Music Sentiment Analysis Based On Multi-Modal Intelligent Computing And Deep Learning

Lu Huang*

JiLin Provincial Institute of Education, Changchun, Jilin 130022, China

Corresponding author. E-mail: lu_huang79@outlook.com, 13086879796@163.com

Received: Feb. 26, 2026; Accepted: Apr. 04, 2026

The study of intelligent computing and deep learning has become a prominent research topic among both industrial and academic researchers in recent years. As a typical form of intelligent computing and deep learning, with ongoing advancements in affective computing, the close connection between deep learning, multi-modal information, and emotion has gradually garnered the attention of researchers. Existing methods still exhibit many shortcomings in the perception, understanding, and expression of machine emotions. A computational model of emotion that integrates emotion perception, information fusion, and deep learning is proposed. The model is a deep learning-oriented network perception model that accepts visual, auditory, and textual inputs to achieve an understanding of uncertain emotions. Experiments demonstrate that the model performs well in various multi-modal emotion computations. The studies presented in this paper provide important guidance for the application of both multi-modal intelligent computing and deep learning.

Keywords: Intelligent Computing; Deep Learning; Music Sentiment Analysis; Multi-modal Information

http://dx.doi.org/10.6180/jase.202609_32.039

1. Introduction

Due to the fast development of multimedia technologies, music has become very accessible and a necessary element of everyday life to relax and control the mood [1]. Music is not only a reflection of the artistic expression, but it also reflects a variety of human emotions, which can produce a powerful emotional appeal. Emotion is thus an important element of music that has great research and practical importance [2]. Emotion analysis is an interdisciplinary study that includes artificial intelligence, computer vision and natural language processing. Ekman determined six fundamental emotions such as anger, disgust, fear, happiness, sadness and surprise and subsequent research included contempt [3, 4]. Multi-modal music is a medium of expression of emotions, and therefore sentiment analysis is important and challenging [5, 6]. Music is a mirror of human emotions in real life and a good tool of expressing psychological emotions of happiness, anger, and sorrow [7]. As the amount of music data grows exponentially, it has become a significant research topic to organize and retrieve music efficiently. The use of emotion-based classification is important in enhancing the accuracy and efficiency of music retrieval, but it has a number of challenges [8]. With the growing significance of emotions in music, numerous recommendation systems are trying to comprehend emotional content since emotion-based recommendations can be used to attract listeners and improve playlist generation. Gudivaka introduced an AI-based model of music teaching based on the analysis of big data and BiLSTM to process musical characteristics and allow intelligent assessment. In the proposed work, this method is modified by using multimodal feature extraction and BiLSTM-based fusion, which enhances accuracy and robustness in music sentiment classification [9]. The classification of emotions is more related to human perception because melody is an

expression of emotion and lyrics are a reflection of mood. Music has low level and high-level features [10] and is applied in event monitoring and video analysis [11]. This paper is concerned with the incorporation of melody and lyrics to enhance the classification of sentiments. Despite significant advancements, existing multi-modal sentiment classification methods often struggle to effectively integrate diverse inputs such as visual, auditory, and textual features, resulting in suboptimal performance. Many models fail to capture complex interactions between these modalities, limiting accurate emotion detection in music. The proposed model addresses this issue by integrating emotion perception, information fusion, and deep learning techniques. It aims to improve emotion classification by combining auditory, visual, and textual modalities and exploring their synergy, thereby overcoming limitations in multi-modal fusion and enhancing the accuracy of music emotion recognition.

2. Theory and formula

2.1. Literature review

The most popular method of music emotion classification is the audio-based method, which derives features of the music and applies machine learning to detect emotions [12]. Conventional approaches are based on audio features that are engineered such as Melfrequency cepstrum and time-domain features (e.g., short-time energy, mean amplitude, and autocorrelation) [13]. Other researchers have also demonstrated that lyrics have a great influence on emotion classification [14]. Schuber E studied how musical emotions and audio features interacted with time based on a two-dimensional emotion model [15]. Li Tao and Ogihara M used Euclidean distance to measure similarity and sentiment detection in music with smaller distance implying a greater similarity between two songs. Sun proposed the Information Cell Mixture Model (ICMM), a fuzzy-based approach that is useful in highdimensional sentiment classification. Huq applied linear regression and support vector machines in recognizing emotions. Chin Y suggested a two-layered SVM model that categorized emotions as angry, happy, sad, and calm [16, 17]. Grekow subsequently investigated the best combinations of audio features to use in music emotion recognition [18]. The melody is not always sufficient to differentiate emotions, and lyrics are important in the analysis of music sentiment [19]. Lyrics offer abundant information of emotion and when combined with tone, it forms a full expression of emotion. Lyric emotion features can be obtained through text processing. In 2012, Jing Li and Hongfei Lin suggested that an emotion vector space could be used by extracting emotional words and classifying songs by similarity [20]. In 2013, Mahn M proposed an approach based on the Plutchik model, which concentrated on intro and chorus lyrics. The intro establishes emotional tone, and the chorus focuses on the repeated emotional content, enhancing the accuracy of emotion classification [21–24]. Music analysis is adopting deep learning models instead of feature engineering. Lidy suggested a Music Information Retrieval Evaluation Exchange of audio processing [25]. Jeon applied dual CNNs to sentiment classification, whereas Huang applied a Boltzmann machine to connect audio and lyrics.

3. Experimental setup

3.1. Research Design

3.1.1. Music Sentiment Analysis

Music emotion involves visual, auditory and textual features, making extraction complex. Figure 1 shows the pipeline from input and feature extraction to fusion and sentiment classification.

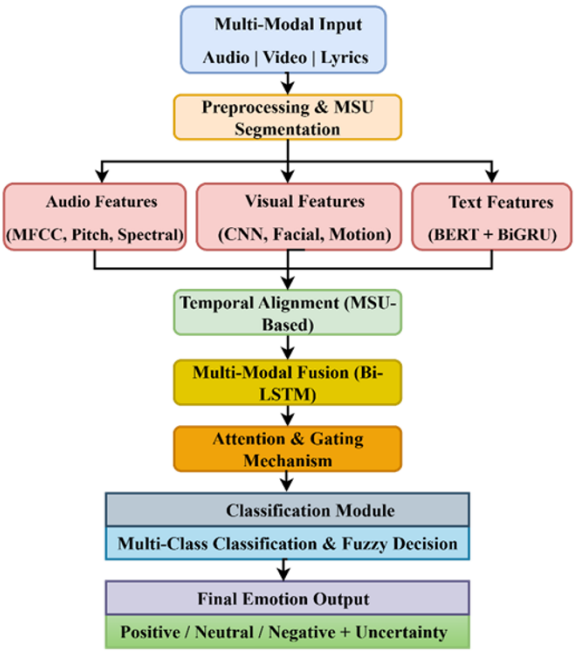

Fig. 1. Workflow of the proposed multi-modal music sentiment analysis system

Therefore, it is necessary to slice and dice the process of musicians playing music. The concept of minimal sentiment unit (MSU) was proposed. The MSU is the smallest part of a music composition that is effective in discriminating emotional states, which allows proper classification of sentiments. It is associated with a section that has a steady emotional sound, like a verse or a chorus. The segmen-

tation process determines emotional transitions based on the analysis of audio and lyrics, boundaries based on spectral flux peaks, pitch change, and boundaries based on the sentences of the lyrics. These boundaries are synchronized with milliseconds timestamps to provide uniform MSU segments. The dataset has 1,000 songs, divided into positive, negative, and neutral emotions, 60 percent of which are audio, 20 percent video, and 20 percent lyrics. It is labeled by experts and automated sentiment tools. Normalization, noise reduction, temporal alignment, and text tokenization are used to preprocess the data. The Adam optimizer is used to train the model with a learning rate of 0.001 and a batch size of 32 to achieve stable performance.

3.1.2. *Music Sentiment Analysis based on auditory features*

Audio features are typically derived out of video clips, such as spectral bandwidth, zerocrossing rate, MFCC, timbre features (through constant Q transform), spectral roll-off, and musical features such as beat, pitch, and spectral bandwidth. MFCCs have been selected due to their ability to capture the perceptual properties of sound, which is important in the detection of emotions. There are also spectral features and filter banks, which give finer frequency detail to capture emotional detail in melody and rhythm. These properties are effective in sentiment classification, which is consistent with human hearing and enhances the accuracy of emotion recognition. The paper separates the vocal melody to concentrate on the emotional type of the melody and eliminate the effect of the difference in vocal range to enable more analysis of the relationship between the melody and the lyrics. The isolation of vocal melody is performed with the help of such frameworks as Wave-U-Net and Open-Unmix that isolate vocals and accompaniment with the help of encoder-decoder structures and spectral masking. This minimizes background interference and improves feature extraction. It is difficult to concentrate on one instrument and disregard the others in musical emotion classification. To solve this, the audio input is treated in two forms: trying to find the finer granularity and examining the relation between melody and lyrics to make them more interpretable. It is based on a hierarchical structure that is inspired by the main song and chorus theory. The context is described in the main part of the song, whereas the emotional focus is highlighted in the chorus, which is a successful way of expressing emotion in a limited time. As Shown in Figure 2.

A 16 kHz sampling rate and frequency are used to process audio input. The feature vectors of 400 dimensions are produced in 25 ms time frame, which is too dense. Timefrequency analysis is instead performed by a spectrogram, which transforms the signal in the time domain into

Fig. 2. Music Sentiment Analysis Framework based on auditory features, including MFCC, DSNN, and fusion processing

a power spectrum through Fast Fourier Transform. This narrows the feature vector down to 80 dimensions. The next step is the Log operation and Discrete Cosine Transformation which leads to a 39-dimensional Mel Frequency Cepstrum Coefficient (MFCC) feature vector. The MFCC 39 dimensions are selected because of its efficiency and strength, which represents the spectral envelope without redundancy. Delta and delta-delta MFCCs are more dimensional and sensitive to noise, hence the base MFCC setup guarantees consistent training and prevents overfitting in the multi-modal fusion setup. Most methods use MFCC and filter bank representations for audio features, achieving good results.

3.1.3. *Music Sentiment Analysis based on visual features*

The paper identifies three visual features: (1) shot characteristics, such as shot switching rate and average duration; (2) color characteristics, such as luminance, color temperature, hue, saturation and RGB color histogram; (3) inter-frame motion intensity characteristics, which compute the mean and variance of key frames between adjacent frames. Video frames are extracted at 30 frames per second, resized to 224×224 pixels. Facial region detection is done using a pretrained model to isolate the face for more accurate emotion recognition. The model employs three CNN sub-networks of 3 , 5 and 10 convolutional layers. These sub-networks receive facial images in the initial stage. The second stage involves combining the outputs to forecast expressions. LBP features and CNN-extracted features are combined with a better VGG-16 network, and the combined features are inputted into a SoftMax classifier to classify the six basic expressions. As Shown in Figure 3.

Fig. 3. Music Sentiment Analysis Framework based on visual features, including CNN, feature fusion, and fusion processing

Following the separation of the vocals, the paper discusses the problem of matching the lyrics and melodies, which may differ in length. In order to attain proper alignment, the timestamps are used to synchronize the lyrics

at the sentence level to enable the lyrics to be analyzed in terms of emotion. This method guarantees that the lyrics of the song are consistent so that the study of emotional changes can be done without compromising the integrity of the song as was done in the past where the whole song or random parts were used.

3.1.4. Music Sentiment Analysis based on textual features

Textual features are extracted using Word2Vec for basic word vectorization, while BERT is used for more advanced feature extraction, providing context-aware representations of the lyrics. The BERT model is trained on a learning rate of $2e-5$, a batch size of 16, and a sequence length of 128 tokens. The BiGRU architecture is used with 256 hidden units, which is optimal to the sentiment classification. These environments provide a stable training and are necessary to reproducible experimental results. BERT is more effective at sentence-level classification than traditional word vectors such as word2vec, which use a weighted average of word vectors, because it uses contextual information. BERT introduces a CLS token at the start of the sentence and SEP token at the end of the sentence. The last latent state of CLS is the sentence vector, which is applied in classification tasks. The approach has a more detailed semantic meaning than the conventional word vectors. The BERT and BiGRU are optimized end-to-end, with the BiGRU encoder gradients backpropagated to optimize the BERT parameters. This allows contextual embeddings to be dynamically adapted to downstream sentiment classification tasks. This type of joint optimization strategy increases semantic consistency and the ability of textual features to be represented. So, we encode to obtain the embedding vector and then obtain the text memory bank, Mt, by BiGRU:

$$\mathrm{Mt} = \mathrm{fBiGRU}(\ \mathrm{Embedding}\ (\mathrm{T})), \mathrm{Mt} \in \mathrm{RLt} \times 2\mathrm{dt} \quad (1)$$

where T is a text sequence, Lt is the maximum length of a padded text sequence, and dt is the dimension of hidden units in the BiGRU.

3.1.5. Music Sentiment Analysis Based on Multi- modal Information

Multi-modal sentiment analysis improves accuracy, but performance depends on effective fusion. After extracting features, proper fusion methods are needed. Audio features (pitch, intensity), video (temporal frames), and text (word2vec vectors) are processed and combined using deep models to classify emotions (e.g., positive/negative).

After extracting visual, audio, and text features, a bidirectional LSTM is used for multimodal fusion. Audio and text features are combined to Multi-modal sentiment analysis enhances accuracy; however, performance is based on successful fusion. Once features have been extracted, they require appropriate fusion techniques. Deep models process and combine audio (pitch, intensity), video (temporal frames), and text (word2vec vectors) to classify emotions (e.g., positive/negative). Multi-modal fusion is performed after the extraction of visual, audio, and text features with the help of a bidirectional LSTM. Audio and text characteristics are merged to enhance emotion recognition (e.g., anger vs happiness). SVM and CNN are used to fuse and analyze features, demonstrating that modalities are more accurate when combined. The most important step in multi-modal sentiment analysis is modal fusion. The bidirectional LSTM hidden representations from auditory, visual, and textual streams are temporally synchronized using a shared time index t, corresponding to aligned MSU segments across modalities. Let h_t^A, h_t^V, h_t^T denote the synchronized hidden states, the fused representation is computed as $h_t^{\text{fusion}} = \phi\left(h_t^A \bigoplus h_t^V \bigoplus h_t^T\right)$, where $\oplus$ denotes concatenation and ϕ is a nonlinear mapping. This formulation ensures coherent cross-modal sequence modeling by preserving temporal consistency prior to integration.

Mathematically, feature-level fusion can be represented as:

$$F_{\text{fused}} = \text{concat}\left(F_{\text{audio}}, F_{\text{visual}}, F_{\text{text}}\right) \quad (2)$$

where F_{audio}, F_{visual}, and F_{text} represent the feature vectors extracted from the auditory, visual, and textual modalities, respectively. For decision-level fusion, the individual decisions from each modality are combined using a weighted sum:

$$D_{\text{final}} = w_1 \cdot D_{\text{audio}} + w_2 \cdot D_{\text{visual}} + w_3 \cdot D_{\text{text}} \quad (3)$$

where D_{audio}, D_{visual}, D_{text} are the outputs from each modality's classifier, and w_1, w_2, w_3 are the corresponding weights. Multi-modal fusion improves decision accuracy by combining more information. For each MSU, bidirectional LSTM produces multiple outputs, including forward and backward predictions and classifications, capturing complete temporal information. Bidirectional LSTM is a type of Recurrent Neural Network (RNN) that processes data in both forward and backward directions to capture temporal dependencies before and after a given point in time. Assuming that the current slice is Xi , the forward short-term memory classification Yi and the forward short-term memory prediction Yi +1 are obtained from the forward LSTM; the reverse short-term memory classification Y′ and the re-verse short-term memory prediction $\mathrm{Yi}' - 1$

are obtained from the reverse LSTM; and the difference between the prediction and the true value is used to Y classification output in Loss. The optimization problem is formulated as a composite loss that combines classification and prediction errors of both forward and reverse streams with weighted terms that control their relative importance. The gradient propagation is done in collaboration with the bidirectional structure, where the parameters are updated in both directions of the temporal dependencies. This design imposes predictive and classification consistency and convergence stability in training. The formula is as follows:

$$\begin{aligned} \mathrm{Y_{i,dinalm}} = 1 - \left(\frac{\hat{Y}' - 1 - \mathrm{Y_{i-1}}}{\mathrm{Y'_{i-1}} - \mathrm{Y_{i-1}} + (\mathrm{Y_{i+1}} - \mathrm{Y_{i+1}})\,, \mathrm{m'_i}} \right) \\ + (1 - \mathrm{Y'})(\mathrm{Y'_{i-1}} - \mathrm{Y^{i+1}_{i-1}} - \mathrm{Y'_{i+1}} \\ + \mathrm{Y_{i+1}} - \mathrm{Y'_{i+1}} \mathrm{m'_i}) \end{aligned} \tag{4}$$

Where, Y_i represents the forward short-term memory classification, Y_{i+1} represents the forward short-term memory prediction, Y' represents the reverse short-term memory classification, $Y_{i'-1}$ represents the reverse short-term memory prediction. From equation (4), we know that the final output classification $\hat{Y}_i^{\text{final}}$ is the confidence component of the positive classification $\hat{Y}'_i$ with its $i+1$ prediction, plus the inverse $\hat{Y}'_i$ categorization with its $i-1$-predicted confidence score. This is expressed as:

$$\hat{Y}_i^{\text{final}} = \hat{Y}'_i + \left(1 - \frac{Y_{i+1} - Y'_{i+1}}{\left(Y'_{i-1} - Y_{i-1}\right) + (Y_{i+1} - Y_{i+1})} \right) \tag{5}$$

This equation combines audio and text with time-varying visual information in featurelevel fusion to analyze sentiment. In particular, first, on each expression of multi-modality, the LSTM is applied to encode each expression with contextual features and encode the information in the unimodality individually. The independent information of these unimodalities is then fused and the interaction information between the multi-modality is obtained using the LSTM to create the multi-modal feature representation. Lastly, the dimensionality of the multi-modal representation is reduced by applying the maximum pooling strategy to build the sentiment classifier.

Bimodal and trimodal LSTM models are built for emotion decision-making. In bimodal LSTM, two modalities share their hidden states to form a combined state, then return to their own networks for classification.

$$\begin{aligned} &\left(h^0 t_1 = (h_{t_{-1}} + hB_{-1})\,/2\right. \\ &\mathrm{i_t} = \tan \mathrm{h}\,(\,\mathrm{W_{xi}x_t} + \mathrm{W_{hi}ht_{-1}t_1} + \mathrm{b_i}' \\ &\mathrm{j_t} = \mathrm{sigm}\left(\mathrm{W_{zj}x_t} + \mathrm{W_{hj}ht^0t_1} + \mathrm{b'_j}\right. \\ &\mathrm{Bi-ISTMA} \\ &\left\{ \begin{array}{l} f_t = \mathrm{sigm}\left(W_{\mathrm{kj}} x_t + W_{hf} h^0 \cdot_1^t + b'_f\right. \\ o_t = \tan h\left(W_{oj} x_t + W_{ho} h t^0 t^t + b'_o\right. \\ c_t = c_{t-1} \Theta f_t + i_t \Theta j_t \end{array} \right. \\ &\left('\mathrm{hA} = \tan h(ct) \Theta jt\right. \end{aligned} \tag{6}$$

Where, h_A and h_B are the hidden states of two modalities. h_{tot} is the shared hidden state representation after fusion. i_t, j_t, f_t, o_t are the different gates in the LSTM unit. W and b are the learned weight matrices and biases for each gate. Hidden states are shared once before updating, then each modality returns to its own bidirectional LSTM. The same process is used in the trimodal LSTM. The common hidden-state exchange is achieved by partially sharing the parameters, where the modality-specific gate parameters are independent and the hidden projection layers are collectively trained to build a common latent representation. In order to overcome gradient interference in multi-modal training, a stabilization constraint of gradient normalization and orthogonality regularization is used on the shared states. This process guarantees equal modality contribution and convergence stabilization in both bimodal and trimodal LSTM structures. GRU and attention enhance sentiment analysis by paying attention to significant words and time intervals. Noise in various modalities is minimized with an attention layer and a reinforcement learning-based gate. The reinforcement learning-based input gate controller is formulated with the multi-modal feature vector at time t as the state, while the action corresponds to adaptive gating weights assigned to each modality. The reward function is defined based on the improvement in sentiment classification accuracy and noise suppression effectiveness across modalities. This design enables dynamic modulation of modality contributions, effectively reducing noise interference during multi-modal training. Sentiment analysis is performed with an end-to-end RNN model, which takes into account context, speaker-Listener emotion, and modality relationships. It models emotional states and sentiments using sGRU and cGRU units. As Shown in Figure 4.

The bidirectional LSTM updates forward and backward states in the same way. A multimodal system with 7 classifiers is used for integrated emotion decision-making. Fuzzy logic is added to handle uncertainty, and the emo-

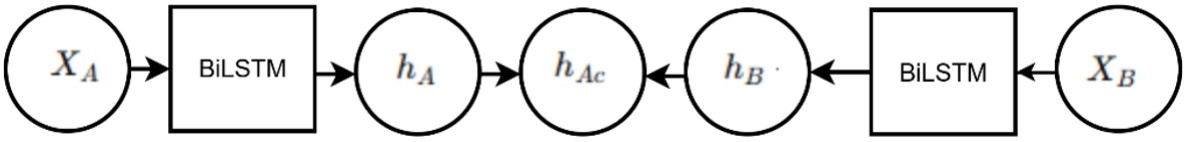

Fig. 4. Music Sentiment Analysis Framework based on multi-modal information, utilizing bidirectional LSTM for feature extraction and fusion.

tion pipeline (based on confidence) converts fixed inputs into uncertain, more realistic emotional outputs. The 7 classifiers give 7 outputs, which are denoted as $\beta i = [A, L, V, VL, AL, AV, AV\ L\ ]$. They represent the Audiovisual Semantic Integrated Classifier, the Audio- visual Integrated Classifier, the Auditory Semantic Inte- grated Classifier, the Visual Semantic Integrated Classifier, the Visual Single Classifier, the Auditory

Semantic Single Classifier, and the Auditory Semantic Single Classifier, respectively. Multiple classifiers define positive, negative, and neutral regions, while their overlaps form fuzzy regions. This helps model real and uncertain emotions, and a final multi-modal fuzzy decision is made using all seven outputs. The fuzzy membership for each affective region is defined as $\mu_k = \frac{\exp(\beta_k)}{\sum_j \exp(\beta_j)}$, where $k \in \{$ positive, neutral, negative $\}$ and β_k denotes the classifier output score. The membership values are further refined using triangular membership functions to model overlapping regions between neutralpositive and neutral-negative emotions. This quantitative scheme enables soft decision boundaries and accurately represents uncertainty in multimodal sentiment integration.

The uncertainty of the input data is measured and analyzed using the fuzzy sentiment integration decision. On this uncertainty, a sentiment pipeline is developed, with the channel size reflecting the degree of uncertainty in an intuitive manner. As Shown in Figure 5.

The overlapping positive-neutral and negative-neutral emotions create fuzzy emotion areas, which are useful in modeling the real and uncertain human emotions. The model simplifies the model by a three-channel linear fusion LSTM to balance the uncertainty of personalization and accuracy of classification. This model is highly accurate (up to 100 percent in the testing, under specific conditions) and has a good balance between accuracy and dealing with uncertainty, which is why it can be used in further experiments.

3.2. Model Regression Results and Analysis

The proposed algorithm is compared to the standard LSTM model to validate it. Decision-level (post) fusion is a fusion

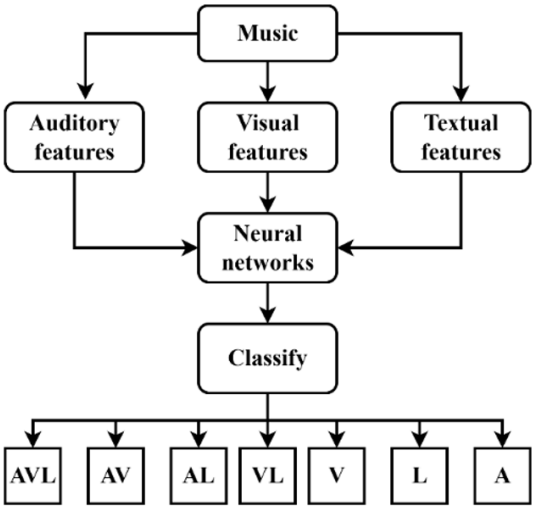

Fig. 5. Complete system framework for multi-modal music sentiment analysis, integrating auditory, visual, and textual features through neural networks for classification

Table 1. Bidirectional trimodal LSTM

mood	Precision	Recall	F1
positive	0.899	0.912	0.900
negative	0.966	0.873	0.915
neutral	0.861	0.930	0.900

of the results of each modality, which occurs after individual analysis and enables decisions to be made even in the absence of one modality, but is complicated and time-consuming because of multiple classifiers. As can be seen in experimental results (Tables 1 and 2), the visual modality addition enhances performance, whereas the bidirectional trimodal LSTM is more effective but can cause overfitting because of the increased complexity.

Table 2. Linear fusion of three single channels LSTM

mood	Accuracy	Precision	Recall	F1
positive	0.869	0.733	0.924	0.893
negative	0.867	0.833	0.912	0.897
neutral	0.860	0.669	0.665	0.757

Table 3. Performance comparison of proposed method with CNN-LSTM and transformer-based models

Model	Accuracy	Precision	Recall	F1-score
LSTM (baseline)	0.842	0.801	0.856	0.828
CNN-LSTM Hybrid	0.873	0.845	0.889	0.866
Transformer (multi-modal)	0.905	0.882	0.918	0.900
Proposed Trimodal BiLSTM	0.918	0.899	0.930	0.910

Table 4. Combined ablation study on modality contributions and architectural components

Configuration	Accuracy	F1-score
Audio only	0.871	0.865
Visual only	0.842	0.836
Text only	0.883	0.878
All modalities (no attention)	0.901	0.895
All modalities + attention (no gating)	0.907	0.900
All modalities + attention + sGRU/cGRU (no fusion)	0.912	0.905
Full model (with shared-state fusion)	0.918	0.910

The proposed model is the most effective in music sentiment analysis in all measures in this seven-class classification task. It is clear that the bidirectional trimodal LSTM is better than the linear fusion LSTM. Both audio and lyrics can be used to identify emotions, but audio is more accurate at predicting intensity (arousal) since it includes rhythm and energy whereas lyrics are more effective at identifying emotions such as happiness and calmness.

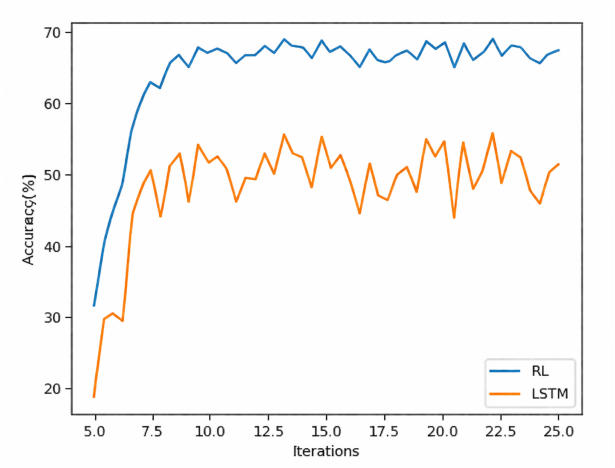

Fig. 6. Accuracy with the number of iterations

Figure 6 indicates that bidirectional trimodal LSTM is highly effective. It simulates the continuous emotional change (emotional coherence) and models the response of various personalities to various stimuli, which assists in modeling human-like emotions in machines.

4. Results and discussion

This paper proposes a multi-modal emotion model with experiments showing improved personalized emotion classification using LSTM, integrated confidence decision, and emotion transfer modeling. Benchmark models (CNN-LSTM and transformer-based) are compared in Table 3, where the trimodal BiLSTM achieves superior accuracy and F1-score, proving effective multi-modal fusion and temporal modeling.

However, due to various objective conditions including hardware, the sentiment classification model and the sentiment state expression model in this paper only utilize the loworder features of the dataset and have not been experimented on some large-scale video sentiment analysis datasets. Table 4 shows that modalities and key components improve performance, with the full model achieving the highest accuracy and F1-score.

5. Conclusion

The construction of artificial sentiment simulation in this paper is only a constructive exploration, and there is still much work to be done to further improve it. Its generalization ability needs to be verified with a large amount of data or more new problems for sentiment integration decision making. In addition, the relationship and significance of reinforcement learning for emotion generation is significant, but the reinforcement learning used in this paper is relatively simple, and subsequent research on reinforcement learning is needed.

Declarations

Funding: This work was supported by the Jilin Provincial Department of Education Project (JJHK 20191047 SK)
Conflicts of interests: The author declared no conflicts of interest regarding this work.
Data Availability Statement: The data used to support the

findings of this study are available from the corresponding author upon request.

Code availability: Not Applicable.

Authors' Contributions: Lu Huang, is responsible for designing the framework, analyzing the performance, validating the results, and writing the article.

Ethical Approval: This article does not contain any studies involving human participants or animals performed by any of the authors.

Consent to Participate: Not applicable.

Consent to Publication: All authors have provided consent for publication of this manuscript.

Competing Interests: The authors declare no competing interests

References

[1] D. Han, Y. Kong, J. Han, and G. Wang, (2022) *"A survey of music emotion recognition"* **Frontiers of Computer Science *16***: 1–11. DOI: 10.1007/s11704-021-0569-4.

[2] Y. Hu, (2022) *"Music emotion research based on reinforcement learning and multimodal information"* **Journal of Mathematics *2022***: 1–10. DOI: 10.1155/2022/2446399.

[3] L. M. Gómez and M. N. Cáceres. "Applying data mining for sentiment analysis in music". In: *International Conference on Practical Applications of Agents and Multi-Agent Systems*. Cham: Springer, 2017, 198–205. DOI: 10.1007/978-3-319-61578-3_20.

[4] K. Napier and L. Shamir, (2018) *"Quantitative sentiment analysis of lyrics in popular music"* **Journal of Popular Music Studies *30***: 161–176. DOI: 10.1525/jpms.2018.300411.

[5] S. Shukla, P. Khanna, and K. K. Agrawal. "Review on sentiment analysis on music". In: *2017 International Conference on Infocom Technologies and Unmanned Systems (ICTUS)*. IEEE, 2017, 777–780. DOI: 10.1109/ICTUS.2017.8286111.

[6] W. Chen, (2022) *"A novel long short-term memory network model for multimodal music emotion analysis in affective computing"* **Journal of Applied Science and Engineering *26***: 367–376. DOI: 10.6180/jase.202303_26(3).0008.

[7] R. Kaur and S. Kautish. "Multimodal sentiment analysis: A survey and comparison". In: *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*. IGI Global, 2022, 1846–1870. DOI: 10.4018/978-1-6684-6303-1.ch098.

[8] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya. "Contextual inter-modal attention for multi-modal sentiment analysis". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, 3454–3466. DOI: 10.18653/v1/D18-1382.

[9] B. R. Gudivaka, (2021) *"Designing AI-assisted music teaching with big data analysis"* **Current Science Humanities *9***: 1–14.

[10] J. Liu, P. Zhang, Y. Liu, W. Zhang, and J. Fang, (2021) *"Summary of multi-modal sentiment analysis technology"* **Journal of Frontiers of Computer Science and Technology *15***: 1165. DOI: 10.3778/j.issn.1673-9418.2012075.

[11] H. Wen, S. You, and Y. Fu, (2021) *"Cross-modal context-gated convolution for multi-modal sentiment analysis"* **Pattern Recognition Letters *146***: 252–259. DOI: 10.1016/j.patrec.2021.03.025.

[12] A. S. Alqarafi, A. Adeel, M. Gogate, K. Dashitpour, A. Hussain, and T. Durrani. "Towards Arabic multimodal sentiment analysis". In: *International Conference on Communications, Signal Processing, and Systems*. Singapore: Springer, 2017, 2378–2386. DOI: 10.1007/978-981-10-6571-2_290.

[13] I. Chaturvedi, E. Cambria, R. E. Welsch, and F. Herrera, (2018) *"Distinguishing between facts and opinions for sentiment analysis: Survey and challenges"* **Information Fusion *44***: 65–77. DOI: 10.1016/j.inffus.2017.12.006.

[14] J. Wu, T. Zhu, X. Zheng, and C. Wang, (2022) *"Multi-modal sentiment analysis based on interactive attention mechanism"* **Applied Sciences *12***: 8174. DOI: 10.3390/app12168174.

[15] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, (2021) *"Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN"* **International Journal of Interactive Multimedia and Artificial Intelligence**: DOI: 10.9781/ijimai.2020.07.004.

[16] W. Yuzhu, X. Jun, C. Bo, and X. Xinying, (2021) *"Multi-modal sentiment analysis based on cross-modal context-aware attention"* **Data Analysis and Knowledge Discovery *1***: DOI: 10.11925/infotech.2096-3467.2020.1042.

[17] J. Zhang, Z. Yin, P. Chen, and S. Nichele, (2020) *"Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review"* **Information Fusion *59***: 103–126. DOI: 10.1016/j.inffus.2020.01.011.

[18] A. Kumar and J. Vepa. "Gated mechanism for attention based multi modal sentiment analysis". In: *ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, 4477–4481. DOI: 10.1109/ICASSP40776.2020.9053012.

[19] D. H. Kim, M. K. Lee, D. Y. Choi, and B. C. Song. "Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild". In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, 529–535. DOI: 10.1145/3136755.3143005.

[20] S. Latif, H. Cuayáhuitl, F. Pervez, F. Shamshad, H. S. Ali, and E. Cambria, (2022) *"A survey on deep reinforcement learning for audio-based applications"* **Artificial Intelligence Review**: 1–48. DOI: 10.1007/s10462-022-10224-2.

[21] M. Sivakumara and S. R. Uyyalab, (2022) *"Aspect-based sentiment analysis of product reviews using multi-agent deep reinforcement learning"* **Asia Pacific Journal of Information Systems *32***: 226–248. DOI: 10.14329/apjis.2022.32.2.226.

[22] S. J. Park, D. K. Chae, H. K. Bae, S. Park, and S. W. Kim. "Reinforcement learning over sentiment-augmented knowledge graphs towards accurate and explainable recommendation". In: *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. ACM, 2022, 784–793. DOI: 10.1145/3488560.3498515.

[23] F. Nadeem. *"Multi-modal reinforcement learning with videogame audio to learn sonic features"*. (phdthesis). Massachusetts Institute of Technology, 2020.

[24] E. Acar, F. Hopfgartner, and S. Albayrak. "Fusion of learned multi-modal representations and dense trajectories for emotional analysis in videos". In: *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2015, 1–6. DOI: 10.1109/CBMI.2015.7153603.

[25] B. Schuller, F. Weninger, and J. Dorfner. "Multi-modal non-prototypical music mood analysis in continuous space: reliability and performances". In: *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*. 2011.