Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Theory and practice of AI-based interactivity enhancement strategies in digital media

Guohao Zou[1, ✉]

[1] *School of Humanities and Arts, Nanchang Institute of Technology, Nanchang, Jiangxi, 330099, China*

ABSTRACT

AI technology can accurately capture and feedback user emotions in digital media interaction to realize precise interaction. In this paper, we design an AI emotion interactivity enhancement model based on multimodal fusion, and apply the neural network model of Bi-GRU and dual attention mechanism to fuse the long and short-term emotion classification results of the tested samples at the decision level to obtain the final emotion classification results. Then the weight coefficient vector of each sentiment category is calculated based on the sentiment classification confusion matrix of the classifier, which is used as the a priori knowledge for multimodal sentiment analysis for decision fusion. The performance is examined on the MOSI dataset and the AI-based interaction design strategy in digital media is proposed. Analyzing the interaction design effect, the interaction design applying the model of this paper has better user experience sense, emotional arousal, pleasure level, and emotional feedback effect in subjectivity evaluation than the control group, and 75% of the experimental subjects think that the feedback-adjusted digital media has a better pleasure level.

Keywords: digital media, interaction design, multimodal sentiment analysis, Bi-GRU, dual attention mechanism

1. Introduction

The current Internet and Artificial Intelligence (AI) technology has also been developed significantly, which provides great convenience for human production and enhances the industrial value chain [1-2]. With the continuous development of the Internet, digital media has entered into people's view. In the big data environment, production and work have also entered the digital operation stage. Enterprises using digital technology, can effectively enhance the efficiency of enterprise management of all kinds of information, the use of digital technology in personnel management, promoting the personnel management work to achieve new breakthroughs [3]. With the gradual integration of digital technology into the use of enterprises, some schools will also be introduced, through the application

✉ Corresponding author.
E-mail address: 19179840890@163.com (G. Zou).
Received 05 March 2024; Revised 20 April 2024; Accepted 05 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-327

of these new digital media technology, the school's teaching quality will be greatly improved, but also to enhance the students' ability to learn independently and learning enthusiasm [4-5].

Digital media is a combination of traditional media and information technology, which has brought about a profound revolution in communication media, information carriers, communication methods, audience participation and other aspects [6]. It not only expands the boundaries of traditional media, but also creates new communication methods and content forms. The most important feature of digital media is the efficient use of a large amount of digital data in the chaos of information, and these data are divided and digitized display, the realization of social and other forms of digitization, and can be achieved through the storage network to achieve the efficient popularization of these data, the encryption and storage of the network can also ensure the security of the information [7-8]. In addition, digital media are characterized by reproducibility and interactivity. Unlike traditional media, digital media information can be reproduced and transmitted through digital coding, and there is no loss of quality in the process of reproduction, which makes digital media information can be rapidly disseminated and diffused, greatly improving the efficiency and speed of information dissemination, and at the same time, it also brings about the problem of content piracy and copyright protection, and the digital media industry needs to set up an effective mechanism of copyright protection and revenue mode [9-10]. Traditional media is mainly one-way communication, while digital media realizes two-way communication between users and media through the interaction of network and user interface. Users can participate in the creation and dissemination process of digital media through interactive ways, such as posting comments, sharing content, interactive display, etc., so as to personalize and customize and proliferate the media content [11-13]. Interactivity makes digital media no longer a passive tool for receiving information, but a diversified all-round media platform. Due to the existence of interactivity, digital media can provide a richer experience than traditional media, and digital media has accelerated the development and become an important direction in the fields of culture, business, health care, and education [14-17].

This paper proposes an interactivity enhancement model for AI based on multimodal fusion. Firstly, a multimodal sentiment analysis method based on long and short-time features and decision-level fusion is proposed, and a neural network model based on Bi-GRU and dual attention mechanism is used to classify the measured multimodal information with short-time sentiment, then a neural network model with long-time features and LMF is used to classify the measured video with long-time sentiment, and finally, the results of the long and short-time sentiment classification are fused with decision-level fusion. Experiments are conducted on the dataset to verify the model classification performance. The AI-based interaction design strategy in digital media is proposed, and the interaction design effect is analyzed in terms of subjectivity and objectivity.

2. Interactivity enhancement model for AI based on multimodal fusion

2.1. Application of Artificial Intelligence Technology to User Interaction and Experience

2.1.1. The role and application of AI in user interaction. AI is one of the important applications of artificial intelligence technology in user interaction and experience. They simulate human-to-human conversations and exchanges through technologies such as natural language processing and machine learning to provide users with intelligent questions and answers, advice and guidance [18].

A chatbot is a computer program that can interact verbally with the user. It can answer questions posed by the user, provide relevant information and advice, and mimic the way and style of human responses. Chatbots can be used in several fields, such as customer service, education, and healthcare, to provide users with 24-hour online Q&A services.

Virtual assistant is a kind of chatbot with voice interaction function. It can have real-time dialog and interaction with users through voice input and speech synthesis technology. The virtual assistant can receive voice commands from users, answer questions, and provide services such as schedule

management and voice search. It can also connect with other smart devices to realize smart home control and personal assistant functions.

2.1.2. Applications of AI in digital media. AI is one of the most important applications of artificial intelligence in the field of digital media. Intelligent speech recognition technology can convert human language into computer-recognizable text, while speech synthesis technology can convert text into human-understandable speech. Intelligent speech recognition technology uses sound signal processing and pattern recognition and other technical means to analyze and decode the input speech so as to convert it into text information. It can help users realize a more convenient and fast input method and improve the accessibility and user experience of digital media [19].

Speech synthesis technology, on the other hand, converts textual information into human-understandable speech output. It uses algorithms and acoustic models of synthesized speech to process and synthesize the input text to produce human speech with natural fluency.

2.1.3. The role of multimodal sentiment analysis in user experience. Multimodal sentiment analysis is one of the important applications of artificial intelligence technology in user interaction and experience. It analyzes the user's language and behavior through techniques such as natural language processing and machine learning to understand their emotional state and mood changes [20].

Sentiment analysis recognizes and understands emotional information in text, speech and images. Combining emotional information with data about the user's behavior, feedback, and preferences allows for more personalized services and suggestions. Sentiment analysis, on the other hand, addresses the segmentation of emotions and can identify different emotional states of the user during the interaction.

2.2. Multimodal data feature extraction

2.2.1. Text Feature Extraction. Among the text sentiment feature extraction methods, the mainstream word vector models are Word2Vec, Glove, BERT and so on. Word2Vec model's mainly contain Skip-gram model and CBOW model. The main idea of CBOW model is to predict the target word based on the context word, while the main idea of Skip-gram model is that the model predicts the context of the target word based on the target word. However, Word2Vec model takes into account the local contextual information of the word, but ignores the global information between the word and the whole text. In order to solve the above problem, researchers proposed Glove model. The main idea of Glove model is the number of occurrences of each word in the corpus, constructing the word co-occurrence matrix, and then optimizing the word vector representations by minimizing the loss function, so that the words with similar meanings are closer to each other in the vector space.

However, the Glove model is still insufficient in semantic understanding and context modeling, for this reason, researchers proposed the BERT model, which is a deep bi-directional pre-trained language representation model based on the Transformer architecture. Unlike traditional unidirectional language models, BERT utilizes the encoder structure of Transformer, takes contextual information into account in the pre-training phase, and learns a rich feature representation through self-supervised tasks such as masked language modeling (MLM) and next sentence prediction (NSP).

2.2.2. Speech feature extraction. Speech, as the most widespread and natural way of communication in our real life, carries rich acoustic and semantic information. Acoustic information includes the frequency, intensity, and time-domain characteristics of the audio signal, while semantic information covers the content of speech and the meaning and emotion behind it. In order for computers to accurately understand and process speech information, speech features need to be extracted first, of which spectral features, tonal features and rhythmic features are closely related to emotions.

Spectral features are a reflection of the correlation between changes in the shape of the vocal tract

and vocal motion, and MFCC features, as one of the spectral features, are widely used in speech recognition. Based on the research on the auditory properties of the human ear, in order to simulate the human ear's perception of the sound frequency, a logarithmic scale transformation is used for the part above 1khz and a linear scale transformation is used for the part below 1khz in the Mel filter.The transformation between the Mel frequency and the Hz frequency can be shown in equation (1).

$$Mel(f) = 2595\log_{10}(1 + f/700) \tag{1}$$

where f is the frequency.

The original speech signal is first pre-emphasized to highlight the characteristics of the high frequency part. Next, the pre-emphasized speech signal is segmented into short time windows, and then a Fourier transform is applied to each window to convert the signal to the frequency domain.

Sound quality features are specific attributes of a sound that are used to measure the clarity and quality of the sound. These features are usually extracted by analyzing the spectrum and waveform of the sound signal.

Rhythmic features are the information contained in the speech signal related to speech rhythm, pitch and speech beat. The emotion recognition ability of rhythmic features has been widely recognized by researchers in the field of speech emotion recognition. Commonly used rhythmic features are time duration, fundamental frequency, energy and so on.

2.2.3. Image Feature Extraction. Image feature extraction is an important technique in the field of computer vision, aiming to extract significantly representative and distinguishable features from image data to support subsequent image analysis tasks. The traditional methods for image feature extraction are scale invariant feature transform, accelerated robust features, and directional gradient histogram. With the development of deep neural networks, a series of neural network models are widely used in image feature extraction. Among them, convolutional neural network (CNN) is the most widely used class of neural network, which realizes the effective extraction and abstraction of image features through the combination of convolutional layer and pooling layer.

ResNet introduces the concept of residual learning, which solves the problem of gradient vanishing or gradient explosion that occurs easily in deep neural networks by connecting across layers, and obtains a network structure that can be trained to greater depths.The Inception module, on the other hand, adopts multiple convolutional kernels of different sizes to work in parallel, which helps the network to learn the features of different scales at the same time, and improves the network's representation capability.

2.3. Multimodal fusion methods

Although many research results have been achieved in emotion recognition based on a single modality, emotion itself is a combination of features from many different modalities, and emotion recognition using only a single modality cannot fully capture the rich expression of emotion, so the development of multimodal fusion technology is crucial. The core of sentiment analysis utilizing multiple modalities lies in the fusion of different modalities. Multimodal fusion refers to filtering and extracting the desired features from data received from multiple sources and fusing them using different fusion strategies.

2.3.1. Model-independent fusion methods. Early fusion, also known as feature fusion combines all the features of each modality into a single feature vector, which is then fed into the classification algorithm.

In late fusion, the features of each modality are first processed and classified independently. The classification results are then fused to form a final decision vector, which subsequently produces a sentiment prediction.

Hybrid fusion refers to an approach that uses a combination of different fusion strategies to better

integrate multimodal data and improve system performance. The advantage of hybrid fusion is that it can synthesize the strengths of multiple fusion methods and compensate for the shortcomings of their respective single methods, but due to the combination of multiple different fusion strategies, the structure of the model may become more complex, resulting in a more difficult training process.

2.3.2. Model-based fusion methods. Multi-Kernel Learning (MKL) methods, as an extension of Kernel Support Vector Machines (SVMs), allow different data modalities to use different kernel functions for better fusion of heterogeneous data. Each modality enhances data fusion by selecting a specific kernel function to improve model performance.MKL has the advantage of a convex loss function, which can be used to train the model using standard optimization packages and obtain a globally optimal solution.

Deep neural networks have been widely used for multimodal fusion tasks. Neural network models such as recurrent neural networks and Transformer are widely used in multimodal fusion tasks.RNNs excel in processing time series data or data with temporal nature through their ability to model sequences. Compared to the RNN model, Transformer is able to efficiently capture dependencies over longer distances using a self-attentive mechanism. The combination of these neural network models makes the fusion of multimodal data more efficient and accurate. One of the important advantages of deep neural network methods in data fusion is their ability to learn from large-scale data.

2.4. Multimodal Sentiment Analysis Method Based on Long and Short-Term Features

In this paper, the long and short-time sentiment features extracted from the long and short-time sequences as a starting point that is, there are complementarities and differences, so how to effectively utilize the complementarities and differences between them becomes a problem that has to be thought about. Decision-level fusion can take the emotion classification results from different classifiers as a priori knowledge and effectively improve the accuracy of emotion classification of the tested video.

In this paper, a multimodal sentiment analysis method based on long and short-time features and decision-level fusion is proposed. The short-time sentiment features and long-time sentiment features of text, expression and speech are extracted from short-time and long-time sequences, respectively, and the tri-modal feature fusion is performed to obtain the short-time fusion features and the long-time fusion features, and then the classification confusion matrices corresponding to the short-time sentiment classification results and the long-time sentiment classification results are obtained.

The method proposed in this chapter uses a neural network model based on Bi-GRU and dual-attention mechanism with a neural network model based on long-time features and LMF to classify the short-time and long-time sentiment of the video under test, respectively, and performs a decision-level fusion of the classification results using the corresponding classification confusion matrices to obtain the final sentiment category of the video under test.

The following section describes the decision-making strategy proposed in this paper, which is applied to the multimodal sentiment analysis method based on short- and long-time features and decision-level fusion. Five trained neural network models based on Bi-GRU and dual-attention mechanism are used to classify the test set samples with short-time sentiment respectively, and the results of the five classifications are counted [21]. The short-term sentiment classification confusion matrix S is obtained.

$$S = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,C} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ s_{C,1} & s_{C,2} & \cdots & s_{C,C} \end{bmatrix} \tag{2}$$

where element $s_{i,j}$ in S represents the probability that a sample of category i affective category is predicted to be category j affective category by the neural network model based on Bi-GRU and

dual-attention mechanism, $i=1,2,\cdots,C$, $j=1,2,\cdots,C$, and C denote the total number of affective categories, and in this paper, C takes the value of 7, category 1 affective category is very negative, category 2 affective category is relatively negative, category 3 affective category is negative, category 4 affective category is neutral, category 5 affective category is positive, category 6 affective category is more positive, and category 7 affective category is very positive. Then the elements on the main diagonal of matrix S are normalized to obtain the weight coefficients $\beta_{i,i}^{s}$ of the short-time emotional features of the tested video that are judged as the ith emotional category.

$$\beta_{i,i}^{s}=\frac{s_{i,i}}{\sum_{j=1}^{C}s_{j,j}},i=1,2,\ldots,C \tag{3}$$

According to Eq. (3), $\beta_{1,1}^{s}$, $\beta_{2,2}^{s}$, $\beta_{3,3}^{s}$, $\beta_{4,4}^{s}$, $\beta_{5,5}^{s}$, $\beta_{6,6}^{s}$ and $\beta_{1,7}^{s}$ are calculated respectively.

Similarly, five trained neural network models based on long-time features and LMF are used to classify the test set samples with long-time sentiment respectively, and the five classification results are counted to obtain the long-time sentiment classification confusion matrix L.

$$L=\begin{bmatrix} l_{1,1} & l_{1,2} & \cdots & l_{1,C} \\ l_{2,1} & l_{2,2} & \cdots & l_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ l_{C,1} & l_{C,2} & \cdots & l_{C,C} \end{bmatrix} \tag{4}$$

where element $l_{i,j}$ in L represents the probability that a sample of category i sentiment category is predicted as category j sentiment category by the neural network model based on long-time features and LMF, and $i=1,2,\cdots,C$, $j=1,2,\cdots,C$, and C denote the total number of sentiment categories. The elements on the main diagonal of matrix L are normalized to obtain the weight coefficients $\beta_{i,i}^{L}$ for the long-time emotional features of the tested video to be adjudicated as the i th emotion category.

$$\beta_{i,i}^{L}=\frac{l_{i,i}}{\sum_{j=1}^{c}l_{j,j}},i=1,2,\ldots,C \tag{5}$$

According to equation (5), $\beta_{1,1}^{L}$, $\beta_{2,2}^{L}$, $\beta_{3,3}^{L}$, $\beta_{4,4}^{L}$, $\beta_{5,5}^{L}$, $\beta_{6,6}^{L}$ and $\beta_{7,7}^{L}$ are calculated respectively.

Using the neural network model based on Bi-GRU and dual-attention mechanism to categorize the N discourse units of the tested video with short-time emotions respectively, the predicted probability vector of short-time emotion categories for N discourse units is obtained $\left\{p_{short,1},p_{short,2},\ldots,p_{short,N}\right\}$, which is averaged to obtain the predicted probability vector of short-time emotion categories for the tested video $p_S=\begin{bmatrix} p_{S,1} & p_{S,2} & p_{S,3} & p_{S,4} & p_{S,5} & p_{S,6} & p_{S,7} \end{bmatrix}^{T}$, where $p_{S,i}$ denotes the probability that the tested video is predicted to be in the i th category of emotions, subscript S denotes the short-time sentiment category, and superscript T denotes the transpose.

Using the neural network model based on long-time features and LMF to categorize the long-time sentiment of N discourse unit of the tested video respectively, we obtain the long-time sentiment category prediction probability vector of N discourse units $\left\{p_{long,1},p_{long,2},\ldots,p_{long,N}\right\}$, and average it to obtain the long-time sentiment category prediction probability vector of the tested video $p_L=\begin{bmatrix} p_{L,1} & p_{L,2} & p_{L,3} & p_{L,4} & p_{L,5} & p_{L,6} & p_{L,7} \end{bmatrix}^{T}$, where $p_{L,i}$ denotes the probability that the tested video is predicted to be in the ith category of sentiment, the subscript L denotes the long-time sentiment category, and the superscript T denotes the transpose of the vector [22].

The long-time and short-time emotion category prediction probability vectors (p_L and p_S) of the tested video are weighted and fused to obtain the emotion category prediction probability vector p of the tested video.

$$p=\begin{bmatrix}p_1\\p_2\\p_3\\p_4\\p_5\\p_6\\p_7\end{bmatrix}=\begin{bmatrix}\beta_{11}^{s}p_{S,1}\\\beta_{2,}^{s}p_{S,2}\\\beta_{33}^{s}p_{S,3}\\\beta_{4,4}^{s}p_{S,4}\\\beta_{5,5}^{s}p_{S,5}\\\beta_{6,6}^{s}p_{S,6}\\\beta_{7,5}^{s}p_{s,7}\end{bmatrix}+\begin{bmatrix}\beta_{L,1}^{L}p_{L,1}\\\beta_{2,2}^{L}p_{L,2}\\\beta_{3,3}^{L}p_{L,3}\\\beta_{4,4}^{L}p_{L,4}\\\beta_{5,5}^{L}p_{L,5}\\\beta_{6,6}^{L}p_{L,6}\\\beta_{7,5}^{L}p_{L,7}\end{bmatrix}=\begin{bmatrix}\beta_{11}^{S}p_{S,1}+\beta_{11}^{L}p_{L,1}\\\beta_{2,}^{S}p_{S,2}+\beta_{22}^{L}p_{L,2}\\\beta_{3,3}^{S}p_{S,3}+\beta_{33}^{L}p_{L,3}\\\beta_{4,4}^{S}p_{S,4}+\beta_{4,4}^{L}p_{L,4}\\\beta_{5,5}^{S}p_{S,5}+\beta_{5,5}^{L}p_{L,5}\\\beta_{6,6}^{S}p_{S,6}+\beta_{6,5}^{L}p_{L,6}\\\beta_{7,7}^{S}p_{S,7}+\beta_{7,3}^{L}p_{L,7}\end{bmatrix} \tag{6}$$

where p_i is the probability that the video under test is predicted to be of type i sentiment category.

3. Experimental results and comparative analysis

3.1. Experimental data set

The experiments in this chapter use the MOSI dataset, which is one of the current mainstream datasets for multimodal sentiment analysis tasks, advancing the field of multimodal sentiment analysis. It contains 96 Youtube movie review videos from 92 speakers; these 92 individuals contain 44 males and 48 females, aged between 20 and 30 years old. The dataset segments the movie review videos in one sentence and takes the segmented short videos as the raw data, then processes the short videos using speech, text image and other related feature extraction tools to obtain text, speech and image data samples with a sample capacity of 2205 for all the three modalities.The MOSI dataset utilizes manual annotation to score the sentiment of each piece of data, with a score of -3 to a decimal between 3, where a lower score indicates a more negative emotion, a higher score indicates a more positive emotion, and 0 indicates neutrality, and the distribution of data in different categories in the MOSI dataset is shown in Figure 1.

It can be seen that the number of Very positive and Very negative distributions are in the first and second places, which are 483 and 442, respectively.

In this experiment, for the two-categorization task, we labeled the samples with labels between 0 and 3 (including 0) as positive categories, and between -3 and 0 as positive categories. For the seven-classification task, the original dataset labels were used directly for training.

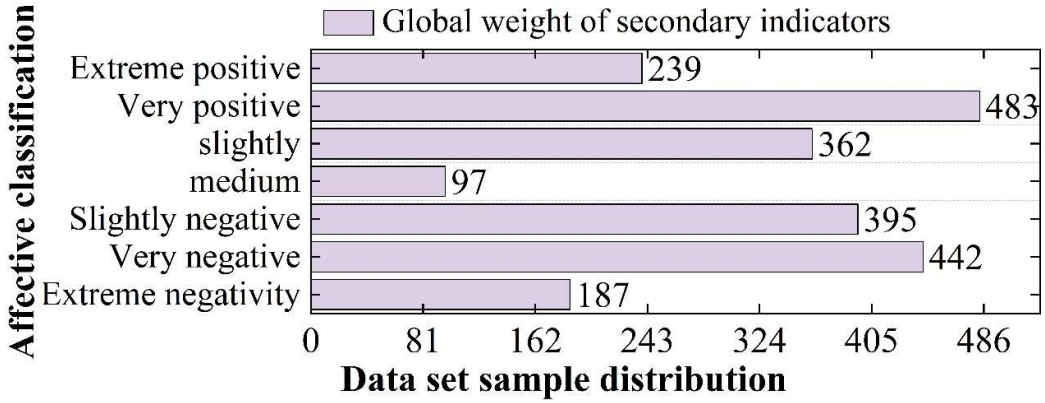

Fig. 1. Sample distribution of MOSI data set

3.2. Experimental setup

The experiments in this chapter are based on Facebook's open-source Pytorch deep learning platform, the CPU is Intel(R) Xeon(R) Silver 4214, and a single RTX2080Ti GPU is used for training.The

experiments are conducted using 64-bit Ubuntu version 20.04.3 operating system, CUDA version 10.1, cudnn version v7.6.5, and Pytorch version 1.7. During the training process, we use the Adam optimizer to optimize the model parameters with a learning rate of 1e-4, a training period of 100 cycles, and a batch size set to 64.

The experiments in this chapter mainly use the MOSI dataset, we first divide the MOSI dataset into training samples, validation samples and test samples, of which 1,355 training samples are used for model training, 309 validation samples are used to monitor the training process of the model and adjust the relevant parameters, and 724 test samples are used to test the effect of the trained model. The purpose of the experiments in this chapter is to validate the effectiveness of the multimodal sentiment analysis method based on uncertainty estimation, therefore, in order to validate the unimodal sentiment analysis part of the model, we carry out experimental validation of the text, speech, and image modalities, respectively; in addition, in order to validate the multimodal fusion part of the model, we carry out experimental validation of the combination of multiple bimodal modalities as well as multimodal modalities, respectively. Table 1 shows the comparison results of the accuracy rate of the seven-classification task, as well as the comparison results of the accuracy rate of the two-classification task and the F1 value under different modalities.

As shown in Table 1, T denotes the sentiment analysis of text modality, A denotes the sentiment analysis of speech modality, V denotes the sentiment analysis of image modality, T+A denotes the bimodal sentiment analysis of text and speech, T+V denotes the bimodal sentiment analysis of text and image, A+V denotes the bimodal sentiment analysis of speech and image, and T+A+V denotes the multimodal sentiment analysis, and all of the above experiments were conducted using the uncertainty estimation of the sentiment analysis model.

From the data shown in Table 1, it can be seen that for the experiments using uncertainty estimation-based sentiment analysis for unimodal sentiment analysis, the recognition accuracy rate of text modality is the highest, with its seven-classification accuracy rate of 35.44%, its two-classification accuracy rate of 79.69%, and the F1 value of 0.8052, while the speech modality has the lowest effect on sentiment analysis. In the bimodal sentiment analysis experiments where uncertainty estimation was used, text and image based bimodal sentiment recognition was the most effective, with a seven classification accuracy rate of 39.3%, a binary classification accuracy rate of 83.84%, and an F1 value of 0.8391, while speech and image based bimodal sentiment recognition was less effective. It can be seen that among text, speech and image modalities, text modality is more important for sentiment recognition, and introducing information from other modalities on the basis of unimodal sentiment analysis can help to improve the effect of sentiment recognition. In addition, our multimodal sentiment analysis for text, speech, and image achieves 40.57% and 84.58% for the seven-classification and two-classification accuracies, respectively, and the F1 value reaches 0.8414, which further improves the recognition effect relative to the bimodal sentiment analysis. Based on the experimental results, we verify the effectiveness and feasibility of the unimodal sentiment recognition method based on Dirichlet distribution and subjective logic theory and the multimodal fusion method proposed in this chapter.

Table 1. Mosi data set of modal experimental results

Model	Acc-7(%)	Acc-2(%)	F1
T	35.44	79.69	80.52
A	30.04	67.63	65.79
V	32.52	73.98	73.88
T+A	37.01	81.59	81.32
T+V	39.3	83.84	83.91
A+V	32.32	77.36	75.97
T+A+V	40.57	84.58	84.14

In order to further verify the recognition effect of the multimodal sentiment analysis model based on uncertainty estimation proposed in this chapter, this paper compares several more classical multimodal sentiment analysis methods in recent years. The experimental results are shown in Table 2.

From the data in Table 2, it can be seen that the ICCN model and the MISA model outperform the other models both in the seven-classification task and the two-classification task, thanks to the fact that both have the ability to take into account the intra- and inter-modal feature representations. In contrast, the model in this paper introduces uncertainty estimation to evaluate and participate in the fusion of the prediction results of different modalities, and achieves the best results in the accuracy index in both the seven-classification and the two-classification tasks, in which the seven-classification accuracy rate reaches 40.75%, and the two-classification accuracy rate reaches 83.37%, which demonstrates that the multimodal sentiment analysis model based on uncertainty estimation proposed in this chapter is able to ensure the classification performance in the and also outputs uncertainty estimation of the prediction results.

Table 2. Mosi data set compared with experimental results

Model	Acc-7(%)	Acc-2(%)	F1
TFN	34.92	80.85	80.74
MFN	34.18	77.41	77.35
MFM	35.45	81.78	81.66
ICCN	39.07	83.09	83.28
MISA	37.59	82.16	82.04
This model	40.75	83.37	83.14

3.3. Sample analysis

The multimodal sentiment analysis model based on uncertainty estimation proposed in this chapter has the advantage of being able to output the uncertainty estimation of the prediction results while ensuring the recognition accuracy. Therefore, to further validate the reliability and accuracy of the model's estimation of uncertainty, we selected three test samples from the MOSI dataset, and their outputs are shown in Table 3.

In Table 3, the prediction results of the model and the uncertainty estimates for unimodal and multimodal are demonstrated. Where ut, uv, and ua denote the uncertainty of text, image, and speech modal sentiment analysis results, respectively, and u denotes the uncertainty of multimodal sentiment analysis results. Uncertainty estimation can be used to assess the reliability of the prediction results. In Sample 1, the text, image and speech modalities express different degrees of positive emotions, so the uncertainty of the prediction results is low in both unimodal and multimodal modalities. In Sample 2, only the text modality has obvious negative emotions, and there is enough evidence to support the prediction results, so the text modality has low uncertainty while the speech and image modalities have high uncertainty, and due to the existence of modalities with low uncertainty, the final prediction results will not have high uncertainty according to Dempster's fusion rule. In Sample 3, when faced with words with ambiguity, the textual modality has higher uncertainty, when the speech and image modalities do not have strong enough emotional tendencies, and although the ambiguity of the text leads to a wrong prediction result, the final uncertainty estimation suggests that the credibility of this prediction result is low.

Table 3. Test the output of the sample

N	Text	Image	Speech	ut/uv/ua	u	Forecast	Tags
1	It looks really good.	Smile	Rising tone, normal volume	0.06/0.08/0.11	0.02	Positive	Positive
2	I thought that film was awful	Blink	The tone is flat, the narrative tone	0.08/0.36/0.43	0.08	Negativity	Negativity
3	It not bad in the way I was expecting it to be bad	Bow down	Flat tone, normal volume	0.48/0.22/0.45	0.47	Negativity	Negativity

4. AI-based interaction design strategies in digital media

4.1. Basic Principles of Interaction Design for Digital Media

4.1.1. Principle of consistency between design elements and emotional memory. Consistency in interaction design can make the interface more predictable, can reduce the user's learning cost, etc. It is one of the most common principles in design. However, the consistency in this paper is the consistency between design elements and emotional memory, which wants to keep the main features

of objects or things or feelings in design elements and emotional memory consistent. The consistency of design elements and emotional memory requires that human emotional memory be followed in the interaction logic, interface language or organization level, so as to better help people form and evoke emotional memory.

4.1.2. Design principles for target user differentiation . Interaction design is human-centered, focusing on users' behavior and psychological needs. Different people have different needs, and people with different needs also have different experiences and emotional memories, so interaction design needs to clarify the target users. Taking the application as an example, before carrying out the interaction design, the designer needs to analyze the functional requirements of the target users, and determine the target user's preferred mode of operation and the expected effect. The principle of target user differentiation, which takes into account gender differences, special age groups, and cultural differences, can help designers better provide users with convenient and effective interaction design.

4.1.3. Principle of technical assistance. Interactive products through the integration of digital media technology and Internet technology for people to build interactive bridges, so that the user in the process of using the product through the action and feedback formed by the exchange of information. At present, many digital media technologies have been developed and matured, we can use these technical means to better assist us to improve the interaction design experience, promote human emotional communication, which is also the advantage of the application of digital media technology in interaction design.

In addition, digital media technology helps to guide users to form intuitive interactive operation behavior, so that users can better use and experience the product. Since user needs are clear, easy-to-use, and in line with the operating habits of the interaction, directly interacting with virtual objects is the most intuitive and effective way shown in the gesture study, avoiding the need for users to try and fail many times before they can find the way of interaction operation hidden under the layers of menus and tutorials. Virtual reality interaction technology allows us to operate products in three dimensions using keyboards, mice, joysticks and touch screens.

4.2. Interaction design methods for different target users

Looking for a breakthrough point from the perspective of differentiation of target users in interaction design, gender, age and culture, as the three basic differences of users, are the factors that we need to differentiate and consider when carrying out interaction design. Based on the previous research on the source and difference of emotional memory content, the author will propose the design principles of emotional memory applied to digital media interaction design from three different perspectives: gender difference, age difference and cultural difference.

4.2.1. Gender-specific interaction design:

1) Interaction Design for Women. Survey research shows that women are good at image thinking, which makes them seem more careful than men and have a strong sixth sense (also known as intuition), so they pay more attention to details and are good at discovering subtle changes, and they prefer to analyze and deal with relatively stable work. When designing for female users in interaction design, the following design methods should be paid attention to:

Utilize delicate and rich images to assist visual interaction. In the previous study, it was said that women react more strongly to emotional stimuli and their emotional memories are more durable and easier to be evoked, so it is necessary to make full use of visual stimulation to enhance the pleasure of emotional experience in the interaction process of women. Therefore, it is necessary to make full use of visual stimulation to enhance the pleasure of the emotional experience during female interaction. It is necessary to consider the shape, color, and style of the design elements, etc. Most women like more

bright and warm color combinations, and soft and sensual graphic images. Take the female favorite "Mushroom Street" as an example, it is a Taobao shopping and communication platform for young women in their 20s.

2) Interaction Design for Men. Research shows that men are good at logical thinking, which makes them naturally like to challenge things with high difficulty coefficients, but they are also often not good at expressing their emotions and feelings. When designing for male users in interaction design, you should pay attention to the following design methods:

Utilize hard and textured images to aid visual interaction. Men prefer black, gray and then white because these colors are more stable and low-profile, followed by cooler blue. In addition, thick lines, textured materials and angular shapes are masculine design elements that are more appealing to men.

4.2.2. Interaction design for different ages:

1) Interaction design for children. The development of science and technology not only affects adults, but also the influence of digital products on children's growth and learning is gradually deepening, especially some educational software or educational digital media has been generally accepted by parents. Based on the above research, the author believes that the interaction design of digital media for children should utilize children's emotional memory characteristics, and pay attention to the following methods in the design:

Use simple graphics, bright colors, friendly images and character settings. Children pay more attention to simple graphics and high purity and high brightness colors, and their living environment is also very simple and they have less contact with people, and most of the people they can remember tend to give them a warm and intimate feeling of character images. Therefore, simple and clear graphics and colors, cute and affectionate images or characters can bring them emotional memory reproduction. As shown in the following figure, the cartoon image of the friendly mother chick uses a simple round shape, and the color is orange-red with high brightness and purity.

2) Interaction design for the elderly. Based on the current development of the aging society in China, the demand for digital media interaction products for the elderly is also greatly increasing. Through the analysis of the emotional memory of the elderly above, the author believes that the digital media interaction design for the elderly should take advantage of the characteristics of the emotional memory of the elderly, and pay attention to the following methods in the design:

Use precise context for expression. Because the era in which the elderly live causes the vocabulary and language they are familiar with to be very different from the popular language of modern society, so they can't understand the modern popular language well, so in the design, pay attention to the use of words that are easy to understand by the elderly, which will help them to understand and operate better when they are learning and using the electronic devices in connection with their past experience.

5. Interaction design effectiveness and user evaluation

Validation of the program proposed in this study. After completing the deployment of the digital media environment, experimental subjects were recruited to experience the digital media design example in two ways. The first one: experiencing digital media with feedback regulation collecting data as an experimental group. The second, experience only traditional digital media and collect data as a control group. At the end of the experience, complete the digital media feedback experience questionnaire registration. Twenty subjects were recruited, with a male to female ratio of 1:1, aged 23-27 years old, of whom five had already experienced digital media.

5.1. Subjective evaluation

There are four main aspects of subjective evaluation: user experience, emotional arousal effect, pleasure level, and emotional feedback effect. Among them, UX is a comprehensive assessment, including immersion, presence, interactivity, and imagination. Figure 2 shows the visualization of UX

assessment.

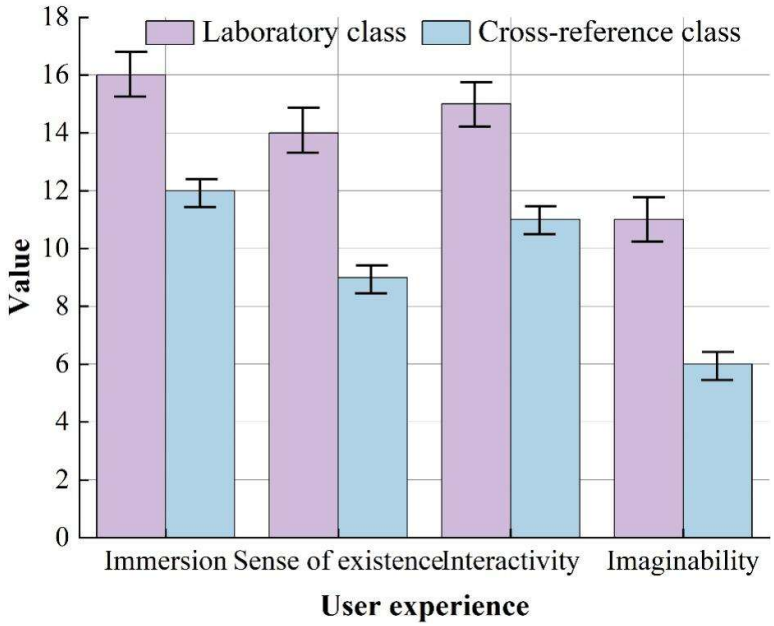

Fig. 2. The visualization of user experience

In the evaluation of the degree of emotional arousal, for the three target emotions of Happiness, Disgust and Surprise, the degree of emotional arousal can be enhanced by changing the brightness and the volume of background music in the digital media environment. The main reason for this is that the subjects' sense of immersion and presence in the digital media environment with emotional feedback is enhanced, which makes it easier to awaken emotions. For the natural state (Neutral), the digital media with emotional feedback makes users deliberately focus on their own body state, which creates a kind of psychological expression of emotion, and thus is lower than the control group data. The visualization of the degree of emotional arousal is shown in Figure 3.

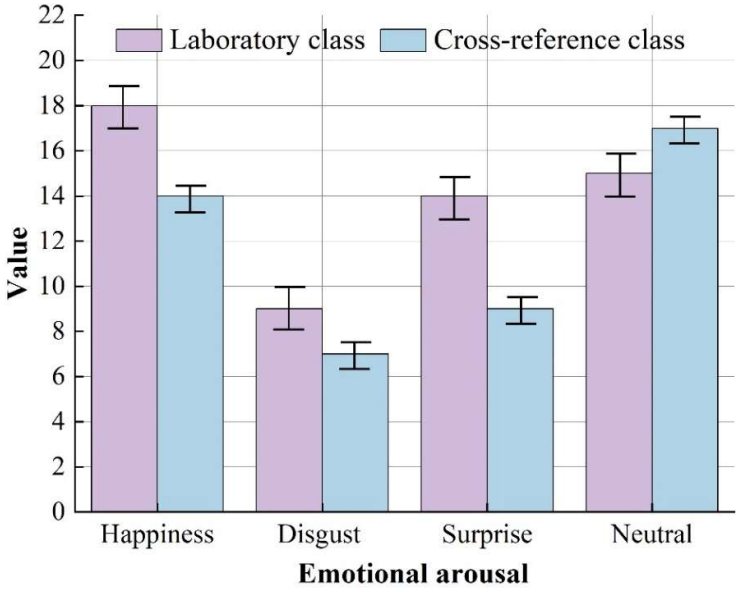

Fig. 3. Emotional arousal

Of the two different digital media experiences, 75% of the subjects perceived the feedback-

moderated digital media as having a better level of enjoyment, which was 25% higher compared to the digital media experience with a lack of feedback mode. In addition, the ratings of "very good" (9-10) were 15% to 20% higher than those of the control group, which can reflect that emotional feedback can increase the enjoyment of digital media. In the experimental group, only 4% to 6% of the subjects thought that there were problems during the experience and that emotional feedback could not increase the fun of digital media. Upon analysis, it was found that the experimental subjects, including those who thought the same in the control group, experienced digital media frequently. There is a strong adaptation to the sense of presence in the digital media environment. Compared to the experimental subjects who did not experience or infrequently experience digital media, emotional arousal requires a higher potency. The experience pleasantness scores are shown in Figure 4.

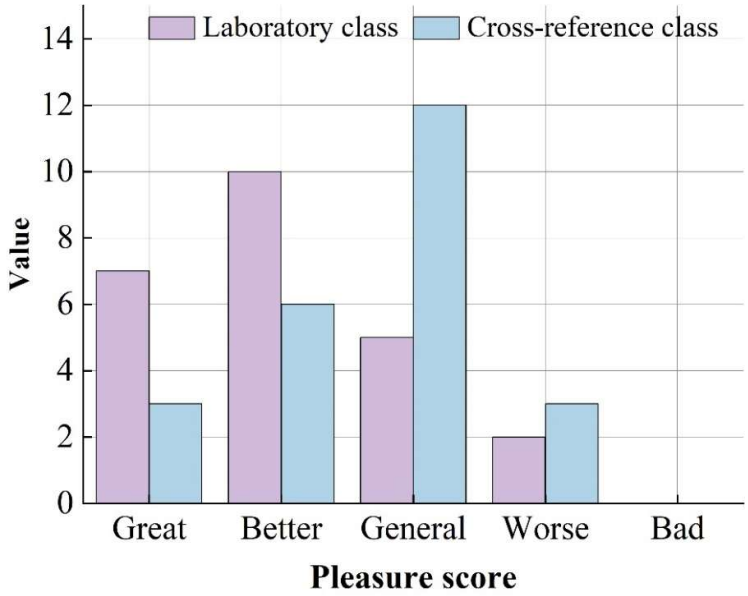

Fig. 4. Experience pleasure

For the analysis of the feedback regulation effect of the target emotion, the subjects all thought that the three emotional states of Happiness, Surprise and Neutral could achieve a good feedback effect. In the low arousal and low valence emotion Disgust does not achieve the expected feedback regulation effect, the main reasons include two aspects: (1) the user state, through the emotional feedback regulation, the experimental subjects to enhance the sense of presence in the virtual space, for the arousal of this kind of emotion caused by a certain degree of difficulty, so the arousal needs to be more valence of the emotional material. (2) In terms of hardware, the feedback regulation can reduce the effect of dazzle, resulting in less likely to produce Disgust emotion due to hardware. Specific data are shown in Fig. 5 Emotional feedback effect.

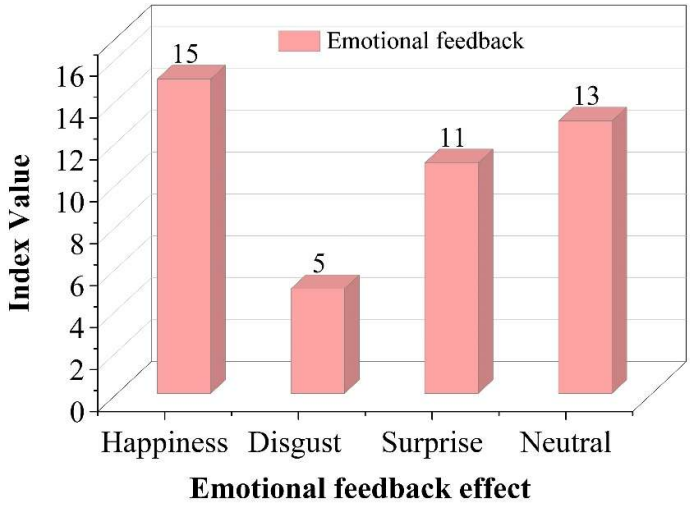

Fig. 5. Emotional feedback effect

5.2. Objectivity analysis

In order to further investigate whether the effect of feedback regulation is achieved. The user's emotional state during the digital media was recorded, and the time step was set to 1 minute, i.e., the emotional changes of all subjects were recorded within 1 minute. In the design of digital media task duration, it was not realized as expected plan, mainly because the experimental subjects needed time to adapt to the environmental changes in the VR digital media, even though the familiarization of the digital media environment had been carried out for the subjects who had no or infrequent experience. As a result, 5 minutes more time than pre-planned was spent on digital media elapsed. Data were collected for each emotional state of the 20 experimental subjects during the digital media time, and after calculating the mean values, they were visualized and expressed. Figure 6 shows the emotional changes of the subjects.

In summary, it can be said that this paper basically realizes the digital media interaction enhancement function by recognizing the user's emotional state.

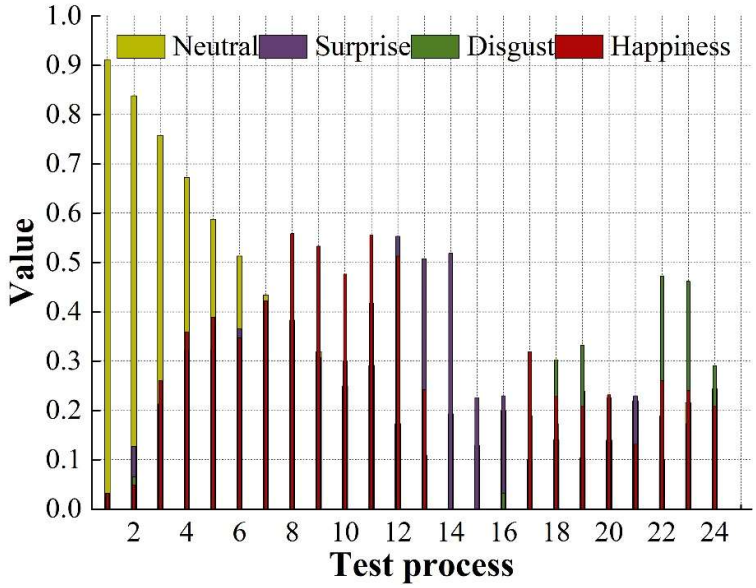

Fig. 6. The emotional changes of subjects during the feedback games

6. Conclusion

In this paper, we design an interactivity enhancement model based on multimodal fusion of AI, and utilize the multimodal fusion sentiment analysis algorithm to realize emotional interaction in digital media. The performance of the algorithm is verified on the dataset, and the interaction design strategy in digital media is proposed to analyze the interaction design effect. Uncertainty estimation is introduced into the model in this paper to evaluate the prediction results of different modalities and participate in the fusion, and the best results are achieved in the accuracy index in both the seven-classification and two-classification tasks, in which the seven-classification accuracy rate reaches 40.75%, and the two-classification accuracy rate reaches 83.37%, which demonstrates that the multimodal sentiment analysis model based on uncertainty estimation proposed in this chapter is capable of ensuring the classification performance while This shows that the multimodal sentiment analysis model proposed in this chapter is able to ensure the classification performance while also outputting the uncertainty estimation of the prediction results. Subjective evaluation of the interaction effects of digital media designed based on the strategy of this paper shows that the user experience, emotional arousal, pleasure level, and emotional feedback are all higher than those of the control group. It is able to perform digital media interaction enhancement functions by recognizing the emotional state of the user.

About the Author

Guohao Zou was born in 1993 in Jiangxi, China. He graduated from Jingdezhen Ceramic University with a PhD. He is now working in the College of Humanities and Arts, Nanchang Institute of Technology. His research interests include computational intelligence, digital media and big data analytics.

References

[1] Liu, J., Jiang, X., Shi, M., & Yang, Y. (2024). Impact of artificial intelligence on manufacturing industry global value chain position. Sustainability, 16(3), 1341.

[2] Liu, Y., Tong, K., Mao, F., & Yang, J. (2020). Research on digital production technology for traditional manufacturing enterprises based on industrial Internet of Things in 5G era. The International Journal of Advanced Manufacturing Technology, 107(3), 1101-1114.

[3] Krupa, V., Oliinyk, I., Bazaka, R., Shtangret, A., & Sylkin, O. (2024). Technical And Technological Support for Personnel Management: Digital Transformation of Enterprise Competitiveness Through Artificial Intelligence. International Journal of Religion, 5(11), 260-270.

[4] Rustan, A. (2021). Digital Communication and Social Media Interaction to Improve Academic Quality of Islamic Higher Education Lecturers. Journal of Social Studies Education Research, 12(4), 144-169.

[5] Bond, M., Marín, V. I., Dolch, C., Bedenlier, S., & Zawacki-Richter, O. (2018). Digital transformation in German higher education: student and teacher perceptions and usage of digital media. International journal of educational technology in higher education, 15(1), 1-20.

[6] Plantin, J. C., & Punathambekar, A. (2019). Digital media infrastructures: pipes, platforms, and politics. Media, culture & society, 41(2), 163-174.

[7] Schroeder, R. (2018). Towards a theory of digital media. Information, Communication & Society, 21(3), 323-339.

[8] Mihelj, S., & Jiménez‐Martínez, C. (2021). Digital nationalism: Understanding the role of digital media in the rise of 'new'nationalism. Nations and nationalism, 27(2), 331-346.

[9] Liu, P. L. (2020). COVID-19 information seeking on digital media and preventive behaviors: the mediation

role of worry. Cyberpsychology, behavior, and social networking, 23(10), 677-682.

[10] Agyekum, K. O. B. O., Xia, Q., Liu, Y., Pu, H., Cobblah, C. N. A., Kusi, G. A., ... & Gao, J. (2019). Digital media copyright and content protection using IPFS and blockchain. In Image and Graphics: 10th International Conference, ICIG 2019, Beijing, China, August 23–25, 2019, Proceedings, Part III 10 (pp. 266-277). Springer International Publishing.

[11] Boulianne, S., & Theocharis, Y. (2020). Young people, digital media, and engagement: A meta-analysis of research. Social science computer review, 38(2), 111-127.

[12] De Paula, B. (2021). Reflexivity, methodology and contexts in participatory digital media research: making games with Latin American youth in London. Learning, media and technology, 46(4), 435-450.

[13] Kirkorian, H. L. (2018). When and how do interactive digital media help children connect what they see on and off the screen?. Child development perspectives, 12(3), 210-214.

[14] Llinares, D., Fox, N., & Berry, R. (Eds.). (2018). Podcasting: New aural cultures and digital media. Springer.

[15] Lobstein, T., Landon, J., Thornton, N., & Jernigan, D. (2017). The commercial use of digital media to market alcohol products: a narrative review. Addiction, 112, 21-27.

[16] Ineji, P. U., & Ogar, I. P. (2021). Impact of digital media on effective healthcare delivery in Cross River State. International Journal of Communication Research, 11(1), 73-79.

[17] Alelaimat, A. M., Ihmeideh, F. M., & Alkhawaldeh, M. F. (2020). Preparing preservice teachers for technology and digital media integration: Implications for early childhood teacher education programs. International Journal of Early Childhood, 52(3), 299-317.

[18] Jing Yuan,Yuan Zhang,Daqing Li,Cailin Yang,Yuxin Xing & Zhanhao Jiang.(2025).Immersive human–computer interaction and digital entertainment new media application in English e-learning mode.Entertainment Computing100878-100878.

[19] Fredrick Stephanie S.,Ettekal Idean,Domoff Sarah E. & Nickerson Amanda.(2025).Patterns of child and adolescent digital media use: Associations with school support, engagement, and cybervictimization..Psychology of Popular Media(1),54-65.

[20] Zuhe Li,Panbo Liu,Yushan Pan,Jun Yu,Weihua Liu,Haoran Chen... & Hao Wang.(2024).Text-dominant multimodal perception network for sentiment analysis based on cross-modal semantic enhancements.Applied Intelligence(2),188-188.

[21] Buchi Reddy Ramakantha Reddy & Ramasamy Lokesh Kumar.(2024).A Fusion Model for Personalized Adaptive Multi-Product Recommendation System Using Transfer Learning and Bi-GRU.Computers, Materials & Continua(3),4081-4107.

[22] Chunlin Huang,Ting Zhou,Weide Li,Haijiao Yu,Rongxia Li & Jinjie Fang.(2024).A coupled model integrating dual attention mechanism into BiGRU-RED for multi-step-ahead streamflow forecasting.Journal of Hydrology(PA),132137-132137.