https://doi.org/10.31449/inf.v49i5.6864

Multimodal Sentiment Perception for Intelligent Music Generation Using Improved Transformer Architectures

Yingshi Jiang*, Zuodong Sun
School of Music, Tonghua Normal University, Tonghua 134001, China
Email: 18704352221@163.com, sunzuodong126@163.com
*Corresponding author

Keywords: multimodal sentiment, music generation model, emotion perception, deep learning, sequence analysis

Received: August 1, 2024

To further examine the interrelationship between music, emotion, and scene, as well as to furnish novel technical assistance for music creation, the study devised a multimodal sentiment analysis model for auditory and visual features with deep learning. Based on this model, a new music content generation model was proposed, which improved upon the traditional Transformer architecture. The validation of the research-designed MSA dataset, as well as the three publicly available datasets IEMOCAP, CMU__MOSI, and CMU-MOSEI, confirmed the accuracy of the sentiment analysis of this multimodal sentiment analysis architecture. The mean absolute error, the root means square error, and the mean absolute percentage error were found to be 0.149, 0.166, and 0.140, respectively, and the goodness-of-fit R-squared reached 0.961. The model performed well on Precision-Recall curve, receiver operating characteristic curve. The sentiment recognition accuracy was up to 0.98, and the recognition efficiency was high. Additionally, the pitch variation of the music generated by the improved Transformer structure was the smallest compared to the Classical Piano MIDI dataset, taking a value of 1.29%. Moreover, the melodic variation was the smallest compared to the Nottingham dataset, taking a value of 0.66%. The generated music performed better in terms of smoothness, coherence, and percentage of completeness. Using this model for music generation, the highest values of hit rate and normalized discounted cumulative gain could be 93.984% and 91.566%. The mean inverse rank could be up to 0.89. This study deepens the mechanism of music emotion generation, captures the emotion and context of music more accurately, and promotes the development of the fields of emotion computing and sentiment recognition.

Povzetek: Raziskava uvaja izboljšan jezikovni pretvornik za multimodalno analizo sentimenta ter generiranje glasbe, kar izboljša čustveno zaznavanje in kakovost glasbe.

1 Introduction

Emotion is an important driver of human behavior and thinking, and accurately understanding and describing emotion can lead to a better understanding of the impact of emotion on decision-making and behavior. The need for emotion research in human-computer interaction, mental health, education, and business applications is steadily growing as society and intelligent technology evolve. Multimodal sentiment analysis (MSA) is an intelligent study of recognizing and understanding human emotions through inputs such as speech, text, and images. By elucidating the nuances of human emotion, MSA can enhance the capacity of computer systems to comprehend and respond to human emotions, thereby optimizing the quality of human-computer interaction [1-2]. The expression of emotion has multi-dimensional characteristics, involving language, voice, facial expression, and other forms. MSA can synthesize multiple sensory inputs to understand and analyze emotional expressions more comprehensively and accurately, enriching the application scenarios and possibilities of emotion perception. Meanwhile, MSA involves the crossover of multiple disciplinary fields such as linguistics, psychology, computer vision, etc., which promotes the cooperation and integration between different disciplines [3-4]. In recent years, “artificial intelligence + music” has become a hot research topic. Auto-generated music technology can create novel and unique music works and bring a fresh listening experience [5-6]. However, the current auto-generated music content is single, lacks user stickiness, and is difficult to generate music content with emotional style. Additionally, MSA methods still face the dilemma of shortage of dataset resources and lack of multimodal data characterization capability [7-8].

To generate music with emotional and melodic characteristics, the research centers around MSA and music automatic generation (MAG) techniques. Firstly, a multi-label sentiment analysis dataset for MSA is constructed, then the MSA framework is constructed based on deep learning (DL), sequence analysis, and feature extraction models, and finally the model is designed based on the improved Transformer. The study unfolds the design of the model based on MSA, which improves the theoretical research level of MSA. It fills the gap of lack of emotion analysis in and expands the human understanding of music generation. The study is expected to provide new technical support for music creation, making music creation more emotionally expressive and situationally perceptive.

The study is divided into a total of four parts. The first part completes the review of the current state of research related to MSA and MAG. The second part designed the MSA framework and constructed the model based on MSA. The performance test of the MSA framework and model is examined in the third section. The study's key findings and future directions are outlined in the fourth part.

2 Related works

As artificial intelligence technology advances, DL is being utilized extensively in several domains such as speech recognition, natural language processing, image recognition, and more. MSA has gradually become a popular research direction. MSA belongs to the popular trend of social research. In order to cope with the interference of mixed emotions, Jiang et al. designed a context-self-attentive fuzzy temporal convolutional network for MSA. The network contained fuzzy emotion affiliation functions, which could obtain the dependency relationship between the key information of itself and the information of the external context within the target discourse. The dataset validated the effectiveness of the method [9]. Existing Multimodal sentiment recognition methods are prone to bring redundant information and low recognition efficiency. Fu et al. designed an innovative asymmetric transformer with cross-modal blocks for complementary learning of the local structure of sequences with different modalities. The method's usefulness and superiority were confirmed by the experimental results [10]. To utilize various physiological signals for sentiment recognition, Chen et al. proposed a multilevel multimodal dynamic fusion network by considering cross-modal interactions. The method utilized this network to independently extracted potential and essential interactions between different modal features and split the multistage fusion network using correlations. Experimental results indicated that the network could utilize finer-grained single-, bimodal, and tri-modal interactions and outperforms single-stage multimodal sentiment recognition methods [11].

Kumar et al. classified the emotions contained in speech and images into discrete classes, designed a hybrid fusion based multimodal sentiment recognition system. Experimental results indicated that the system achieved 83.29% accuracy in sentiment recognition [12]. Currently, there is a lack of labeled datasets for speech sentiment recognition due to the time-consuming nature of recognizing emotion categories. Therefore, Yi et al. tried to apply the wav2vec2.0 model to speech feature extraction for speech sentiment recognition task. The technique created a multimodal two-branch transformer network for emotion categorization. Two databases, IEMOCAP and CASIA, verified the recognition accuracy of the method [13]. Due to the large feature dimension, there is multimodal sentiment recognition in multimodal sentiment recognition. Using a bidirectional gated recurrent cell recognition network based on bidirectional gated recurrent units, Tang et al. developed a novel multimodal sentiment and presented the process of multimodal self-attention. According to experimental data, the study's proposed technique might successfully raise the network's multimodal sentiment recognition accuracy [14]. Multimodal sentiment recognition involves various different application areas. Dixit and Satapathy first used three publicly available benchmark datasets, ISEAR, RAVDESS and FER-2013, to train independent models in three modes of emotion: text, audio and image. The study then stacked deep neural network (NN), recurrent NN, and 2D convolutional NN based on the idea of integration for effective emotion prediction. The method's prediction accuracy was found to be 86.6% with an F1 value of 0.84, as per the experimental findings [15].

Automated music generation using artificial intelligence offers musicians of all levels an innovative way to enhance the creative process, but also creates a way for ordinary music lovers to "make music" for their personal music needs. Automated music creation has received a lot of attention, and the use of AI technology in creative music creation is still an area that can be explored. Shukla and Banka changed the traditional way of generating music using predefined music parameters. The study created a brand-new genetic algorithm-based approach for creating music that features enhanced crossover and mutation operators. The method was flexible to change the musical parameters can be generated by inputting short segments of the melody to generate a musical melody, and comparative analysis verified the validity of the model [16]. Automatic generation of music relies on contextual representation capabilities. Wu et al. first delineated music by setting up a fragment range localization module. Validated on a public MIDI dataset, the method outperformed other comparative models in short- and long-term quantitative music generation metrics [17]. DL techniques have application advantages in creative music generation. Li S et al. proposed a pre-trained model for multitasking music generation to learn melodic and rhythmic representations by taking advantage of pre-training to overcome remote dependency limitations in natural language processing. Experimental results indicated a significant improvement in the performance of the method under the HITS@k evaluation metric [18].

The relevant work is summarised in Table 1. The research on MSA and technology at home and abroad has made technical breakthroughs in the main performance aspects. However, the robustness of MSA under the interference of data noise and unlabeled dataset still needs to be improved. Meanwhile, Meanwhile, most of the automatic music generation models use text sequence methods, ignoring the correlation between chords and melodies, and its difficult to generate music segments that contain emotional information. Therefore, in this regard, MSA-based music generation is investigated.

Table 1: Summary of related work

Literature	Author	Methodology	Data set	Results	Insufficient research
[9]	Jiang et al.	Fuzzy temporal convolutional network based on contextual self-attention	/	Model outperforms state-of-the-art models on tested datasets.	More complex model design
[10]	Fu et al.	Efficient neural network to learn modality-fused representations with CB-Transformer and Innovative asymmetric transformer with cross-modal blocks	IEMOCAP, CMU-MOSI and CMU-MOSEI	Superiority and efficiency in word-aligned and unaligned experiments	Higher model complexity
[11]	Chen et al.	A multi-stage multimodal dynamical fusion network	Multimodal benchmark DEAP	The method outperforms related single-stage multimodal emotion recognition methods	Insufficient attention to original features
[12]	Kumar et al.	A multimodal emotion recognition system based on hybrid fusion	IIT-R SIER dataset	The accuracy of emotion recognition reaches 83.29%	Only image and voice features are fused
[13]	Yi Y, Tian et al.	Self-supervised multimodal dual-branch transformer network	IEMOCAP and CASIA datasets	The method outperforms the accuracy of existing state-of-the-art methods	Higher model complexity
[14]	Tang et al.	Multimodal emotion recognition based on multi-head self-attention mechanism and bi-directional gated	/	Attentional mechanisms can effectively improve the accuracy of multimodal emotion recognition	Unvalidated robustness on unlabelled datasets

		recurrent units			
[15]	Dixit and Satapathy	Deep neural networks, recurrent neural networks, and 2D convolutional neural network stacking	ISEAR, RAVDESS and FER-2013 datasets for training text, CMU-MOSEI benchmark multimodal dataset for testing	The accuracy obtained on the CMU-MOSEI dataset is 86.60% with an F1 score of 0.84	Model training and stacking are too cumbersome
[16]	Shukla and Banka	Music generation based on improved genetic algorithms	/	Algorithm performance is better and the model generates music with better efficacy	Neglected emotional information in music data
[17]	Wu et al.	Hierarchical Transformer model	Open MIDI datasets	The method outperforms other comparative models in quantitative metrics of short- and long-term music generation	Neglects emotional information in music data
[18]	Li and Sung	Multi-task music generation based on pre-trained models	/	0.09-13.10% and 0.02-7.37% improvement in the performance of the generation task under the HITS@k evaluation metric, respectively.	Ignores sentiment information in music data

3 Key technology of automatic music generation modeling

With the goal of automatic music generation, the research task is divided into the construction of MSA dataset, the construction of MSA model and the design of music generation model based on sentiment analysis.

3.1 Construction of a multimodal multilabel sentiment analysis dataset

MSA dataset is the basic task of MSA work, and a high-quality MSA dataset should be characterized by diversity, annotation accuracy, balance, and multimodal consistency to ensure model optimization and performance improvement. However, there is a clear imbalance in the contribution of different modal information. That is, auditory and visual representation feature engineering cannot adequately extract and utilize modal affective information such as images and audio. This results in the textual modality usually contributing more than the auditory or visual modalities. This also leads to missing information from other modalities, incomplete emotion prediction, and weakened inter-modal complementarity, which exacerbates the modal imbalance phenomenon, and thus results in limited application scenarios for MSA models [19-20]. Meanwhile, the construction of MSA dataset faces serious challenges of data resource quality [21]. In order to train and optimize the MSA model, the study firstly launched the construction of MSA dataset. The study helps the model learn the correlations between different modalities and the diversity of emotional cues more deeply by constructing MSA datasets that contain rich, diverse, and accurately labeled data.

The study uses social media platforms, online video resources, and professional performance databases as data sources. During the acquisition process, the meta-video modality is required to have standard Chinese Mandarin language, clear speaker images, and accurate text transcription. The video player Plot Player is used to complete the video cropping. The cropped video clips should cover rich video scenes, such as various interviews, video blogs, and variety shows. During the capture process, the face scenes with different angles and lighting, and slight noise are preserved as much as possible. To analyze as much as possible ambiguity, irony, innuendo and other moods, the cropped video clips should possibly focus on non-verbal behaviors. The final study obtains a total of 10,364 unsupervised resources and 4,631 supervised instances. Table 2 displays the statistical data and details of the modality extraction of the MSA dataset. In the course of data labeling, all samples are assigned acoustic, visual, textual, and multimodal affective values. A total of ten annotators participated in labeling the dataset, deleting the highest and lowest scores, or taking the average of the labeled scores as the final labeling results, and mapping the results of labeling to seven different affective levels. Additionally, the study adopts a modal isolation strategy in the annotation process to ensure that different modal information does not interfere with each other.

3.2 Construction of a model for multimodal sentiment analysis

Traditional MSA mostly adopts a multimodal fusion framework, which is easy to ignore the original features [22]. In this regard, the study proposes an auditory and visual MSA analysis framework (AV-MSA). The structural composition is shown in Figure 1.

Table 2: Statistical information and modal extraction details of msa dataset

Type	Unsupervised resources	Supervision examples	Modality	Feature dimension	Extraction tool
Video	163	175	Hearing Visual Text	920*25 230*180 50*750	OpenSMILE TalkNet Pre training model
Fragment	4631	10364	/	/	/
Average fragment duration	3.26	4.64	/	/	/
Average number of clause words	16.1	19.4	/	/	/
Maximum fragment duration ratio	0.5/19	0.3/29	/	/	/
The ratio of the maximum word count in a clause	2/46	2/89	/	/	/
Standard deviation of sentence word count	8.2	10.1	/	/	/

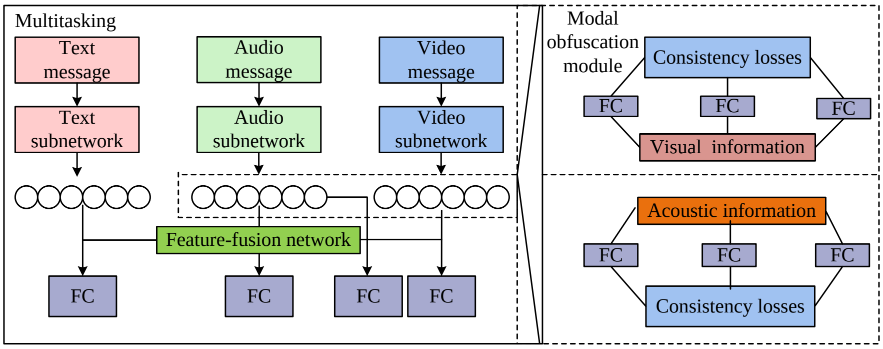


Figure 1: Structure diagram of auditory and visual msa analysis framework

In Figure 1, AV-MSA contains four components: text, audio, video, and feature fusion. Text, audio, and video are individually supervised and then trained with the feature fusion network. The text modal representation learning network extracts text features through Bert pre-training model with feature dimension of 50. The audio modal and video modal representation learning networks complete feature extraction and optimization through long short-term memory networks (LSTM) and multilayer perceptron (MLP). The basic structure of the resulting audio and video modal representation learning network is shown in Figure 2.

In Figure 2, the LSTM responsible for the hidden layer feature extraction contains the forgetting gate (FG) f_t, the input gate (IG) i_t, the output gate (OG) O_t, and the memory cells. The computation process of the internal state S_t of the LSTM with the hidden layer y_t transitions is shown in Equation (1) [23].

$$\begin{cases} S_t = f_t \times S_{t-1} + i_t \times S_t \\ y_t = \mathrm{O}_t \tanh\left(S_t\right) \end{cases} \quad (1)$$

In Equation (1), S_t denotes the intermediate state of the current internal state. t denotes the time series.

The expression of the FG f_t is shown in Equation (2).

$$f_t = \sigma\left(W_f\left[y_{t-1}, x_t\right] + b_f\right) \quad (2)$$

In Equation (2), W_f and b_f denote the weight and bias of the FG, respectively. σ denotes the activation

In Equation (5), $J\left(w,b\right)$ denotes the loss function. The parameter update process is shown in Equation (6).

$$\begin{cases} w_{ij}^l = w_{ij}^l - \lambda \dfrac{\partial J\left(w,b\right)}{\partial w_{ij}^l} \\ b_i^{\,l} = b_i^{\,l} - \lambda \dfrac{\partial J\left(w,b\right)}{\partial b_i^{\,l}} \end{cases} \quad (6)$$

function. The calculation process of IG i_t and OG O_t is shown in Equation (3).

$$\begin{cases} i_t = \sigma\left(W_i\left[y_{t-1}, x_t\right] + b_i\right) \\ O_t = \sigma\left(W_o\left[y_{t-1}, x_t\right] + b_o\right) \end{cases} \quad (3)$$

In Equation (3), W_i and b_i denote the weights and bias of the IGs, respectively. W_o, b_o are the weights and bias of the OGs respectively. After obtaining the hidden layer features of the audio subnetwork, the MLP performs feature optimization. MLP belongs to a feed-forward neural network (FNN) based DL model. It consists of an input layer, a hidden layer, and an output layer, and contains two processes, forward propagation (FP) and back propagation. The FP process of MLP is shown in Equation (4).

$$z_j^{l+1} = \sum_{i=1}^{n_l} w_{ij}^{\,l} a_i^l + b_j^{\,l+1} \quad (4)$$

In Equation (1), n_l is the neurons in layer l, i and $j \in n$. a is the computed output of neurons. z is the output of the network layer. w is the connection weights. b is the bias parameter. To minimize the error between the predicted and true outputs, the MLP back propagation method computes the gradient by applying the chain rule to compute the derivatives of the loss function with respect to the parameters and changes the weights and biases [24–25]. The residual δ_i^l of the neuron is calculated in Equation (5).

$$\delta_i^l = \frac{\partial J\left(w,b\right)}{\partial z_i^l} \quad (5)$$

In Equation (6), λ denotes the learning rate. The acquired modal sequences are then fed into the unimodal encoder $S\left(\cdot\right)$. The encoding process is described in Equation (7).

$$F_k = S_k\left(I_k\right) \quad (7)$$

In Equation (7), F_k denotes unimodal intermediate features. I_k denotes modal original sequence feature. k denotes the modality. Finally, the text intermediate feature F_t, acoustic intermediate feature F_a, and visual intermediate feature BBB are connected. Equation (8) depicts the calculating procedure.

$$F_m = Concat([F_t;F_a;F_v]) \tag{8}$$

In Equation (8), F_m denotes multimodal features. Fully connect feed-forward network (FCFN) is utilized as unimodal and multimodal sentiment classifier. The sentiment prediction result y'_k is calculated in Equation (9).

$$y'_k = FCFN(F_k) = FFN(FFN(FFN(BN(F_k)))) \tag{9}$$

In Equation (9), FFN denotes FNN. BN denotes batch data. The two supervised cases are mixed for training, and the mixed unimodal sentiment prediction results of y''_k are calculated in Equation (10).

$$y''_k = Clf(F_k) = FFN(FFN(FFN(BN(F_k)))) \tag{10}$$

In Equation (10), Clf denotes the classifier. The modal obfuscation module of AV-MSA is mainly responsible for enhancing and optimizing the feature representation of unimodal subnetworks using the mixup method. The working mechanism of the modal confusion module is shown in Figure 3.

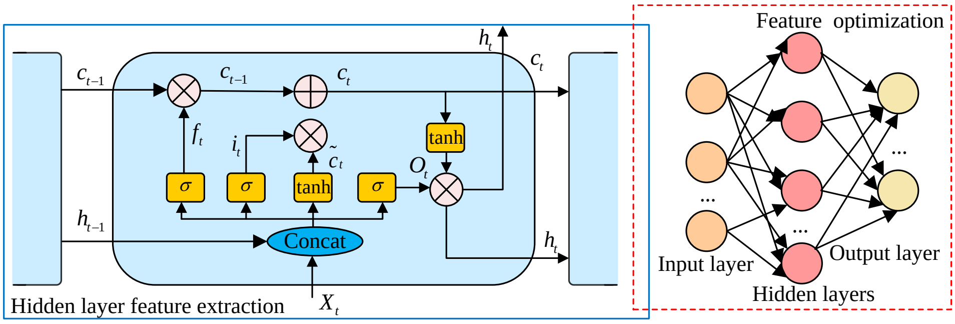
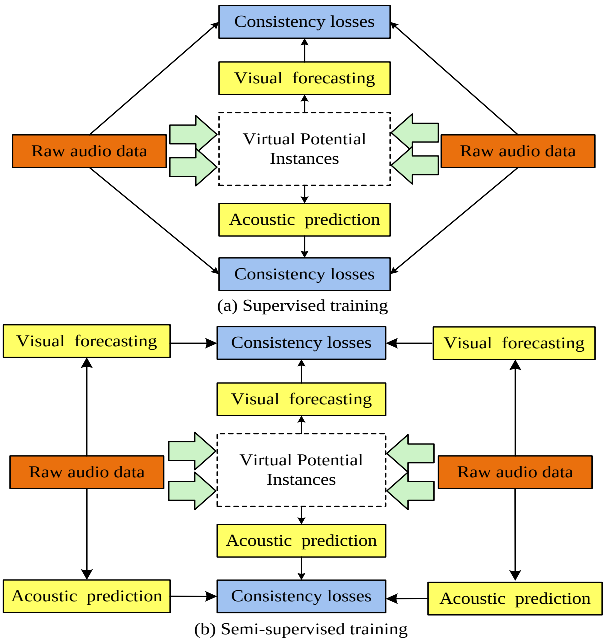


Figure 2: Composition of feature extraction results for audio modal



(a) Supervised training

(b) Semi-supervised training

Figure 3: Working mechanism of modal obfuscation module

In Figure 3, the obfuscation of supervised and unsupervised data representations further improves the fine-grained sentiment prediction performance. Supervised training directly obfuscates the raw data labels and supervises the prediction results. Unsupervised training is assisted by supervised training. The original data labels, original data prediction results are also utilized to generate new labels. The process of mixing instances of modal obfuscation module is shown in Equation (11).

$$X_i^{''} = \beta \cdot X_i + (1-\beta) \cdot X_i^{'} \quad (11)$$

In Equation (11), $X_i^{'}$ and X_i denote different instances before mixing. $X_i^{''}$ denotes the instances after mixing. β denotes the random variable. Meanwhile, the same way of mixing the prediction results is used and the modal confusion module is shown in Equation (12).

$$\left(X_1^{''}, y_1^{''}\right), \ldots, \left(X_n^{''}, y_n^{''}\right) = Mixup\left\{\left(X_1, y_1\right), \ldots, \left(X_n, y_n\right)\right\} \quad (12)$$

3.3 Modeling for automatic music content generation

Automatic music content generation based on AV-MSA framework is to complete the generation of music content based on user sentiment analysis, and the study analyzes music in MIDI format as an example. The unfolding, repetition, and change of music segments naturally constitute time series data, which is regarded as a sequence model [24]. Based on the similarity between music time series data and natural language sequences, the study adopts the Transformer model as the technical framework for automatic MIDI music generation. Figure 4 schematically depicts the Transformer model's construction.

In Figure 4, the core idea of Transformer is the attention mechanism with the ability to process input sequences in parallel. The input sequences constitute the coding part through a multi-layered self-attention mechanism, which is processed by the decoder to generate the output sequences. Transformer does not need to consider the sequence order problem, and it can capture the long-distance dependencies better [25-26]. As a result, the automatic music generation framework constructed by the research is shown in Figure 5.

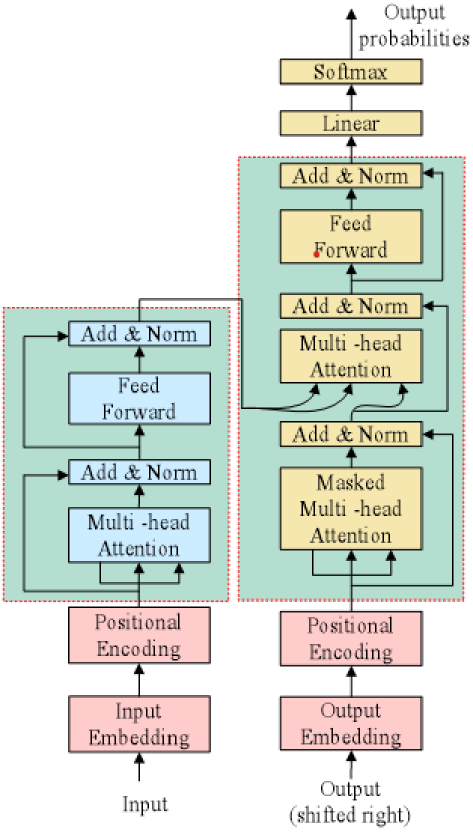


Figure 4: Structure diagram of transformer feature extractor

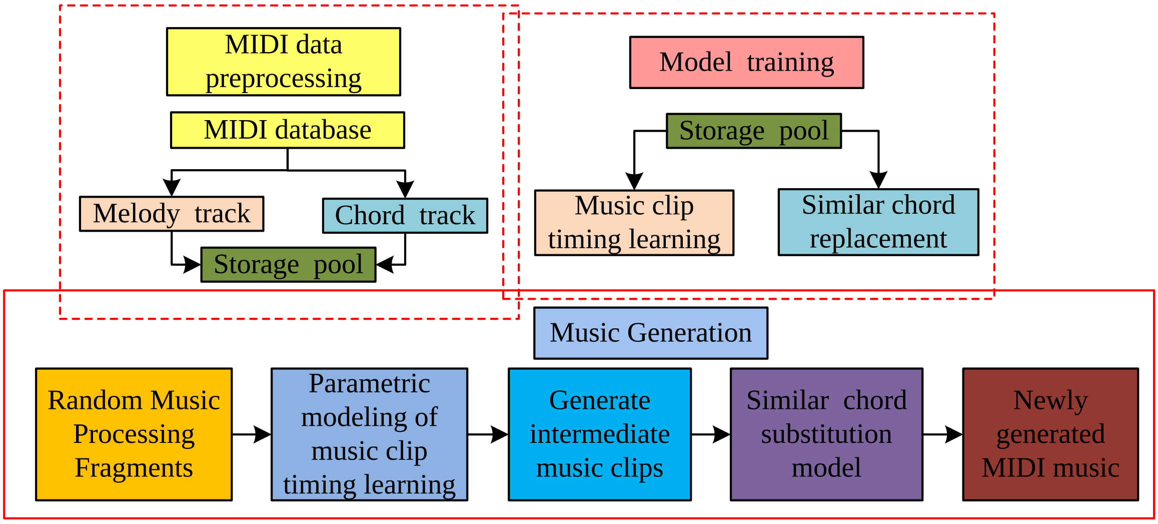

MIDI data preprocessing
MIDI database
Melody track
Chord track
Storage pool
Model training
Storage pool
Music clip timing learning
Similar chord replacement
Music Generation
Random Music Processing Fragments
Parametric modeling of music clip timing learning
Generate intermediate music clips
Similar chord substitution model
Newly generated MIDI music

Figure 5: Schematic diagram of the framework for automatic music generation

In Figure 5, the music generation framework consists of three parts: preprocessing of MIDI music data, model training and music generation. The melody and chords of MIDI music are correlated. The melody is the dominant part of the music and the chords are the collection of notes that support and enrich the melody in the context of the music [27]. The study constructs a new storage structure. The melody and chord tracks are extracted separately and then sequentially spliced to ensure that the music generation model can learn the corresponding information units of both at the same time. In the note extraction and storage process, the study defines a "chord library dictionary", which helps the model to follow the harmonic rules of the music during the generation process and reduces the occurrence of dissonance. The joint modeling of melody and chord extracts the two sequences from the MIDI data and then digitally maps them separately. This allows the model to learn the intrinsic connection between the two sequences simultaneously, thereby facilitating the generation of more harmonious and expressive musical compositions. This allows the model to simultaneously learn the intrinsic connection between the two sequences, helping to generate more harmonic and expressive musical compositions. The model training process is shown in Figure 6. In Figure 6, the melodic sequences and harmonic sequences obtained from the preprocessing of MIDI music data are combined and mapped according to the actual sequences. Moreover, the harmonic and notes are considered as elements for loop training. The chord substitution network learns the similarities and transformation patterns between chords during training to generate more musically logical melodies. The group training approach allows the model to learn both local and global musical structure features. In the similar chord substitution model, a set of tones (octaves) is first divided into twelve chromatic intervals to complete the chord correspondence according to the numerical ladder. Then the basic chord-C chord is replaced using the genitive chord. Finally, the Chord2vec model is used to learn the vector representation of the chords, and similar chords are obtained by model training. Chord2vec model is developed on the basis of Word2Vec model, which is a mathematical representation of musical structure, harmonic properties and chord relationships by mapping chords in music into vector space. The process of chord vectorization for the Chord2vec model treats chords as a series of note combinations, and the vector position expression matrix PE is shown in Equation (13).

$$\begin{cases} PE_{(pos,2i)} = \sin\left|\dfrac{pos}{10000^{\frac{2i}{d}}}\right| \\ PE_{(pos,2i+1)} = \cos\left|\dfrac{pos}{10000^{\frac{2i}{d}}}\right| \end{cases} \tag{13}$$

In Equation (13), d denotes the note vector dimension. pos is the position of the note vector in the music bar. $2i$ and $2i+1$ refer to even and odd numbers, respectively. Harmonically similar chords will be mapped to similar positions. The similarity between chords can then be measured using distances in vector space. Additionally, the study determines the replacement chords based on the knowledge of music theory, and the chord replacement is accomplished by combining the calculated similar chords. In the music generation and optimization section, MIDI music clips are randomly input and trained with a sequence learning model to obtain the predicted music clips. Then the chord substitution model is trained and the music clip is deposited again to get the new music.

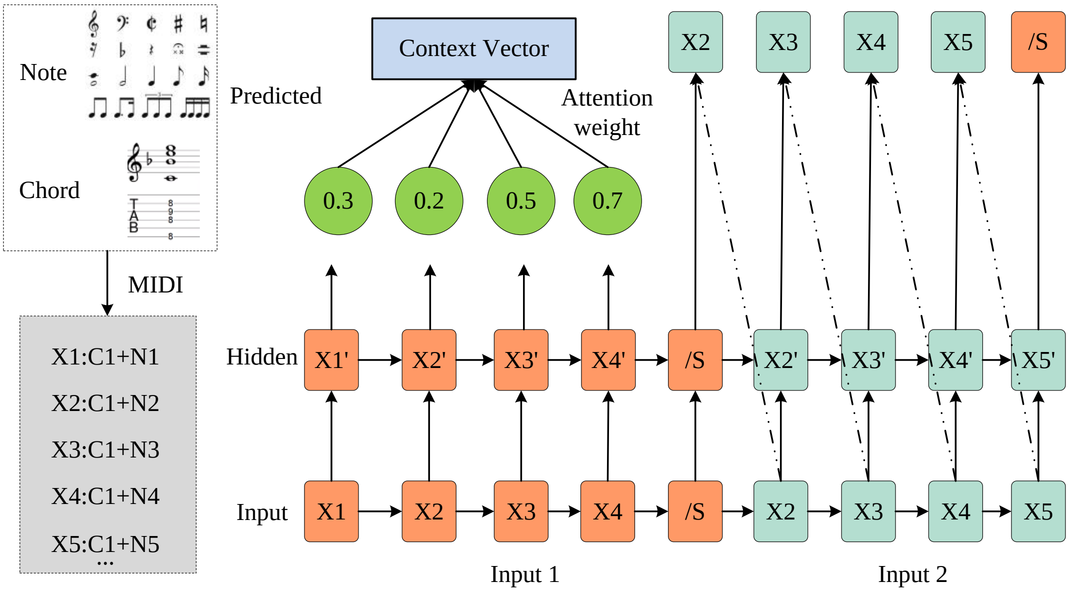


Figure 6: Schematic diagram of the network structure for music sequence generation in the transformer model

4 Performance analysis of MSA and automatic music content generation models

To validate the effectiveness of the MSA-based model for automatic music content generation, the study centers on MSA and automatic music generation, and the results are analyzed and discussed.

4.1 Performance testing of multimodal sentiment analysis models

The hardware environment operating system used for the experiment is Windows 10, the processor is Intel(R) Xeon(R) CPUE5-2603 v3, the GPU is Nvidia GTX 1080Ti, and the memory is 64G. The DL framework in the software environment is based on the implementation of Nvidia GTX 1080Ti, and the programming language is Python 3.7.1. The experiment is conducted using the research-constructed MSA dataset, the interactive emotional dyadic motion capture (IEMOCAP), carnegie mellon university multimodal opinion sentiment and emotion intensity (CMU__MOSI) and its extended version, the CMU-MOSEI dataset. A 12-hour period of audiovisual data, including speech, video, text transcription, and facial motion capture, can be found in the action, multimodal, and multicamera database known as the IEMOCAP dataset. Many annotators have annotated it with dimensional labels like potency, activation, and dominance as well as categorical categories like anger, happiness, sadness, and neutrality. CMU__MOSI contains movie clips, audio, text, and emotion tags data from YouTube covering a wide range of emotional expressions and subjective comments. CMU-MOSEI includes video, audio, and text information covering more than 16,000 subjective comments and sentiment labels. Depending on the requirements of the experiment, the dataset is split into an 8:2 ratio training set and test set.

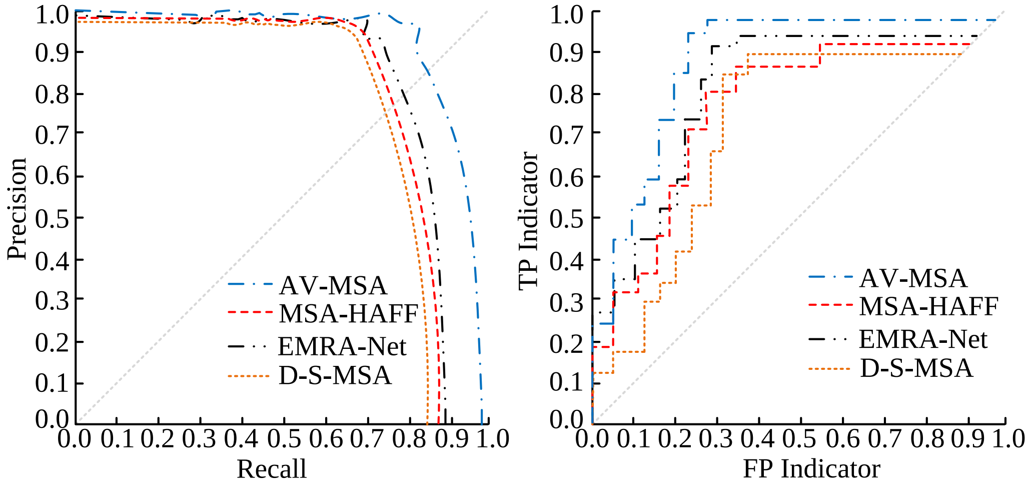
This experiment selects the optimized ensemble multi-scale residual attention network (EMRA-Net) [28], the MSA method based on hierarchical adaptive feature fusion network (MSA-HAFF) [29], and the MSA analysis method based on Dempster-Shafter (D-S) theory [30]. Mean absolute percentage error (MAPE), root mean square error (RMSE), mean absolute error (MAE), and R-squared metrics of different MSA methods are compared. Table 3 displays the outcomes of the experiment.

Table 3: Comparison of analytical performance of different MSA methods

Model	Index	MSA-study	IEMOCAP	CMU__MOSI	CMU-MOSEI
EMRA-Net	MAE	0.398	0.449	0.438	0.429
	RMSE	0.404	0.447	0.454	0.431
	R-squared	0.743	0.716	0.715	0.736
	MAPE	0.434	0.439	0.466	0.450
MSA-HAFF	MAE	0.449	0.461	0.475	0.458

	RMSE	0.452	0.509	0.513	0.502
	R-squared	0.704	0.669	0.668	0.673
	MAPE	0.388	0.494	0.465	0.420
D-S-MSA	MAE	0.391	0.562	0.423	0.407
	RMSE	0.388	0.429	0.398	0.366
	R-squared	0.781	0.735	0.722	0.744
	MAPE	0.343	0.439	0.417	0.400
AV-MSA	MAE	0.149	0.219	0.216	0.183
	RMSE	0.166	0.310	0.270	0.212
	R-squared	0.961	0.851	0.873	0.902
	MAPE	0.140	0.296	0.267	0.228

In Table 3, the research-designed AV-MSA framework takes the smallest values on three different error metrics. Among them, it performs best on the research-designed MSA-study dataset. the MAE, RMSE, and MAPE metrics are 0.149, 0.166, and 0.140, respectively. Secondly, the CMU-MOSE dataset takes better values of error than the CMU_MOSI dataset. Meanwhile, the pattern of taking values between different datasets is reflected in EMRA-Net, MSA-HAFF, and D-S-MSA models. In summary, the MSA-study dataset can better train and optimize the MSA model, which helps the model learn the correlation between different modalities and the diversity of emotional cues in a deeper way, and reduces the error of sentiment analysis. Additionally, EMRA-Net, MSA-HAFF, and D-S-MSA are the error takeoffs of the three MSA models, which all fluctuate in the 0.3-0.5 range. The surface model is slightly less predictive of sentiment. The R-squared take of the AV-MSA framework is up to 0.961, which is significantly better than the other models in explaining emotion. Figure 7 displays the receiver operating characteristic curve (ROC) and precision-recall (P-R) curve results for various models.



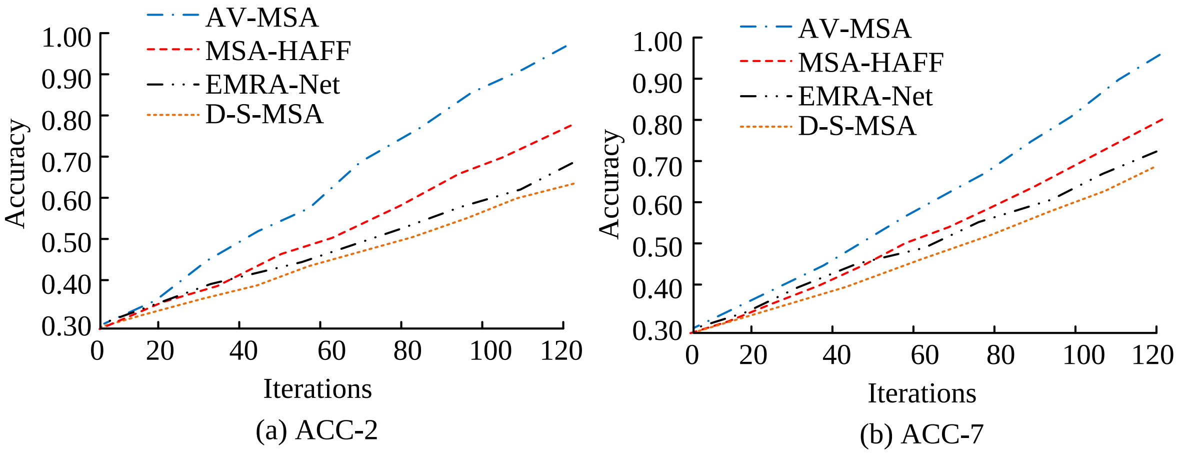
(a) PR

(b) ROC

Figure 7: Comparison of P-R curves and ROC curves for different MSA methods

In Figure 7(a), the PR curve of the AV-MSA framework designed by the study performs the best and is located at the top right-hand side of the axes at a precision of 0.9. The recall of the AV-MSA framework is 0.98. The recalls of EMRA-Net, MSA-HAFF, and D-S-MSA are 0.88, 0.86, and 0.82, respectively. The AV-MSA framework in the same experimental environment has a relative advantage in the precision rate and recall reconciliation average, and the accuracy of the model is superior. In Figure 7(b), the ROC curve of the research-designed AV-MSA framework has the best performance. Area under the curve (AUC) takes the values of 0.94, 0.90, 0.84, and 0.80, respectively. The larger the AUC value the better the comprehensive performance of the model, which displays that the AV-MSA framework has obvious superiority. The results of the 7-category accuracy and 2-category accuracy analysis of the model are shown in Figure 8.

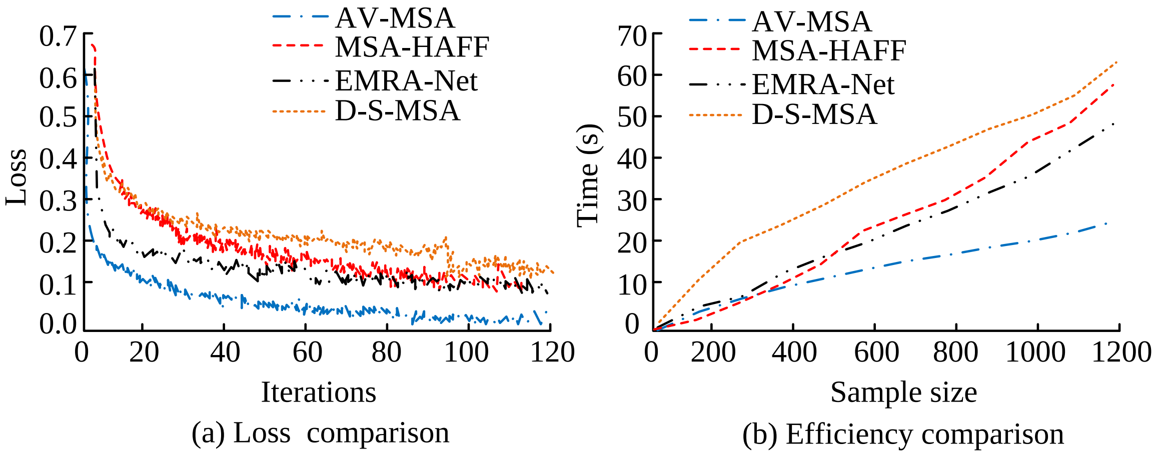
In Figure 8, the AV-MSA framework designed by the study has significant superiority in both 7-category accuracy and 2-category accuracy. The 2-category accuracy reaches 0.98 in Figure 8(a). The 7-category accuracy reaches 0.96 in Figure 8(b). In contrast, the 2-category accuracy of EMRA-Net, MSA-HAFF, and D-S-MSA are 0.69, 0.78, and 0.62, respectively. The 7-category accuracy are 0.72, 0.80, and 0.69, respectively. The loss functions of the different models curves are compared with the analytical efficiency as shown in Figure 9.



(a) ACC-2

(b) ACC-7

Figure 8: Comparison between 7-category accuracy and 2-category accuracy



(a) Loss comparison

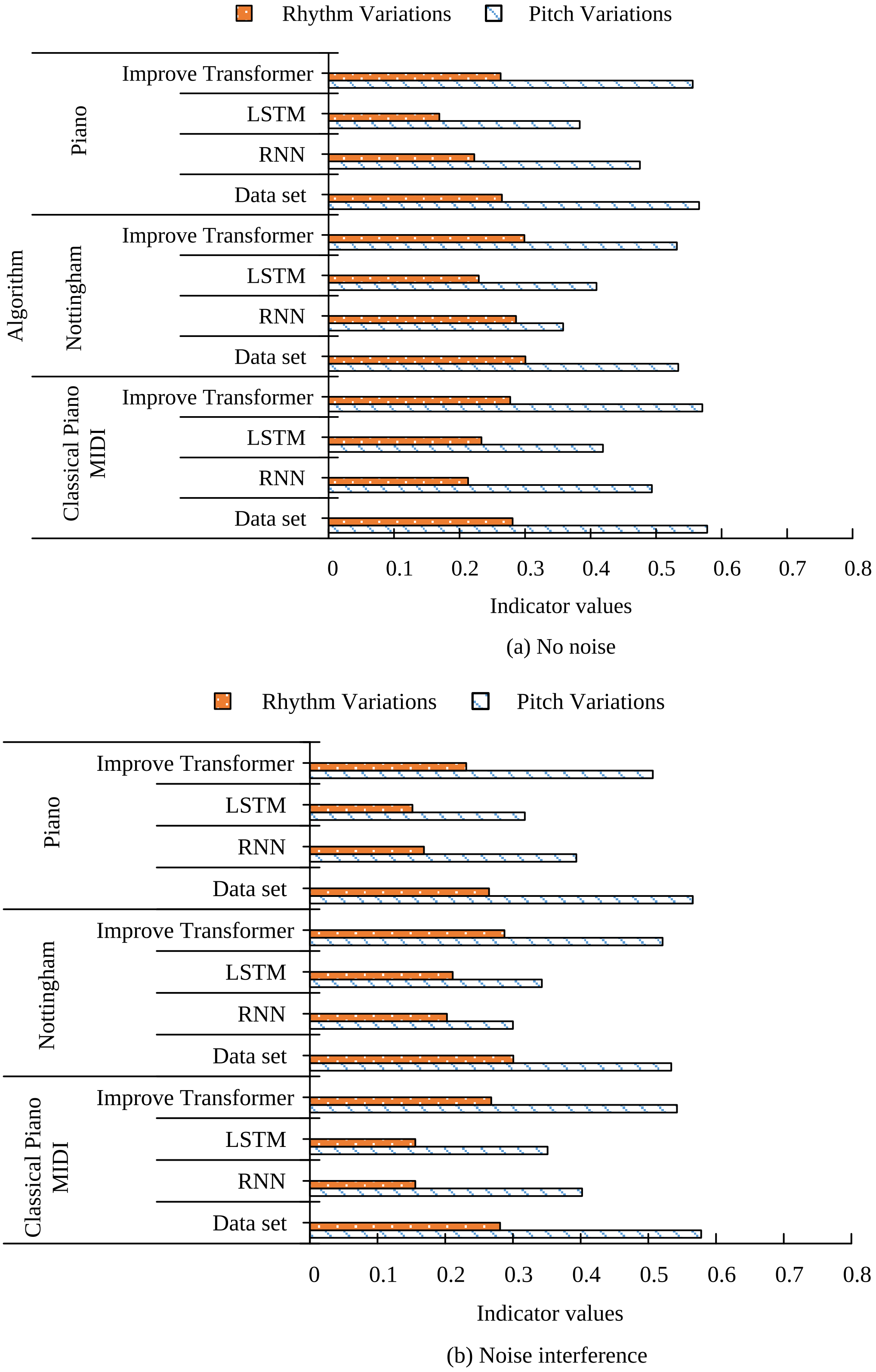
(b) Efficiency comparison

Figure 9: Comparison of loss function and analysis efficiency

In Figure 9(a), the loss function curve of the AV-MSA framework designed for the study converges the fastest, with the smallest convergence value of 0.03. The loss function curves of EMRA-Net, MSA-HAFF, and D-S-MSA, converge slower, with the convergence value roughly around 0.20. In Figure 9(b), the efficiency of the AV-MSA framework for analyzing the multimodal sentiment is remarkable. When the number of analyzed samples is 1200, the analysis time is 24.2s. Compared with other models, its maximum time reduction value can reach 39.4s. Comprehensively analyzing, the MSA analysis framework designed by the research achieves better results in terms of accuracy, efficiency and other indicators, which provides a solid technological foundation for music content generation.

4.2 Performance testing of automatic music content generation models

The classical piano music dataset Classical Piano MIDI, the folk song ballad dataset Nottingham, and the MIDI piano music dataset Piano in classical and jazz styles are selected for experimental analysis. According to the experimental needs, it is divided into training set and test set according to the ratio of 7:3. Commonly used baseline sequence analysis models are selected for comparison, including LSTM, recursive neural network (RNN). The pitch variations (PV) and rhythm variations (RV) of the generated music for different models are compared as shown in Figure 10.



(a) No noise

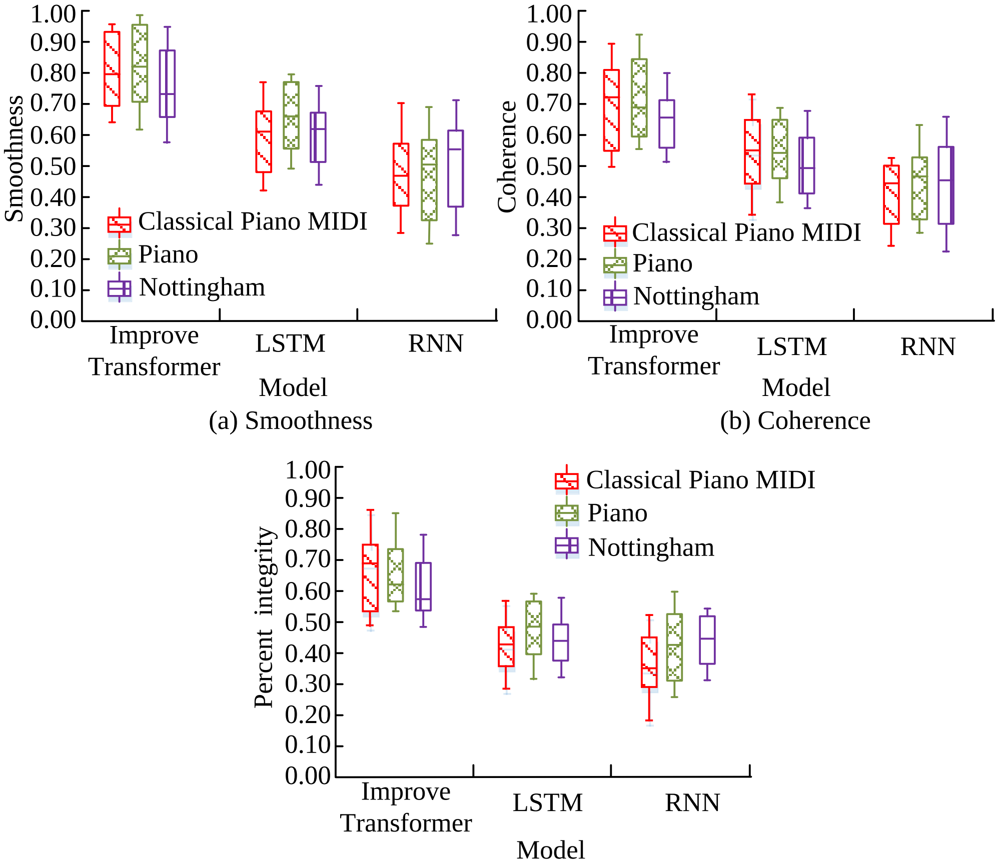
(b) Noise interference

Figure 10: Comparison of pitch and melody changes in different models

PV and RV are related metrics for note analysis. PV is the variation in the ebb and flow of notes across the range, i.e., the variation in note elevations and intervals, obtained by comparing the pitches to the total notes. RV is the change in rhythm, i.e., how many different note durations occur. In Figure 10 (a), the PV and RV values of the improved Transformer structure designed by the study are closest to the metrics taken from the dataset. The PV values are taken with a difference of 1.29%, 8.01%, and 1.78% compared to the Classical Piano MIDI, Nottingham, and Piano datasets, respectively. The RV values are taken with a magnitude of difference of 1.29%, 0.66%, and 0.86% when compared to the Classical Piano MIDI, Nottingham, and Piano datasets, respectively. The magnitude of difference between the PV and RV values of the other models can be up to 32.92% and 36.14%. Improved Transformer structure is better for feature learning on music dataset. Additionally, to test the performance of the models under noise conditions, the study artificially added noise segments in different styles of music datasets, and the results are shown in Figure 10(b). It can be observed that background noise impedes the parsing of the audio signal by RNN and LSTM models, thereby reducing the quality of the generated music and increasing the discrepancy between the PV and RV values. However, the improved Transformer structure still maintains more consistent PV and RV values, and the robustness of the model is superior.

Comparison of smoothness, coherence, and percentage of fragmented bars of music generated by different models is shown in Figure 11.

In Figure 11(a), the music generated by the improved Transformer structure has a better smoothness, with median fetch levels above 0.70. The smoothness of the music generated by the other two methods is below the 0.60 value level. Smoothness can be used to describe the smoothness and harmony of the music. The higher the smoothness level, the better the handling of connections and transitions between notes. In Figure 11(b), the music generated by the improved Transformer structure has a better coherence and is at least 0.2 fetch level higher compared to the other models. The model handles the degree of articulation and completeness of the music better. The percentage of fragmented bars can refer to the presence or absence of fragmented bars in the music. The experiment uses the integrity percentage to refer to the fragmented bar percentage. In Figure 11(c), the integrity percentage of the improved Transformer structure is significantly higher than the other two models. The research-designed AV-MSA framework with the improved Transformer structure is applied to the multimodal sentiment intelligent music generation analysis. The normalized discounted cumulative gain (NDCG) and hit rate (HR) are chosen as the evaluation indexes of the model's music generation merit. Table 4 displays the findings of the experiment.



(a) Smoothness

(b) Coherence

(c) Percent integrity

Figure 11: Comparison of music quality generated by different models

Table 4: Comparison of NDCG and HR indicators for different models

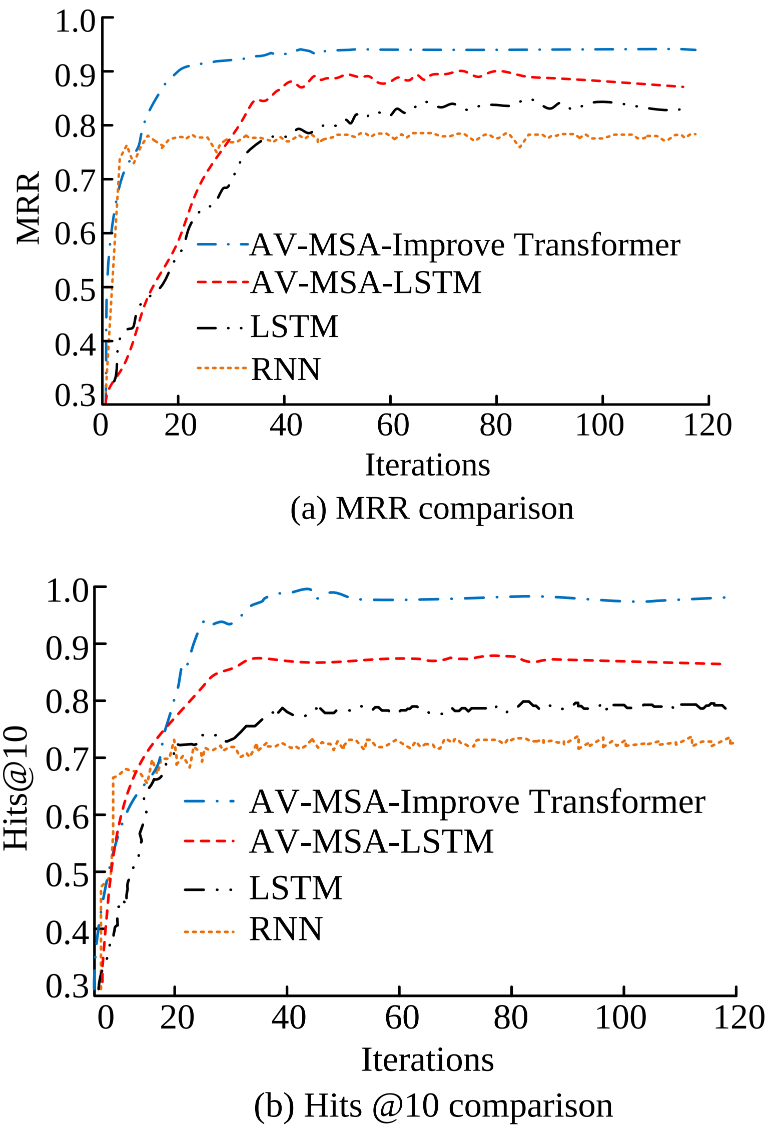
Evaluating indicator	AV-MSA-Improve Transformer	AV-MSA-LSTM	LSTM	RNN
HR#3	44.659%	37.218%	32.734%	30.038%
HR#5	56.888%	57.231%	45.675%	48.490%
HR#5	64.768%	63.547%	59.546%	53.384%
HR#10	87.680%	70.655%	66.449%	68.804%
HR#15	93.984%	83.900%	79.848%	79.490%
Evaluating indicator	AV-MSA-Improve Transformer	AV-MSA-LSTM	LSTM	RNN
NDCG #3	46.314%	37.815%	29.228%	30.806%
NDCG #5	67.420%	42.030%	39.531%	38.345%
NDCG #8	77.749%	56.857%	51.647%	42.556%
NDCG #10	83.335%	70.856%	59.129%	50.095%
NDCG #15	91.566%	83.304%	63.322%	62.644%

In Table 4, with the increase of k-rank, the values of the indicators taken by different models showed an upward trend. The larger the value of k, the larger the NDCG and HR take values. The maximum HR value of AV-MSA-Improve Transformer model is 93.984% and the maximum NDCG value is 91.566%, which is significantly higher than the other three methods. Additionally, comparing the AV-MSA-LSTM model with the LSTM model, the HR and NDGG of AV-MSA-LSTM under different k-ranks are improved to some extent. It can be concluded that music generation based on MSA is more likely to get the music that meets the demand and hits the user's needs more easily. The experimental results of mean reciprocal rank (MRR) with Hits@10 are shown in Figure 12. In Figure 12(a), the AV-MSA-Improve Transformer model has the highest MRR curve of 0.91. The AV-MSA-LSTM model reaches 0.89, which is higher than the other two music generation models. In Figure 12(b), the results of the different models for the values of Hits@10 show the same pattern. The AV-MSA-LSTM model has the best performance with Hits@10 taking values up to 0.97. Additionally, to verify the effects of feature fusion, LSTM-MLP feature extraction module and MIDI music data preprocessing and model training on the music generation effect in the AV-MSA framework, the model ablation experiments are first initiated. The results of the ablation experiments are shown in Table 5. In Table 5, the AV-MSA-Improve Transformer model takes the highest level of values on the three metrics of HR#10, NDCG #10, and Hits@10, and the model has the best performance. The traditional Transformer music generation architecture takes values around 0.75 for all metrics. After introducing the feature fusion and LSTM-MLP feature extraction modules to build the AV-MSA framework, the fetch level increases to around 0.80. The feature fusion module helps to alleviate the noise or missing problem that may exist in single-modal data, and promotes inter-modal complementarity. LSTM-MLP can effectively capture the emotion change trend and contextual information of each modality, and improves the accuracy of emotion recognition. Then the AV-MSA framework improves the emotional and melodic characteristics of the generated music, and the performance of HR, NDCG & Hits indicators is improved. Additionally, the music generated by the Transformer architecture improved by the MIDI data pre-processing and model training module is further improved, and the improvement strategy is set accordingly.

Finally, the qualitative assessment of the model-generated music is carried out by combining the feedback and subjective judgment of the users. Moreover, a total of 30 people is included in the experimental assessment to evaluate the quality of 100 pieces of music. The experimental results are shown in Figure 13.

Table 5: Results of ablation experiments of the model

Transformer	Feature fusion	LSTM-MLP	Data preprocessing	Model training	HR#10	NDCG #10	Hits@10
√	/	/	/	/	0.749	0.731	0.753
√	√	√	/	/	0.805	0.807	0.804
√	√	√	√	/	0.824	0.812	0.820
√	√	√	√	√	0.877	0.833	0.970



(a) MRR comparison

(b) Hits @10 comparison

Figure 12: MRR and Hits@10 value comparison

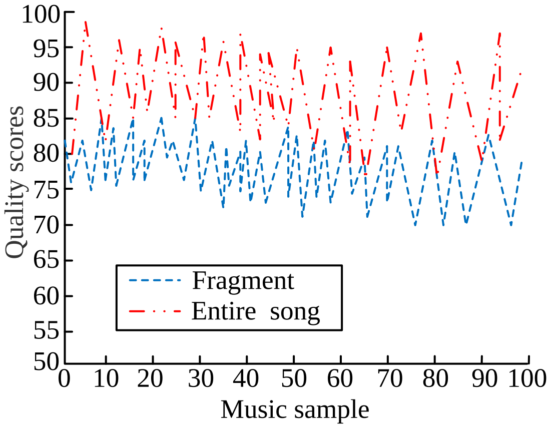


Figure 13: Results of qualitative assessment of music generation quality

In Fig. 13, among the 100 pieces of music tested, the average quality score of music segments is roughly in the range of 75-85. The overall quality score is in the range of 80-100. The results of the qualitative assessment can confirm the excellent performance of the music generation model designed by the study.

5 Discussion

In recent years, the rapid development of the field of "artificial intelligence + music" has opened up new possibilities for music creation, which can not only speed up the process of music creation, but also explore music styles and forms that are unimaginable to humans. However, by analyzing the literature [16], [17], [18], it can be concluded that the existing automatic music generation technology still faces challenges such as single content and difficulty in generating music content with deep emotional style. Therefore, the research combined MSA with automatic music generation techniques to allow music generation frameworks to understand and express more complex and delicate emotions. However, the robust performance of MSA models under noisy data and unlabelled datasets still had more shortcomings. The inter-modal fusion problem and the complexity of feature extraction were still challenges for MSA models. As in literature [11] Chen S et al. designed a multi-stage multimodal dynamic fusion network, and literature [12] Kumar P et al. designed a multimodal emotion recognition system with hybrid fusion. The MSA method was obviously insufficient to focus on the original features. Even though the accuracy rate of emotion recognition reached 83.29%, it only fused the image and speech features, and the multimodal information was not sufficiently considered. Therefore, to address the technical dilemma faced by MSA and automatic music generation technology, the study drew on the design concepts of Yi Y et al. in literature [13]. Firstly, it solved the technical difficulties posed by unlabelled datasets, and constructed a high-quality, accurately annotated multi-label sentiment analysis dataset, which could help the music generation model to correlate multiple sentiment labels. This study then utilized DL models, sequence analysis models, and feature extraction to capture the intrinsic connections between different modalities. The study extracted key information that characterized emotional and melodic features and constructed the MSA framework. Finally, the study introduced the MIDI music data preprocessing technique and model training module to improve the Transformer model to better generate music content with emotional and melodic features.

The results of the study showed that the music generation model incorporating the AV-MSA framework could take the highest HR and NDGG values up to 93.984% and 91.566% on Classical Piano MIDI, Nottingham and Piano datasets. Hits@10 could take values up to 0.97. Compared to the pre-trained model based on literature [18] for multitask music generation, the HITS@k metrics improved significantly. The efficient feature extraction and sequence analysis capabilities improved the performance of the MSA framework. The incorporation of the AV-MSA framework into the automated music generation methodology, in conjunction with the cumulative impact of the preprocessing technique and the model training module, enhances the capabilities of the enhanced Transformer model in generating music with emotional and melodic attributes.

However, the dataset used for the experimental analysis is biased toward popular and classical music, ignoring other niche or regional music styles. The music style bias may cause the model to perform poorly in generating or analyzing non-mainstream music. Furthermore, there are notable discrepancies in the cultural milieu of disparate regions, which give rise to markedly disparate musical expressions and emotional connotations. A single cultural context in the dataset may limit the application of the model in cross-cultural contexts. In future research, it would be beneficial to update the model training dataset in a timely manner to incorporate the latest musical works and popular trends, thus enhancing the model's generalizability.

6 Conclusion

The objective of this study is threefold: Firstly, to gain a comprehensive understanding of how humans perceive emotions and situations. Secondly, to enhance the capacity of intelligent systems to comprehend emotions and situations. Thirdly, to facilitate the advancement of intelligent music content generation and the field of affective computing. To this end, a model is constructed and an analysis conducted of MSA and music intelligent generation. Based on these findings, a novel AV-MSA framework with an automatic music generation model was proposed. The experimental results indicated that the AV-MSA framework took the smallest values on three different error metrics. MAE, RMSE, and MAPE metrics were 0.149, 0.166, and 0.140, respectively. R-squared took the maximum value of 0.961. The performance of the model constructed by the study achieved good performance with the dataset. The recall was 0.98 for AV-MSA framework precision of 0.9 and AUC took the value of 0.94. The 2-category accuracy and 7-category accuracy were 0.98 and 0.96, respectively. The PV and RV values of the improved Transformer structure are closest to the metrics taken from the dataset. The minimum difference was 1.29% and 0.66% respectively. The smoothness, coherence, and completeness of the music performed better. Jointly utilized with the AV-MSA framework, the maximum NDGG value of the generated music was 91.566%, the maximum HR value was 93.984%, the MRR value was 0.89, and the Hits@10 value could be taken up to 0.97. The objective of the study was to generate personalized music content based on multimodal situational sentiment perception. This was done with the aim of matching specific situations and emotional states, thereby enhancing the user's music experience. However, the MSA dataset, as designed in the study, still requires a greater investment of time and human resources. Additionally, the integration of the music generation model is inadequate, the model training is divided into two distinct modules, and the potential for further performance enhancement remains to be investigated.

References

[1] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A survey of deep learning-based multimodal sentiment recognition: Speech, text, and face," Entropy, vol. 25, no. 10, pp. 1440-1462, 2023. https://doi.org/10.1007/s10518-023-01779-8

[2] S. Holiday, J. L. Hayes, H. Park, Y. Lyu, and Y. Zhou, "A multimodal emotion perspective on social media influencer marketing: The effectiveness of influencer emotions, network size, and branding on consumer brand engagement using facial expression and linguistic analysis," Journal of Interactive Marketing, vol. 58, no. 4, pp. 414-439, 2023. https://doi.org/10.1177/109499682311711

[3] B. Pan, K. Hirota, Z. Jia, L. Zhao, X. Jin, and Y. Dai, "Multimodal sentiment recognition based on feature selection and extreme learning machine in video clips," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 3, pp. 1903-1917, 2023. https://doi.org/10.1007/s12652-021-03407-2

[4] J. M. Garcia-Garcia, M. D. Lozano, V. M. Penichet, and E. L. C. Law, "Building a three-level multimodal sentiment recognition framework," Multimedia Tools and Applications, vol. 82, no. 1, pp. 239-269, 2023. https://doi.org/10.1007/s11042-022-13254-8

[5] S. Dong, X. Fan, and X. Ma, "Multichannel multimodal emotion analysis of cross-modal feedback interactions based on knowledge graph," Neural Processing Letters, vol. 56, no. 3, pp. 1-17, 2023. https://doi.org/10.1007/s11063-024-11641-w

[6] J. Liu, Z. Wang, W. Nie, J. Zeng, B. Zhou, J. Deng, H. Li, Q. Xu, X. Xu, and H. Liu, "Multimodal sentiment recognition for children with autism spectrum disorder in social interaction," International Journal of Human-Computer Interaction, vol. 40, no. 8, pp. 1921-1930, 2024. 10.1080/10447318.2023.2232194

[7] Z. Yin, F. Reuben, S. Stepney, and T. Collins, "Deep learning’s shallow gains: A comparative evaluation of algorithms for automatic music generation," Machine Learning, vol. 112, no. 5, pp. 1785-1822, 2023. https://doi.org/10.1007/s10994-023-06309-w

[8] P. Georges, and A. Seckin, "Music information visualization and classical composers’ discovery: An application of network graphs, multidimensional scaling, and support vector machines," Scientometrics, vol. 127, no. 5, pp. 2277-2311, 2022. https://doi.org/10.1007/s11192-022-04331-8

[9] D. Jiang, H. Liu, R. Wei, and G. Tu, "CSAT-FTCN: A fuzzy-oriented model with contextual self-attention network for multimodal sentiment recognition," Cognitive Computation, vol. 15, no. 3, pp. 1082-1091, 2023. https://doi.org/10.1007/s12559-023-10119-6

[10] Z. Fu, F. Liu, Q. Xu, X. Fu, and J. Qi, "LMR-CBT: Learning modality-fused representations with CB-transformer for multimodal sentiment recognition from unaligned multimodal sequences," Frontiers of Computer Science, vol. 18, no. 4, pp. 184314-184332, 2023. https://doi.org/10.1007/s11704-023-2444-y

[11] S. Chen, J. Tang, L. Zhu, and W. Kong, "A multi-stage dynamical fusion network for

multimodal sentiment recognition," Cognitive Neurodynamics, vol. 17, no. 3, pp. 671-680, 2023. https://doi.org/10.1007/s11571-022-09851-w

[12] P. Kumar, S. Malik, and B. Raman, "Interpretable multimodal sentiment recognition using hybrid fusion of speech and image data," Multimedia Tools and Applications, vol. 83, no. 10, pp. 28373-28394, 2023. https://doi.org/10.1007/s11042-023-16443-1

[13] Y. Yi, Y. Tian, C. He, Y. Fan, X. Hu, and Y. Xu, "DBT: Multimodal sentiment recognition based on dual-branch transformer," The Journal of Supercomputing, vol. 79, no. 8, pp. 8611-8633, 2023. https://doi.org/10.1007/s11042-023-16443-1

[14] G. Tang, Y. Xie, K. Li, R. Liang, and L. Zhao, "Multimodal sentiment recognition from facial expression and speech based on feature fusion," Multimedia Tools and Applications, vol. 82, no. 11, pp. 16359-16373, 2023. https://doi.org/10.1007/s11042-022-14185-0

[15] C. Dixit, and S. M. Satapathy, "A customizable framework for multimodal sentiment recognition using ensemble of deep neural network models," Multimedia Systems, vol. 29, no. 6, pp. 3151-3168, 2023. https://doi.org/10.1007/s00530-023-01188-6

[16] S. Shukla, and H. Banka, "Monophonic music composition using genetic algorithm and Bresenham's line algorithm," Multimedia Tools and Applications, vol. 81, no. 18, pp. 26483-26503, 2022. https://doi.org/10.1007/s11042-022-12185-8

[17] G. Wu, S. Liu, and X. Fan, "The power of fragmentation: a hierarchical transformer model for structural segmentation in symbolic music generation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, no. 5, pp. 1409-1420, 2022. https://doi.org/arXiv:2205.08579v2

[18] S. Li, and Y. Sung, "MRBERT: Pre-Training of melody and rhythm for automatic music generation," Mathematics, vol. 11, no. 4, pp. 798-811, 2023. https://doi.org/10.3390/math11040798

[19] S. Choudhuri, S. Adeniye, and A. Sen, "Distribution alignment using complement entropy objective and adaptive consensus-based label refinement for partial domain adaptation," Artificial Intelligence and Applications vol. 1, no. 1, pp. 43-51, 2023. https://doi.org/10.47852/bonviewAIA2202524

[20] V. Beltrán-Palanques, and M. Querol-Julián, "The genre of PechaKucha presentations: Analysis and implications for enhancing multimodal literacy at university," English for Specific Purposes, vol. 75, no. 1, pp. 27-37, 2024. https://doi.org/10.1016/j.esp.2024.05.002

[21] S. S. Hosseini, M. R. Yamaghani, and S. Poorzaker Arabani, "Multimodal modelling of human emotion using sound, image and text fusion," Signal, vol. 1, no. 81, pp. 71-79, 2024. https://doi.org/10.1007/s11760-023-02707-8

[22] J. Wang, A. Sharifi, T. R. Gadekallu, and A. Shankar, "MMD-MII model: A multilayered analysis and multimodal integration interaction approach revolutionizing music emotion classification," International Journal of Computational Intelligence Systems, vol. 17, no. 1, pp. 99-124, 2024. https://doi.org/10.1007/s44196-024-00489-6

[23] I. C. Lu, J. Y. Huang, and W. P. Lee, "An emotion-driven and topic-aware dialogue framework for human–robot interaction," Advanced Robotics, vol. 38, no. 4, pp. 267-281, 2024. https://doi.org/10.1080/01691864.2023.2297902

[24] L. Jing, "Evolutionary deep learning for sequential data processing in music education," Informatica, vol. 48, no. 8, pp. 63-78, 2024. https://doi.org/10.31449/inf.v48i8.5444

[25] H. F. T. Al-Saadawi, and R. Das, "TER-CA-WGNN: Trimodel sentiment recognition using cumulative attribute-weighted graph neural network," Applied Sciences, vol. 14, no. 6, pp. 2252-2281, 2023. https://doi.org/10.3390/app14062252

[26] G. Wang, L. Tan, and Z. Shang, "Multimodal dual emotion with fusion of visual sentiment for rumor detection," Multimedia Tools and Applications, vol. 83, no. 10, pp. 29805-29826, 2024. https://doi.org/10.1007/s11042-023-16732-9

[27] R. Kumari, V. Gupta, N. Ashok, T. Ghosal, and A. Ekbal, "Emotion aided multi-task framework for video embedded misinformation detection," Multimedia Tools and Applications, vol. 83, no. 12, pp. 37161-37185, 2024. https://doi.org/10.1007/s11042-023-17208-6

[28] B. Subbaiah, K. Murugesan, P. Saravanan, and K. Marudhamuthu, "An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network," Artificial Intelligence Review, vol. 57, no. 2, pp. 34-56, 2024. https://doi.org/10.1007/s10462-023-10645-7

[29] H. Zhao, "Research on the recognition of psychological emotions in adults using multimodal fusion," Informatica, vol. 48, no. 9, pp. 155-162, 2024. https://doi.org/10.31449/inf.v48i9.5876

[30] M. Zhang, C. Wang, Y. Sun, and T. Li, "Memristive PAD three-dimensional emotion generation system based on D–S evidence theory," Nonlinear Dynamics, vol. 112, no. 6, pp. 4841-4861, 2024. https://doi.org/10.1007/s11071-023-09264-2