calmness between users with different system frequencies. The new variable decrease in heart rate is calculated as the ratio of the heart rate difference between before and after using the software to the heart rate during anger.

Selecting the frequency of system utilization as the independent variable and the heart rate decrease ratio as the dependent variable, we observe that the significance is 0.367, indicating that it is significant at the 0.5 level. This result indicates that the frequency of system use of the human-music emotional interaction system has a significant effect on the decrease in heart rate, excluding the effects of age.

In this study, correlation analysis is also employed to examine the correlation between user scores and user satisfaction with the system's functions. The correlation between user satisfaction with the sentiment recognition system, the music recommendation system, and the music playing system is significant at the 0.01% level. Therefore, these three aspects require additional research.

5. Conclusion

In this study, we developed an innovative sentiment classification and music interaction system based on deep learning in order to investigate the interaction between music and listener sentiments. A three-dimensional sentiment classification model for input music clips was developed using a dataset comprised of P-A-D three-dimensional sentiment classification label and a one-dimensional convolution neural network. To obtain real-time sentiment classification results, the late fusion method with a 3 : 7 weight rule is adopted to incorporate the results of text sentiment and picture sentiment classification. The combination of individual and group surveys is utilized to verify that music sentiment can regulate and control the negative sentiment of listeners.

In the future, for the purpose of enhancing the accuracy of the research, in-depth explorations need to be conducted on two aspects: the extraction of listener data and the classification of the emotional label. Specifically, it would be recommendable to introduce video capture for automatic extraction to replace the existing manual control and to integrate an audio feature as a mode for measuring the listener's sentiments. Additionally, the labels of sentiment classification should be updated from positive-negative dimensions to three or more dimensions to more accurately describe music and listeners' sentiments. Further research also has to be performed on the positive feedback regulations that music has on the sentiments of listeners.

Data Availability

Previously reported CK+ data were used to support this study and are available at http://www.consortium.ri.cmu.edu/ckagree/. These prior studies (and datasets) are cited at relevant places within the text as references [21]. Previously reported Weibo_senti_100k data were used to support this study and are available at https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/weibo_senti_100k/intro.ipynb. These prior studies (and datasets) are cited at relevant places within the text as references [22]. The JCMD Dataset (Jiang's classical music dataset) data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This paper was supported in part by the National Natural Science Foundation of China (Grant No.71774084).

References

[1] IResearch, *China's Music Industry Development Research Report – Digital*, 2020, https://report.iresearch.cn/report_pdf.aspx?id=3668.

[2] International Federation of the Phonographic Industry (IFPI), *Global Music Industry Report 2021*, 2021, https://max.book118.com/html/2021/0407/8027066034003072.shtm.

[3] NetEase, "IPO Prospectus of NetEase Cloud music," 2021, https://www.doc88.com/p-23047076740709.html.

[4] P. N. Juslin and J. A. Sloboda, *Music and Emotion: Theory and Research*, Oxford University Press, Oxford, 2009.

[5] S. Nag, M. Basu, S. Sanyal, A. Banerjee, and D. Ghosh, "On the application of deep learning and multifractal techniques to classify emotions and instruments using Indian classical music," *Physica A: Statistical Mechanics and its Applications*, vol. 597, article 127261, 2022.

[6] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *Engineering Science and Technology, an International Journal*, vol. 24, no. 3, pp. 760–767, 2021.

[7] P. Ekman, *Facial Action Coding System*, Consulting Psychologists Press, Palo Alto, CA, 1978.

[8] R. Hablani, "Facial expression recognition using transfer learning on deep convolutional network," *Bioscience Biotechnology Research Communications*, vol. 13, no. 14, pp. 185–188, 2020.

[9] L. Zhang, Y. Zhou, X. Duan, and R. Chen, "A hierarchical multi-input and output bi-GRU model for sentiment analysis on customer reviews," *IOP Conference Series: Materials Science and Engineering*, vol. 322, article 062007, 2018.

[10] Y. Han, M. Liu, and W. Jing, "Aspect-Level drug reviews sentiment analysis based on double bigru and knowledge transfer," *IEEE Access*, vol. 8, pp. 21314–21325, 2020.

[11] S.-Y. Tseng, H. Li, B. Baucom, and P. Georgiou, "Honey, I learned to talk," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, Boulder, CO, USA, 2018.

[12] J. Huang, Y. Li, J. Tao et al., "Continuous multimodal emotion prediction based on long short term memory recurrent neural network," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, Mountain View, CA, USA, 2017.

[13] S. Gebhardt, I. Dammann, K. Loescher, P. M. Wehmeier, H. Vedder, and R. von Georgi, "The effects of music therapy on the interaction of the self and emotions–an interim analysis," *Complementary Therapies in Medicine*, vol. 41, pp. 61–66, 2018.

[14] M. Furuya, H.-H. Huang, and K. Kawagoe, "Music classification method based on lyrics for music therapy," in *Proceedings of the 18th International Database Engineering &amp; Applications Symposium on - IDEAS '14*, Porto, Portugal, 2014.

[15] L.-W. Ko, Y.-T. Chen, and M.-C. Chiu, "Integrating music therapy and music information retrieval using music pattern analysis," in *Moving Integrated Product Development to Service Clouds in the Global Economy*, pp. 678–687, IOS Press, 2014.

[16] A. Mehrabian, "Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[17] X. Zou, *Renovamen/speech-emotion-recognition: Speech emotion recognition implemented in Keras (LSTM, CNN, SVM, MLP)*, GitHub, 2021, https://github.com/Renovamen/Speech-Emotion-Recognition.

[18] C. Jin, Z. Song, J. Xu, and H. Gao, "Attention-based Bi-DLSTM for sentiment analysis of Beijing Opera Lyrics," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 1167462, 8 pages, 2022.

[19] F. Miao, X. Wang, F. Feng, C. Jin, and L. Jin, "A Renovated CNN-Based model enhances KGC task performance," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 5968047, 12 pages, 2022.

[20] M. Yan, S. Li, C. A. Chan, Y. Shen, and Y. Yu, "Mobility prediction using a weighted Markov model based on mobile user classification," *Sensors*, vol. 21, no. 5, p. 1740, 2021.

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* -, San Francisco, CA, USA, 2010.

[22] Jinhuakst, *Inprove introductions for datasets(Weibo_senti_100k)*, GitHub Inc, 2018, https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/weibo_senti_100k/intro.ipynb.

[23] K. M. Harris, J. S. Gottdiener, S. S. Gottlieb, M. M. Burg, S. Li, and D. S. Krantz, "Impact of mental stress and anger on indices of diastolic function in patients with heart failure," *Journal of Cardiac Failure*, vol. 26, no. 11, pp. 1006–1010, 2020.