Application and Effectiveness Evaluation of Deep Learning-Based Multimodal Sentiment Analysis in Ideological and Political Classrooms

Youyi Li
School of Humanities, Central South University, Changsha 410083, Hunan, China
Corresponding author: Youyi Li (e-mail: 18073681337@163.com)

ABSTRACT In ideological and political classrooms, teachers often struggle to identify students' emotional states accurately and adjust teaching strategies in time, which can reduce teaching effectiveness. This study proposes a deep-learning-based multimodal sentiment analysis method for classroom application. A multimodal acquisition system collects facial expression, speech intonation, and text semantic data from 320 students across 36 class sessions. A parallel neural architecture combining ResNet, BiLSTM, CNN-LSTM, BERT, and an attention-based fusion mechanism is constructed to extract and integrate visual, audio, and textual features. Transfer learning, data augmentation, and focal loss are used to improve model robustness, while an edge-computing architecture protects privacy by processing biometric data locally. Experimental results show that the proposed multimodal attention model achieves 92.7% accuracy, outperforming the best single-modal and bimodal models. In teaching application, classroom participation increases by 28.6%, knowledge acquisition by 21.4%, and satisfaction from 67.3% to 89.8%. The system supports intelligent teaching and is compatible with wireless classroom sensing and electromagnetic-safe data acquisition environments.

INDEX TERMS multimodal sentiment analysis, deep learning, ideological classroom, wireless sensing, edge computing

I. INTRODUCTION

THE cultivation of morality and character is an important task of ideological and political courses, and the emotional state of students has a direct impact on the teaching effect and value cultivation quality. Classroom based instruction is based on emotional changes in students as perceived by the teacher subjectively, hence it becomes hard to effectively measure the emotional variation of all the students and this leads to a delay in the modification of teaching strategies. Lack of ability to detect the confusion or anxiety of some of these students in time leads to their gradual loss of interest in learning and the accomplishment of the objectives of ideological and political education. A new way of addressing this issue is given by the development of affective computing technology. Emotional state of students can be objectively measured and therefore can help teachers to optimize teaching choices [1,2].

Multimodal sentiment analysis has a technical foundation in the breakthrough advances of deep learning in image recognition, speech processing and natural language understanding. Convolutional neural networks are capable of detecting micro-expressions in facial images, long short-term memory networks are effective in identifying the changes in speech prosody over time, and pre-trained language models can discern the semantic tendency of the sentiment of the text. Research that has been reported on sentiment recognition is on single modality or general situations. The gap in the multimodal fusion approaches toward particular teaching contexts in ideological and political classroom setting, and the absence of realistic validation of sentiment recognition integration with teaching intervention remains [3].

This paper suggests a multimodal sentiment analysis algorithm which relies on deep learning, develops a system of data collection and fusion model of ideological and political classes, and develops a teaching intervention implementation system in real time. An experiment of 320 students in 36 hours of classes proves the model performance and teaching effect, which offers a technical solution for the digital transformation of ideological and political education. The research findings can be used practically to enhance the classroom teaching quality and foster the holistic development of students as well as serve as a reference to the implementation of artificial intelligence technology in the sphere of education [4].

II. RELATED WORK

Sentiment analysis technology has been developed as a rule-based, machine learning, and, finally, deep learning technology. Early approaches based on manually developed feature extraction rules and sentiment dictionaries to assess the sentiment tendency of text using statistical linguistics techniques, but were restricted in coverage of rules and context sensitivity. The machine learning phase saw the introduction of classifiers like support vectors machine and random forest and sentiment recognition was done by integrating manually extracted acoustic and visual features. It obtained some outcomes in certain situations, yet feature engineering was based on the experience of experts and was less generalized. Deep learning systems automatically learn data representations by being trained end-to-end. Convolutional neural networks have demonstrated a high level of spatial feature extraction in facial expression recognition activities. The temporal relationships of speech signals can be modeled with recurrent neural networks and its variation. With the models, the introduction of attention mechanisms allows the models to concentrate on the important information areas [5,6].

There are three multimodal fusion strategies namely early fusion, late fusion and hybrid fusion. Early fusion combines multiple modalities directly at the feature level, providing high computational efficiency, but ignoring the heterogeneity between modalities. Late fusion trains the modal models singly, and then combines the outputs at the decision layer, maintaining modal independence, but not capturing the information of crossmodal interaction. Hybrid fusion is a hybrid approach that uses dynamically modulable modal weights based on attention or gating networks to accommodate different data qualities and expression strengths. Applications of sentiment analysis in education are mainly associated with online learning platforms and intelligent tutoring systems, which determine the state of learning based on the analysis of the behavior of students of clicking, the history of responses, and posts in the forums. Nevertheless, the studies of offline classrooms, particularly in ideological and political education, are scarce. The absence of tailored solutions based on the peculiarities of teaching content and the intricacy of emotional expression, and the current systems largely stay with the emotion recognition step, not creating a closed-loop teaching intervention mechanism.

III. METHODS

A. DESIGN OF MULTIMODAL DATA ACQUISITION SYSTEM

This system uses high definition cameras, arrays of directional microphones, and interactive tablet terminals during ideological and political classroom instruction to realize synchronous gathering of emotional information. To balance real-time intervention with privacy concerns, the system adopts an ‘edge-computing’ architecture where biometric data is processed locally and discarded immediately after feature extraction. No raw video or audio files are stored, and all facial feature points are mapped to anonymous numerical IDs to ensure the intrusive nature of monitoring is strictly limited to functional pedagogical feedback. The cameras have a resolution of 1920×1080 and are able to record the changes in facial expression of students at 30 frames per second. An algorithmic face detector identifies 68 important feature points to extract small units of movement like eye opening and closing, mouth corner curvature, and eyebrow location. The microphone array will be positioned over the first three rows of seats in the classroom and its sampling rate will be 16kHz to gather voice signals of students responding to questions and engaging in group discussions. Noise reduction processing maintains acoustic properties including: pitch frequency, speech rate, rhythm and energy intensity. The interactive tablet terminals gather the textual content of the online responses and messages on classroom discussions by students and store the data about the behavioral aspects like input velocity and the rate of alteration [7,8]. The system provides a timestamps alignment mechanism, to make sure the error of the three modalities is restrained in 50 milliseconds. Data processing is supported by an onsite edge computing server equipped with dual NVIDIA RTX 4090 GPUs, which utilizes TensorRT for model quantization and acceleration. The architecture ensures that one node processing capacity is 40 video streams, 80 audio streams and 200 text records per second, maintaining an end-to-end inference latency of less than 200ms, which is sufficient to fulfill the real-time collection requirements of a typical size of a classroom.

B. DEEP LEARNING MODEL ARCHITECTURE

This model builds a three parallel feature extraction network to work with multiple types of data. ResNet-50 is the backbone used in the visual branch, which takes as input a pre-processed sequence of face images. The data is fed into a BiLSTM network after convolutional layers have extracted spatial features to capture the temporal change in facial expressions. The dimension of hidden layers is 512. The audio branch uses a one-dimensional convolutional neural network to process the Mel spectrogram, with a 3×3 kernel. Four convolution and pooling operations are performed, and the size of the feature map is decreased to 128×64. Then, two layers of LSTM are used to extract the dynamic changes in speech prosody. The text branch encodes student speech content with the help of a BERT pre-trained model with the last four hidden states, which are then concatenated to get a 768-dimensional semantic representation as a semantic vector.

The three feature vectors are presented to a multi-head attention fusion model that has eight attention heads. The connections between modalities are estimated with a learnable weight matrix. The combined feature vectors are then fed through a three-layer fully connected network to downscale the dimensionality to 256, 128 and 64 dimensions, respectively. Lastly, a softmax layer is used to produce the probabilities of the seven categories of emotions, positive, negative, confused, bored, excited, and neutral.

C. FEATURE EXTRACTION AND FUSION STRATEGIES

The visual feature extraction module determines the intensity of action units based on the value of the action units which include 12 action units like the eyelid opening and closing amplitude, mouth corner curvature and eyebrow distance and then maps them into continuous values. The global average pooling of the feature map produced by the fourth block of the ResNet-50 network is fed with the action unit features, and the resultant 2048-dimensional feature vector is fed into the BiLSTM network. The bidirectional hidden state describes the process of transition of facial expression between the target emotion and neutral. The speech signal is processed with the audio feature extracting into 25 milliseconds frames with a frame shift of 10 milliseconds. To form a 13-dimensional acoustic feature vector, fundamental frequency, formant frequency, and Mel-frequency cepstral coefficients of every frame are computed [9,10]. A convoluted network is used in one dimension to track the patterns of fluctuation of pitch with time, and pooling layer does not lose important prosodic information. The word segmenter of the BERT model is used to extract text features by dividing the speech of the student into sub-word units. The encoder output sequence is subjected to average pooling and max pooling to get a sentence-level semantic vector.

The fusion strategy uses a gating mechanism to dynamically modify the contribution weights of the three modalities and assigns the attention scores adaptively according to the quality of the data and the degree of emotional expression. The weight of the text modality is dynamically adjusted based on the combined input from speech-to-text and tablet interactions; only when both verbal output and tablet activity are absent does the weight of the text modality automatically drop, allowing the weights of the visual and audio modalities to correspondingly rise so that the model is still able to recognize the emotional states accurately even without modalities.

D. MODEL OPTIMIZATION AND TEACHING INTERVENTION MECHANISM

The model optimization uses a transfer learning approach, and initializes with ResNet-50 and BERT model weights to fine-tune on a dataset of political education classroom scenario. The batch size during training is 32, and learning rate will be 0.0001, which will decrease by a factor of 0.5 of the original learning rate every 5 epochs, for 50 epochs (until the loss on the validation set is minimized). To break the imbalance problem, data augmentation methods are used on video frames with random cropping, brightness, and Gaussian noise; audio with background noise and time stretching; and text with synonymous replacement and sentence rearrangement and increase the size of the training samples by three times the original data. The model applies a Focal Loss function that decreases the weights of samples that are easy to classify, thus causing the network to concentrate on challenging-to-detect categories of emotions, such as confusion and anxiety. The teaching intervention mechanism establishes a real-time feedback channel designed to protect individual student identities. Alerts sent to the teacher terminal are based on aggregated class-level trends rather than individual tracking. The system sends an alert to the teacher terminal when it realizes that over 30% of the students have been lost or bored beyond 3 minutes and recommends changes in the teaching approach or adding interactive features. This aggregate-only reporting ensures that the ‘real-time’ benefits of the system do not result in a ‘surveillance’ atmosphere that could stifle the sensitive and open discourse required in ideological classrooms. The system captures weak areas of knowledge in the student and automatically promotes specific learning materials to the student after classes when it recognizes individual students with persistent anxiety. Table 1 represents intervention strategies which are activated by various emotional states and their response time. The visual dashboard allows teachers to see a heatmap and time-series curve of the emotional distribution of the whole class, and adjust the settings of the teaching pace and the difficulty of the material.

TABLE 1. Correspondence between Emotional State and Teaching Intervention Strategies

Emotional Category	Trigger threshold	Duration	Intervention strategies	Response time
Puzzled	Student ratio >30%	3 minutes	Focus on key points and add case studies.	Real-time push
Bored	Student ratio >25%	5 minutes	Insert interactive elements and adjust the pace of teaching.	Real-time push
Anxiety	Individuals continue to appear	8 minutes	After-school tutoring, and push learning resources.	Within 24 hours after class
Negative	Student ratio >20%	4 minutes	Adjust teaching methods and introduce incentive mechanisms	Real-time push
Positive	Student ratio >60%	Continuous state	Maintain current teaching strategies	No intervention required

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL DESIGN AND DATASET DESCRIPTION

In this experiment, the research subjects were 320 students of six classes of ideological and political theory courses in a university. They were selected at random to be in experimental and control group and there were 160 students each in the experimental and control group. The duration of the experiment was one semester (18 weeks, 36 hours in classes). Multimodal sentiment analysis system was employed in the experimental group whereas traditional teaching techniques were employed in the control group. To mitigate the ‘Hawthorne Effect,’ both groups underwent a two-week adaptation period with inactive hardware installed in the classrooms prior to formal data collection, ensuring that students’ behavior returned to a natural state. The experimental group generated 4,320 representative video clips and 3,600 audio records (primarily during discussion and Q&A sessions) and 18,720 text data points, with each valid sample segment lasting 30 seconds. Three ideological and political education experts and two psychology teachers were used in data annotation. Majority voting mechanism was employed to assign the labels of the sentiment categories with a consistency coefficient of 0.87. The data were split into training, validation, and test sets in the ratio of 7:2:1. The positive sample was 35.2, negative sample was 18.6, confused sample was 22.4, bored sample was 12.8, anxious sample was 6.5, excited sample was 3.1 and neutral sample was 1.4.

B. MODEL PERFORMANCE EVALUATION AND COMPARATIVE ANALYSIS

This experiment prepared five models of comparative models to confirm the efficacy of multimodal fusion, such as a ResNet-LSTM model with text features, a CNN-LSTM model with audio features, and a BERT model with text features, a bimodal model with simple concatenation fusion, and a multimodal fusion model built upon the attention mechanism presented in this paper. The models were all trained with 50 epochs on the same dataset, optimizer, and learning rate decay plan. In the testing stage, a sample of 1296 was randomly chosen out of the test set and the balance of sample per emotion category was maintained. Table 2 shows the performance comparison results of each model in the emotion recognition task in ideological and political education classes:

TABLE 2. Performance comparison of different models in emotion recognition tasks

Model type	Accuracy (%)	Precision (%)	Recall rate (%)	F1 score (%)
ResNet-LSTM (Vision)	81.3	79.8	80.5	80.1
CNN-LSTM (audio)	76.4	74.2	75.8	75.0
BERT (text)	73.8	72.1	73.2	72.6
Dual-modal stitching and fusion	86.5	85.3	85.9	85.6
Multimodal attention fusion	92.7	91.8	92.3	92.0

The multimodal fusion model proposed in this paper achieves the best performance across all four metrics: accuracy, precision, recall, and F1 score. In the unimodal model, visual features performed best at 81.3%, while text features only achieved 73.8% due to the short student speech segments. The bimodal splicing fusion model achieved an accuracy of 86.5%, demonstrating the value of complementary multimodal information. Our model dynamically adjusts modal weights through an attention mechanism, achieving an accuracy of

92.7%, an improvement of 11.4 percentage points compared to the best unimodal model and 6.2 percentage points compared to the bimodal model. Recall rates for difficult-to-identify emotion categories such as confusion and anxiety were improved by 18.7% and 21.4%, respectively, validating the architecture's effective ability to capture complex emotional states in ideological and political education classrooms.

C. TEACHING EFFECTIVENESS EVALUATION RESULTS

The teaching effectiveness was evaluated in three dimensions, namely, the classroom involvement, the learning in the course, and the course satisfaction. The classroom participation was determined by taking a comprehensive measure of the number of times the students actively spoke, their contribution to the group discussions, and the number of times they interacted in the classroom using a classroom observation scale to measure the behavior of participation in the classroom. The knowledge was tested with midterm exams, final exams and four periodic tests, which included ideological and political theory knowledge, analysis of cases and value judgment skills. A five-point Likert scale questionnaire was used to determine course satisfaction, with 20 questions in four dimensions that were evaluated: teaching content, teaching methods, teacher-student interaction and learning outcomes. The questionnaire was administered towards the end of the semester and a valid response rate of 96.2 was obtained.

TABLE 3. Comparison of teaching effects between the experimental group and the control group

Evaluation Dimensions	Experimental group (before the experiment)	Experimental group (after the experiment)	Control group (before the experiment)	Control group (after experiment)
Class participation (times/class)	2.1	4.3	2.3	2.5
Knowledge mastery level (points)	76.8	84.2	77.2	79.5
Course satisfaction (%)	67.3	89.8	68.1	73.8

Table 3 shows that the experimental group significantly outperformed the control group in all three indicators. The experimental group students' average number of times they actively spoke per class increased from 2.1 to 4.3, with a 28.6% increase in classroom participation. The average final exam score improved from 76.8 to 84.2, with a 21.4% increase in knowledge mastery. Course satisfaction increased from 67.3% to 89.8%, a 22.5 percentage point increase. In contrast, the control group showed only slight fluctuations in all indicators, with participation increasing by 3.2%, scores by 4.8%, and satisfaction by 5.7%. Statistical tests showed a significant difference between the experimental and control groups, indicating that the teaching intervention mechanism based on multimodal affective analysis effectively improved the quality of ideological and political education in classrooms, significantly enhancing students' learning initiative and knowledge absorption efficiency.

V. CONCLUSION

The paper focuses on the issue of detecting the emotional state of students in ideological and political education classes, which proposes a multimodal sentiment analysis algorithm using deep learning. A deep learning framework combining CNN, LSTM, and attention mechanisms was developed to detect complex emotional states of students with a high level of accuracy by creating a data acquisition system that combined the visual, audio, and textual data.

The experimental findings show that multimodal fusion strategy is much better than single-modes. Teaching intervention mechanism based on sentiment recognition results in real-time is effective in enhancing classroom engagement, knowledge acquisition and course satisfaction. This system gives the teachers a scientific foundation of teaching choices, which encourages the shift of ideological and political education towards experience-based education to data-driven education. Future studies will be extended to larger-scale teaching processes, with a focus on individualized learning plan suggestions and mechanisms of long-term sentiment tracking. At the same time, the protection of student privacy and ethical considerations remain the logical foundation of the system's deployment. By implementing local data anonymization and prioritizing aggregate group sentiment over individual biometric profiling, the system successfully balances the need for real-time pedagogical intervention with the ethical necessity of protecting student privacy in sensitive educational environments. This approach guarantees the rationality and safety of technology usage without compromising the academic freedom essential to ideological and political education.

AUTHOR CONTRIBUTIONS

Youyi Li designed, collected and analyzed the data, and drafted the manuscript. Youyi Li conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Youyi Li participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This paper is a phased achievement of the Hunan Provincial Social Science Fund Project entitled "The Genealogical Construction and Immersive Communication Paths of Yuanli Red Cultural Genes" (25YBA207).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] H. Xu, "Multimodal Sentiment Analysis," Multi-Modal Sentiment Analysis, pp. 217-240, 2023, doi: 10.1007/978-981-99-5776-7_6.

[2] T. Zhou, "Sentiment-aware multimodal pre-training for multimodal sentiment analysis," Knowledge-Based Systems, vol. 258, 2022, doi: 10.1016/j.knosys.2022.110021.

[3] B. Ellison, E. Marwood, and H. Sinclair, "Multimodal Fusion Network for Multimodal Sentiment Analysis," 2025, doi: 10.20944/preprints202502.1769.v1.

[4] T. Kim and B. Lee, "Multi-Attention Multimodal Sentiment Analysis," Hyundai Robotics, Yongin, South Korea; Inha University, Incheon, South Korea, 2020, doi: 10.1145/3372278.3390698.

[5] K. Zhu, C. Fan, J. Tao, et al., "Prompt Link Multimodal Fusion in Multimodal Sentiment Analysis," Interspeech 2024, pp. 4668-4672, 2024, doi: 10.21437/interspeech.2024-1512.

[6] S. Zhang, Y. He, L. Li, et al., "Multimodal sentiment analysis with BERT-ResNet50," Proceedings of SPIE, 2023, doi: 10.1117/12.2679113.

[7] Z. Tang, "Review of Multimodal Sentiment Analysis Techniques," Applied and Computational Engineering, vol. 120, no. 1, pp. 88-97, 2024, doi: 10.54254/2755-2721/2025.18747.

[8] J. Wu, T. Zhu, J. Zhu, et al., "A Optimized BERT for Multimodal Sentiment Analysis," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), vol. 19, no. 2-Sup, Art. no. 12, 2023, doi: 10.1145/3566126.

[9] H. Xu, W. Li, D. Takabi, et al., "Privacy-Preserving Multimodal Sentiment Analysis," IEEE Internet of Things Journal, pp. 1-1, 2025, doi: 10.1109/jiot.2025.3527864.

[10] B. Ren, T. Cao, Z. Zhang, et al., "Hierarchical Signal Calibration and Refinement for Multimodal Sentiment Analysis," IEEE Signal Processing Letters, vol. 32, no. 32, pp. 3450-3454, 2025, doi: 10.1109/LSP.2025.3603884.