IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 2, April, 2023, pp. 533-541
Journal homepage: http://iieta.org/journals/ts

Multi-Modal Affective Computing: An Application in Teaching Evaluation Based on Combined Processing of Texts and Images

Shunye Wang[1*], Yin Zhai[1], Guifang Xu[2], Ning Wang[1]

[1] College of Electronic and Information Engineering, Langfang Normal University, Langfang 065000, China
[2] College of Education, Langfang Normal University, Langfang 065000, China

Corresponding Author Email: wangshunye@lfnu.edu.cn

https://doi.org/10.18280/ts.400212

Received: 10 January 2023
Accepted: 7 April 2023

Keywords:
text and image processing, multi-modal affective computing (MAC), teaching evaluation

ABSTRACT

Conventional teaching evaluation emphasizes students' knowledge mastery over their affections. Multi-modal Affective Computing (MAC) can analyze versatile information of students in the classroom, including their facial expressions, gestures, and text feedback, in a comprehensive way, thereby helping teachers discover problems with students' affections in a timely manner, so that they could adjust the teaching methods and strategies accordingly. However, the available MAC technology might make unstable or wrong judgement when dealing with complex affective expressions, then the inaccurate evaluation results of students' affection state might adversely affect the teaching evaluation results. To tackle these issues, this study innovatively applied MAC in teaching evaluation based on combined processing of texts and images. The input texts were divided into two parts: main body and the hash tag, which were subjected to feature extraction respectively. The image features were extracted from two angles: object and scene, since the two angles can give image information of different levels. The MAC model was divided into modal sharing tasks and modal private tasks to attain better adaptability in case of new teaching evaluation scenarios. The effectiveness of the proposed method was verified by experimental results.

1. INTRODUCTION

As AI and deep learning are both developing fast these days, significant progresses have been made in fields of computer vision and natural language processing [1-5], which has created a background for the emergence of MAC for combined processing of texts and images [6-12]. Conventional teaching evaluation emphasizes students' knowledge mastery over their affections [13-15], however, the state of affection has a very important influence on students' learning process and outcome. MAC can analyze versatile information of students in the classroom, including their facial expressions, gestures, and text feedback, in a comprehensive way, thereby helping teachers discover problems with students' affections in a timely manners, so that they could adjust the teaching methods and strategies accordingly [16-18]. The research results attained in this study can be used directly in actual teaching evaluation works, they could assist educators to understand students' learning requirements and affection state more comprehensively and accurately, and provide useful evidences for education policy makers, thereby improving the teaching effect, promoting rational education resource allocation, realizing education equity, and raising the overall education level.

Scholars Pham and Wang [19] proposed *AttentiveLearner2*, a multi-modal intelligent tutor running on unmodified smartphones, to supplement today's clickstream-based learning analytics for MOOCs (Massive Open Online Courses). The *AttentiveLearner2* uses both the front and back cameras of a smartphone as two complementary and fine-grained feedback channels in real time: the back camera monitors learners' photoplethysmography (PPG) signals and the front camera tracks their facial expressions during MOOC learning, and implicitly infers learners' affective and cognitive states during learning from their PPG signals and facial expressions. Barron-Estrada et al. [20] introduced the initial implementation of a multi-modal recognition system of affections using mobile devices and the creation of an affective database through a mobile application. The recognizer works into a mobile educational application to identify user's affections as they interact with the device. The affections recognized by the system are engagement and boredom. The affective database was created with spontaneous emotions of students who interacted with an educational mobile application called Duolingo and a mobile information gathering application called EmoData. Hu and Flaxman [21] proposed a novel approach of using deep neural networks combining with visual analysis and natural language processing to perform multi-modal sentiment analysis, their goal is different from the standard sentiment analysis goal of predicting whether a sentence expresses positive or negative sentiment, instead, they aim to infer the latent affection state of the user, focus on predicting the emotion word tags attached by users to their Tumblr posts, regarding these as "self-reported emotions". Yin et al. [22] pointed out that adopting deep learning methods to analyze multi-modal physiological signals and recognize human emotions has become more attractive these days, but the conventional deep emotion classifiers may suffer from the drawback of the lack of the expertise for determining model structure and the oversimplification of combining multi-modal feature abstractions. So they proposed a multiple-fusion-layer based

ensemble classifier of stacked auto-encoder for recognizing emotions, in which the deep structure was identified based on a physiological-data-driven approach.

After carefully reviewing existing studies, it's found that although MAC does have some potential in teaching evaluation, its defects and challenges are also apparent. The available MAC technology might make unstable or wrong judgement when dealing with complex affective expressions, then the inaccurate evaluation results of students' affection state might adversely affect the teaching evaluation effect. Moreover, the existing MAC technology is mostly based on deep learning models, which have a poor interpretability and may confuse educators in understanding and trusting the evaluation results given by the models. In view of these matters, this paper applied MAC to teaching evaluation based on combined processing of texts and images. In the second chapter, the input texts were divided into two parts: main body and the hash tag, which were subjected to feature extraction respectively. In the third chapter, image features were extracted from two angles: object and scene, since the two angles can give image information of different levels. In the fourth chapter, the MAC model was divided into modal sharing tasks and modal private tasks to attain a better adaptability in case of new teaching evaluation scenarios. The effectiveness of the proposed method was verified by experimental results.

2. TEXT FEATURE EXTRACTION

MAC could exert an important role in teaching evaluation since it integrates the processing of texts and images, detects and judges students' affection state based on the content of submitted text feedback and images of their facial expressions, provides personalized learning resources and suggestions according to their affection state and learning requirements, and figures out students' understanding and acceptance of course content by analyzing text answers and images uploaded by students in online tests or surveys, in addition, it could also evaluate teachers' performance and assist mental health specialists to analyze students' affection state. These applications are conductive to improving teaching quality, education equity, and the all-round development of students.

To apply MAC in teaching evaluation requires to collect these text and image data: 1) Student text data: including assignments and exam answers, text data from classroom discussions, forums, chat tools and other scenarios, and their feedback and comments about teachers, courses, and teaching methods, so as to understand their satisfaction and needs; 2) Student image data, including students' facial expressions in class, their postures and movements, and images they used in their paintings, designs, crafts, and other course works; 3) Teacher text data, including teaching plans and lecture notes, and text data drawn from interactions and communications between the teacher and students; 4) Teacher image data, including the teacher's facial expressions in the class, as well as the postures and movements.

In this study, the texts entered were divided into two parts: main body and hash tag. The text data of teaching evaluation usually has a high dimension, and the word fragment vector expression could convert the high-dimensional text information into the vector form with a lower dimension, thereby reducing computation complexity and improving the efficiency of model training and prediction. When turning texts into a vector form using word fragment-based algorithms, the semantic information of original texts could be well preserved, which is good for increasing the accuracy of MAC models in analyzing text features. The idea of this method is to convert texts into mathematical vectors, so as to support comparison and similarity calculation between texts, and this facilitates the recognition of texts with similar affection features in teaching evaluation scenarios and helps increasing the accuracy of affection calculation. Figure 1 shows the principle of word segmentation algorithm.

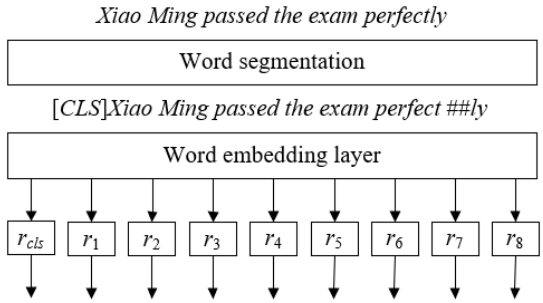

Figure 1. Principle of word segmentation algorithm

Assuming: u represents the u-th sample, $Y^q{}_u$ represents the input main body sequence of teaching evaluation texts, $Y^y{}_u$ represents the hash tag sequence, then the two pars of texts (main body and hash tag) can be written as $Y_u=\{Y^q{}_u,Y^y{}_u\}$; also, assuming q_k represents the k-th word in the main body sequence of teaching evaluation texts, l represents the length of main body sequence, then the input main body sequence of teaching evaluation texts can be written as $Y^q{}_u=\{q_1,q_2,...,q_k,...,q_l\}$; similarly, assuming y_k represents the k-th hash tag in the hash tag sequence, j represents the length of hash tag sequence, then the hash tag sequence can be written as $Y^y{}_u=\{y_1,y_2,...,y_k,...,y_j\}$.

The process of transforming teaching evaluation texts into vectors is described by the following formula:

$$G_0^y = d_{TE}\left(Y_u^q;\varphi_{TE}\right), G_0^y \in E^{l\times f} \tag{1}$$

After the input teaching evaluation texts were preprocessed, they were converted into word vectors in two steps. Assuming: q_k represents the input word, C represents the size of word list generated correspondingly, the value of C equals to the size of one-hot vector, $\{0,0,0,...,0,1,0,...,0\} \in E^C$ represents each word vector in one-hot encoding; if the dimension of the word vector is represented by f, then the weight initialization matrix between the input layer and the output layer can be characterized by a matrix Q_r with a size of $C\times f$.

$$Q_r = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ q_{C1} & q_{C2} & \cdots & q_{Cn} \end{pmatrix} \tag{2}$$

Hash tags (such as: #keyword) may contain important affection information in the teaching evaluation texts, namely keywords or phrases that are closely related to the content of the texts, which are helpful in more accurately analyzing the affection features of the texts. The encoder based on gate control mechanism can effectively extract these key

information and provide more valuable input for subsequent affection calculations. Moreover, keywords in hash tags may have similar or same semantics with words in the main body of the texts. By encoding and distinguishing the hash tags, the weight of these keywords in text vectors could be increased, so as to better capture the semantic information of the texts.

Assuming: y_k represents the k-th hash tag in a sample, j represents the number of hash tags, then the operation of removing the "#"symbol from each hash tag can be written as $Y^g_u=\{y_1...,y_k,...,y_j\}$; in the meantime, assuming G^g_0 represents the hidden state of the generated hash tag, d_{TE} represents the network for processing the texts, ϕ_{BM} represents the parameter to be learnt by the network, G^g_u represents the eigenvector of *tag*, d_{BM} represents the network for encoding hash tags, then the process of hash tag encoding can be expressed by the following formula:

$$G_0^g = d_{BM}\left(Y_u^g;\varphi_{BM}\right), g_0^g \in E^{l\times f} \tag{3}$$

Similar to the processing of main body of the texts, hash tags should be expressed in the vector form as well, assuming C_g represents the number of hash tags, $Q_g \in E^{Cg\times f}$ represents a a learnable matrix, then there is:

$$r_k^g = Q_g \cdot y_k \tag{4}$$

The initialization parameters of the neural network could be attained by uniformly sampling in the interval, let b_{SR} and b_{SC} respectively represent the number of neurons in the input layer and output layer, there is:

$$q \sim I\left(-\sqrt{\frac{6}{b_{SR}+b_{SC}}},\sqrt{\frac{6}{b_{SR}+b_{SC}}}\right) \tag{5}$$

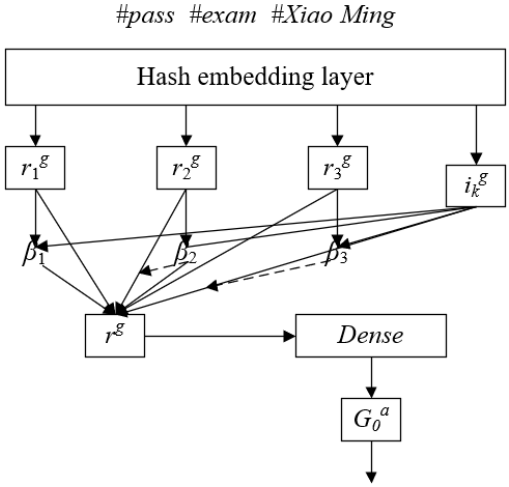

Figure 2. Principle of hash tag extraction algorithm

To apply MAC in teaching evaluation, a necessary work is to ensure which hash tags automatically learnt by the model during training are related to affections or not. In this way, the model can focus on hash tags containing affection information when dealing with teaching evaluation texts, then the accuracy of affection calculation could be improved. To implement the above algorithm, the soft gate control method was adopted to optimize the encoding method based on gate control network. Instead of simply classifying the tags into two categories (pass or not pass), the soft gate control method allows multiple hash tags to pass through with a certain probability, in this way, the relative weight between tags could be kept so that the model can better capture the differences between hash tags, and the effect of affection calculations could be enhanced. Figure 2 shows how the hash tag extraction algorithm works.

By using a vector $i_v \in E^f$ to represent the global context representation of the teaching evaluation texts, the calculation of similarity between the current y_k and the corresponding i_v can be completed through the gate control mechanism, based on which the probability of y_k passing the gate control mechanism can be judged. Specifically, assuming r^g represents the vector form of y_k, which was subjected to non-linear transformation and its hidden expression was learnt; assuming i^g_k represents the hidden expression of the k-th hash tag, then the transformation matrix and bias term can be represented by $Q_{gg} \in E^{f\times f}$ and $y^y_g \in E^f$ respectively, and this process can be described by the following formula:

$$i_k^g = \mathrm{Tanh}\left(Q_{gg}\cdot r_k^g + n_g^y\right) \tag{6}$$

The *sigmoid* function was selected to calculate the probability β_k of each hash tag being selected by the gate control mechanism:

$$\beta_k = \frac{1}{2+e^{-i_k^y\cdot i_v}} \tag{7}$$

Then, get the weighted sum of the hidden expressions of all hash tags:

$$r^g = \sum_{k=1}^{j}\beta_k \cdot r_k^g \tag{8}$$

At last, the expressions of hash tags were subjected to a linear transformation to merge the information of all modes in the same semantic space. Assuming: $G^g_0 \in E^f$ represents the final output hash tag, $Q_{gg} \in E^{f\times f}$ represents the transformation matrix, then there is:

$$G_0^g = Q_{gg}\cdot r^g \tag{9}$$

3. IMAGE FEATURE EXTRACTION

To get more comprehensive affection information, in the task of applying MAC in teaching evaluation, in this study, image features were extracted from two angles: scene and object, since the two angles could provide image information of different levels. Object features focus on specific objects in the image and the relationships between them, this is helpful to identifying objects and events in the image, while scene features reveal the overall background and environment of the image, and can provide information about the background and scenarios. Object features and scene features can complement each other, giving richer semantic representations of the image. In case of teaching evaluation, the image data may contain multiple objects and scenes, by considering features of the two aspects, the generalization ability of the model in processing new image data could be strengthened. Figure 3 shows the principle of image feature extraction algorithm.

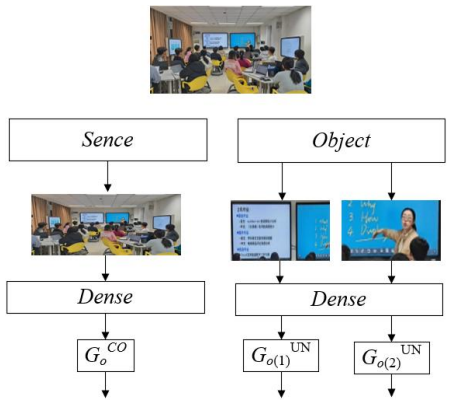

Figure 3. Principle of image feature extraction algorithm

Local object features of teaching evaluation scenario images could be attained by target detection algorithm. Assuming: $G^p{}_0$ represents object-oriented image features, x represents the number of objects, f represents the dimension of features, d_{PB} represents the target detection network, ϕ_{PB} represents the parameter to be learnt by the network, then this process can be expressed as:

$$G_0^p = d_{PB}\left(L;\varphi_{PB}\right), G_0^p \in E^{x\times f} \tag{10}$$

The local object image area of the final teaching evaluation scenario can be written as $R^p=\{r^p{}_1,r^p{}_2,...,r^p{}_x\}$, assuming $G^p{}_0 \in E^{x\times f}$ represents the last output local object features of teaching evaluation scenario, $Q_{gp} \in E^{f\times f}$ represents the transformation matrix, then the non-linear transformation from image features to the semantic space of texts is given by the following formula:

$$G_0^p = Q_{gp} \cdot r^p \tag{11}$$

Assuming: $G^a{}_0$ represents the scene-oriented image features, d_{SC} represents the network for extracting the scenes, ϕ_{SC} represents the parameter to be learnt by the network, then the extraction process of global scene features is given by the following formula:

$$G_0^a = d_{SC}\left(L;\varphi_{SC}\right), G_0^a \in E^f \tag{12}$$

Assuming: $G^a{}_0 \in E^f$ represents finally output object-oriented image features, $Q_{ga} \in E^{f\times f}$ represents the transformation matrix, then the non-linear transformation from image features to the semantic space of texts is given by the following formula:

$$G_0^a = Q_{ga} \cdot r^a \tag{13}$$

4. A MAC METHOD COMBINING TEXT AND IMAGE PROCESSING

After feature extraction of the collected text and image data was completed, to further realize the application of MAC in teaching evaluation based on combined processing and texts and images, the MAC model was divided into modal sharing tasks and modal private tasks to attain a better adaptability in case of new teaching evaluation scenarios. The modal sharing tasks can integrate information between different modes, and extract cross-modal sharing features, which is conductive for the model to better capture the relationship between texts and images, and improve the accuracy of affection calculation. Modal private tasks focus on the internal features of each mode, and could retain the unique information inside the modes, in this way, when the model merges the multi-modal information, it won't lose the features of each mode, and it could better reflect the affection information in teaching evaluation scenarios. Via such task division operation, the generalization ability of the model could be enhanced. This is because modal sharing tasks can learn common features across modes, while modal private tasks focus on the features of each mode. This design enables the model to have a higher adaptability when coping with new teaching evaluation scenarios.

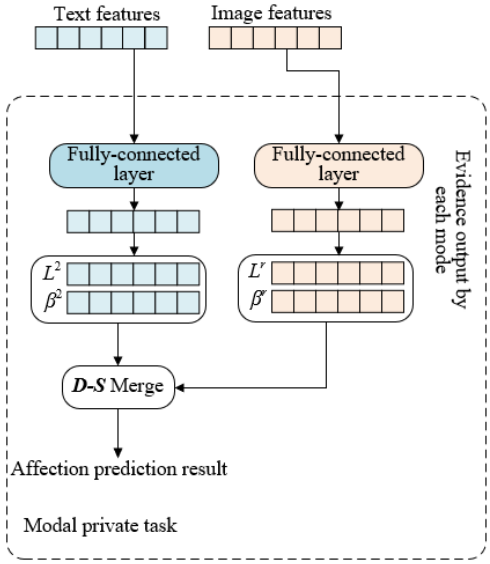

Figure 4. Principle of modal private task

The key to realizing modal private tasks (Figure 4) lies in the prediction and analysis of each mode, in this way, the unique information insides each mode could be dug out, namely the Dirichlet distribution parameters and the subjective opinions. This helps to prevent the loss of modal features when merging multi-modal information, thereby better displaying the affection information in teaching evaluation scenarios. After that, prediction results of each mode were merged according to the D-S Evidential Theory. Since Dirichlet distribution parameters and subjective opinions can provide an uncertainty matric for the prediction results of each mode, so we need to apply the D-S Evidential Theory to effectively processing those incomplete, uncertain, and conflicting information, so that the model can merge the information flexibly between modes, thereby improving the adaptability and accuracy of the model in teaching evaluation scenarios.

Assuming: L^v and L^c represent the output subjective opinions of each mode, β^v and β^c represent the Dirichlet distribution parameters, B represents the sample number, J represents the sample type number, $\hat{t}_v = [\hat{o}_1, \hat{o}_2, \ldots, \hat{o}_J]$ represents the final output prediction results of each mode merged based on the D-S Evidential Theory, i represents the uncertainty estimate, then the loss function can be expressed as:

$$Loss_{UN} = -\sum_{u=1}^{B}\sum_{k=1}^{J} t_{uk} \log\left(\hat{o}_{uk}\right) \tag{14}$$

The key to realizing modal sharing tasks (Figure 5) is to construct a feature fusion network based on tensor decomposition and use it to effectively fuse the features of different modes, and integrate the multi-modal information of texts and images into a unified feature space. This conduces to enhancing the ability of the model in capturing affection information in teaching evaluation scenarios. With the help of tensor decomposition, the computation complexity can be reduced and the running efficiency of the model can be improved during the process of feature fusion. By performing low-rank approximation on multi-modal features, the dimension of fused features could be reduced greatly, so the computation load of the model could be reduced as well. At last, the multi-layer perception (MLP) was adopted for prediction classification based on the feature fusion results. MLP can effectively fuse the multi-modal features, decrease computation complexity, capture high-level interactive information, realize end-to-end training, and has good flexibility and extensibility, by adopting MLP, more valuable model designs could be provided for the application of MAC in teaching evaluation scenarios.

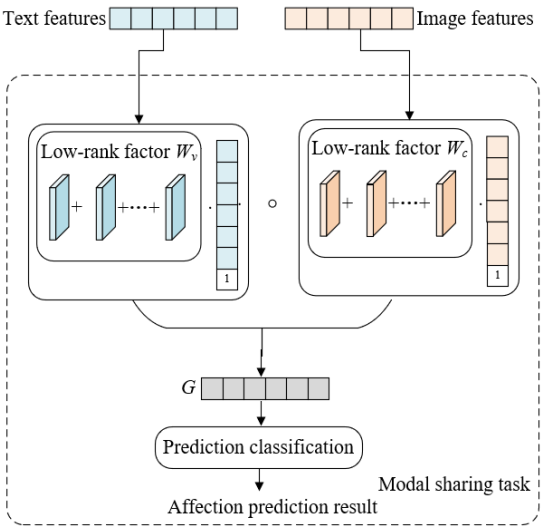

Figure 5. Principle of modal sharing task

Assuming: g_y and g_c represent eigenvectors of texts and images after going through modal representations; after fusion, the feature tensor can be represented by $X=x_y \otimes x_c$, the product of tensors can be represented by $\otimes$, dimension of the eigenvectors can be represented by f_y and f_c, and there is $X \in E^{f_y \times f_c}$, then the eigenvectors of each model can be expanded by one dimension using the above formulas:

$$\begin{aligned} x_y &= \Gamma\left(G_y, 1\right), x_y \in E^{f_y} \\ x_c &= \Gamma\left(G_C, 1\right), x_c \in E^{f_c} \end{aligned} \tag{15}$$

Assuming: Q represents the weight of linear layer, n represents the bias, X represents the third-order tensor, $Q \in E^{f_y \times f_c \times fG}$ represents the fourth-order tensor, the extra dimension is the dimension of output vector f_G, g is output through linear layer $h(")$, then there is:

$$G = h\left(X; Q, n\right) = Q \cdot X + n, g, n \in E^{f_G} \tag{16}$$

Q can be regarded as a superposition of f_G second-order tensor $\tilde{Q}_j \in E^{f_y \times f_c}$. For any $\tilde{Q}_j \in E^{f_y \times f_s \times f_c}$, it can be decomposed with rank E as follows:

$$\tilde{Q}_j = \sum_{u=1}^{E} q_{y,j}^{(u)} \otimes q_{c,j}^{(u)} \tag{17}$$

Assuming: $q^{(u)}_{y,j} \in E^{f_y}$ and $q^{(u)}_{c,j} \in E^{f_c}$ represent the decomposition factors of tensor $\tilde{Q}_j$ with a modal rank of E corresponding to texts and images, here E was set as a constant, the E-rank decomposition factor $\{q^{(u)}_{y,j}, q^{(u)}_{c,j}\}^{E}_{u=1}, j=1,...f_G$ can be used to re-construct tensor $\tilde{Q}_j$. Let $q^{(u)}_y=[q^{(u)}_{y,1}, q^{(u)}_{y,2}, ...q^{(u)}_{y,fG}]$, $q^{(u)}_s=[q^{(u)}_{s,1}, q^{(u)}_{s,2}, ...q^{(u)}_{s,fG}]$, $q^{(u)}_c=[q^{(u)}_{c,1}, q^{(u)}_{c,2}, ...q^{(u)}_{c,fG}]$, then the weight Q in Formula 16 can be represented by the following formula:

$$Q = \sum_{u=1}^{E} q_y^{(u)} \otimes q_s^{(u)} \otimes q_c^{(u)} \tag{18}$$

Further, Formula 16 can be written as:

$$\begin{aligned} g &= \left(\sum_{u=1}^{E} q_{y,j}^{(u)} \otimes q_{s,j}^{(u)} \otimes q_{c,j}^{(u)} \right) \cdot X = \sum_{u=1}^{E} \left(q_{y,j}^{(u)} \otimes q_{s,j}^{(u)} \otimes q_{c,j}^{(u)} \cdot X \right) \\ &= \sum_{u=1}^{E} \left(q_{y,j}^{(u)} \otimes q_{s,j}^{(u)} \otimes q_{c,j}^{(u)} \cdot x_y \otimes x_s \otimes x_c \right) \\ &= \left(\sum_{u=1}^{E} q_y^{(u)} \cdot x_y \right) \otimes \left(\sum_{u=1}^{E} q_s^{(u)} \cdot x_s \right) \otimes \left(\sum_{u=1}^{E} q_c^{(u)} \cdot x_c \right) \end{aligned} \tag{19}$$

Let $Q_v \in E^{fG \times 1}$, the bias is represented by n_v, the affection prediction result is represented by $\hat{t}_v$, at last, the fully-connected layer was used to perform multi-modal affection prediction, and the output result can be expressed as:

$$\hat{t}_v = Q_v^Y G + n_v \tag{20}$$

The loss function of modal sharing tasks can be expressed as:

$$Loss_{CO} = \frac{1}{B} \sum_{u=1}^{B} \left(\hat{t}_u - t_u \right)^2 \tag{21}$$

Assuming: β represents a hyperparameter, then the overall loss function of the MAC model based on combined processing of texts and images can be expressed as:

$$Loss_{MA} = Loss_{CO} + \beta Loss_{UN} \tag{22}$$

5. EXPERIMENTAL RESULTS AND ANALYSIS

Figure 6 gives changes in the affection classification accuracy corresponding to different key area numbers, showing the relationship between key area number and affection classification accuracy. For the given six key area numbers (1, 2, 3, 4, 5, 6), the corresponding affection classification accuracy is 77%, 77.5%, 81.9%, 82.7%, 83.1%, and 83.4%, respectively. Analysis of table data shows that with the increase of key area number, the affection classification accuracy shows an upward trend, indicating more key areas can help the model better capture affection information and improve affection classification accuracy.

When the number of key areas is increased from 1 to 2, the increment of affection classification accuracy is quite limited (rising from 77% to 77.5%), indicating that under certain conditions, a small increase of key area number may not significantly improve affection classification accuracy. But when the number of key areas reaches 3, the improvement of affection classification accuracy becomes obvious. Especially when the number of key areas increases from 3 to 4, the affection classification accuracy rises from 81% to 82.7%, showing a significant improvement, and this indicates that in the MAC task, increasing the number of key areas appropriately has a positive effect on improving the performance of the model. As the number of key areas continues to increase from 5 and 6, the affection classification accuracy is still on the rise, but the increment is smaller, suggesting that after a certain threshold, continuing to increase the number of key areas only has a limited effect on the performance of the model.

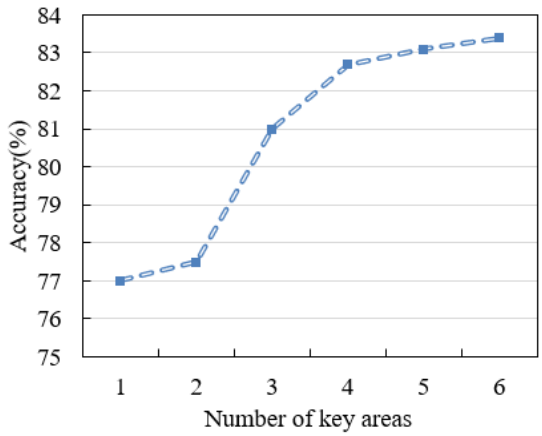

Figure 6. Affection classification accuracy corresponding to different key area numbers

Table 1. Experimental results of different models

Model	Precision (%)	Recall (%)	F1(%)	Accuracy (%)
Basic model	72.16	70.63	71.96	74.02
Basic model + text feature extraction	74.93	71.95	73.48	71.35
Basic model + image feature extraction	78.15	85.16	82.51	85.19
The proposed model	82.49	84.31	83.26	83.42

Table 1 shows the experimental results of different models. The performance of four different models in four aspects of *Precision*, *Recall*, *F1*-value, and *Accuracy* was given. Models participated in the experiment include the basic model, basic model + text feature extraction, basic model + image feature extraction, and the proposed model. During our experiment, *ResNet* 50 was chosen as the basic model, which only performed simple text and image processing and did not involve advanced feature extraction or MAC. According to the data given in the table, the performance of the basic model was the lowest among all models, which has proved the necessity of advanced feature extraction and MAC methods. In terms of basic model + text feature extraction, after text feature extraction was added based on the basic model, precision, recall, and F1 value all showed an improvement, but the accuracy dropped slightly. This may be because text feature extraction helped the model to better understand text information, but due to the lack of support for image information, the model made some misjudgments in some cases. In terms of basic model + image feature extraction, after image feature extraction was added based on the basic model, the performance of the model was enhanced in all indicators, and this has proved that image feature extraction has a great effect on the affection calculation in teaching evaluation scenarios, and the model can well capture the affection information in images. The proposed model combined text feature extraction, image feature extraction, and MAC methods. Judging based on the data in the table, the proposed model outperformed others in terms of all indicators, which has verified the validity of the proposed method in teaching evaluation. The model can accurately classify affections by comprehensively analyzing text and image information.

Figure 7 uses histogram to show the experimental results of different models, as can be known from the figure, the proposed model achieved a higher precision than other models, which means that the proposed model can more precisely recognize the affection type, and reduce the possibility of misjudgment.

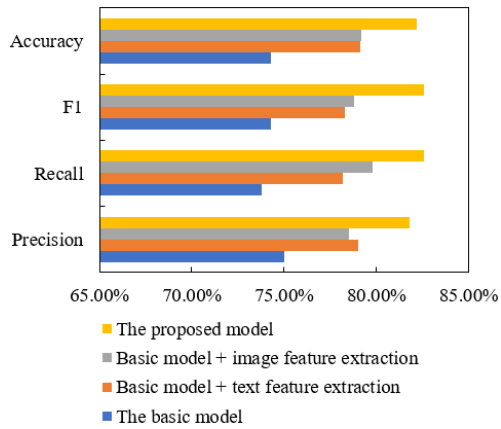

Figure 7. Experimental results of different models

Compared with other models, the recall rate of the proposed model showed a significant improvement, indicating that the proposed model can recognize samples containing certain affections more comprehensively, and the coverage range of the model has been increased. F1-value is the harmonic mean of precision and recall, it is used to evaluate the overall performance of the model. The proposed model outperformed others in terms of F1, indicating it can give a higher performance with both precision and recall taken into account. In terms of accuracy, the proposed model also performed the best among the four models, proving its high accuracy of overall affection classification in teaching evaluation scenarios, the proposed model could well predict the actual affection types, and provide valuable references for teaching evaluation. In summary, analysis from the aspects of multiple indicators shows that, the method of integrating text feature extraction, image feature extraction and MAC proposed in this paper has been proved to be effective in teaching evaluation, compared with other models, it exhibited significance improvements in terms of all indicators, which has also demonstrated its great potential of being applied in teaching

evaluation. Moreover, it offers a new and more effective solution for the teaching evaluation field to better understand students' affection requirements and improve education quality.

Figure 8 shows the experimental results under conditions of 50%, 75%, and 100% data volume. As can be seen from the figure, with the increase of the amount of training data, the accuracy of all models increases as well, but the proposed model outperformed others in terms of accuracy under all data volume conditions, which has demonstrated that the proposed MAC method was significantly effective in teaching evaluation scenarios, and this is mainly due to the MAC technology combining with text and image processing, which enables to model to make use of both text and image information at the same time, so the accuracy of affection classification has been improved. As the data volume rises, the accuracy of all models increases, indicating that more training data is helpful for the models to learn richer and more accurate information, and the prediction ability of the model has been enhanced. By comparing the results of basic model, basic model + text feature extraction and basic model + image feature extraction, it can be found that adding text feature extraction or image feature extraction alone can improve model accuracy to a certain extent, but their effects are not as significant as the overall combined effect of the proposed model, and this has again verified the advantages of combining MAC with text and image processing in teaching evaluation scenarios.

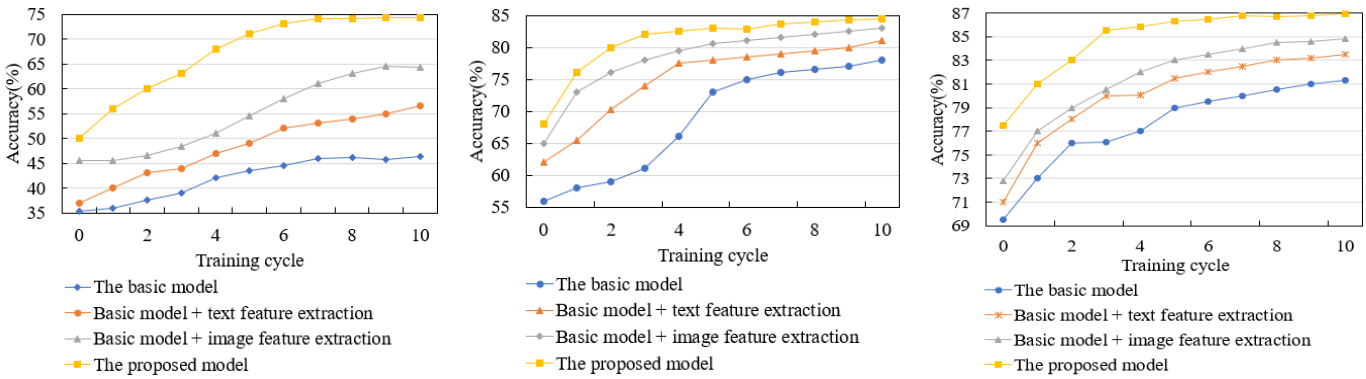

Figure 8. Experimental results in case of 50%, 75%, and 100% data volume

Table 2. Experimental results of MAC datasets (%)

No.	Mode	Modal	Standard Sample Set 1		Standard Sample Set 2		Standard Sample Set 3	
			Acc	*F*1	*Acc*	*F*1	*Acc*	*F*1
1	*Text*	*BERT*	63.58	52.36	69.35	52.48	52.37	52.43
2		*GRU*	61.24	68.42	61.52	69.16	65.42	59.16
3		The proposed method	69.35	60.51	69.47	64.82	69.31	57.35
4	*Image*	Object-oriented	64.52	55.82	60.51	53.62	42.18	42.05
5		Scene-oriented	53.16	59.27	69.34	59.46	46.53	46.15
6		The proposed method	64.15	68.31	61.42	64.81	59.18	49.38
7	*Text+Image*	*ICCN*	68.53	61.81	63.81	69.76	63.18	52.58
8		*MISA*	74.15	60.18	69.52	60.85	64.32	50.16
9		The proposed method	79.52	72.36	63.84	69.15	62.08	57.49

Table 2 gives the performance of different modes and models on three standard sample sets. Following conclusions can be drawn from the analysis of these results. Under text mode, the accuracy and F1 scores of the proposed method (No. 3) were better than those of BERT (No. 1) and GRU (No.2), indicating that the text feature extraction method proposed in this paper can better capture affection information, so the accuracy of affection classification has been improved. Under image mode, the proposed method (No. 6), the object-oriented method (No. 4) and the scene-oriented method (No.5) performed good in terms of accuracy and F1 value, indicating that the image feature extraction method adopted in this paper can more comprehensively capture affection information, and the prediction performance of the model has been enhanced. Under the combined mode of texts and images, the accuracy and F1 value scores of the proposed method (No. 6) were higher than those of *ICCN* (Interaction Canonical Correlation Network, No.7) and *MISA* (Modality Invariant and Specific Representations, No. 8), indicating that by considering text and image information simultaneously, the proposed MAC method can make full use of these information, and the accuracy of affection classification has been enhanced.

To sum up, the proposed model was proved to be highly effective in the experimental results of MAC data sets, thanks to the ability of the proposed text extraction method in well capturing affection information, the affection classification accuracy under text mode has been enhanced. The image feature extraction method adopted in this paper can attain affection information more comprehensively from two aspects of object and scene, so that the prediction performance under image mode could be enhanced. In addition, the MAC method proposed in this paper can make full use of text and image information, and further improve the accuracy of affection classification by merging the two.

6. CONCLUSION

This study explored the application of MAC in teaching evaluation combining text and image processing. At first, the

input texts were divided into two parts: main body and the hash tag, which were subjected to feature extraction respectively. The image features were extracted from two angles: object and scene, since the two angles can give image information of different levels. The MAC model was divided into modal sharing tasks and modal private tasks to attain better adaptability in case of new teaching evaluation scenarios. In the experimental results, the changes in affection classification accuracy corresponding to different numbers of key areas were given, the performance of four models was analyzed based on four indicators, *Precision*, *Recall*, *F1* and *Accuracy*, and the results verified the validity of the proposed method in teaching evaluation. Through comprehensive analysis of text and image information, the model can more accurately classify affections. The experimental results under the conditions of 50%, 75% and 100% data volume were given, which again verified the advantages of the proposed MAC method combining text and image process in teaching evaluation. At last, the experimental results of different models under different modes on MAC datasets were given, and the reasons why the proposed model has exhibited high validity on the datasets were summarized.

FUNDING

This paper was supported by Humanities and Social Science Research Project of Hebei Education Department (Grant No.: SD2022041).

REFERENCES

[1] Singh, G.V., Firdaus, M., Ekbal, A., Bhattacharyya, P. (2022). EmoInt-Trans: A multimodal transformer for identifying emotions and intents in social conversations. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31: 290-300. https://doi.org/10.1109/TASLP.2022.3224287

[2] Samsonovich, A.V., Liu, Z., Liu, T.T. (2023). On the possibility of regulation of human emotions via multimodal social interaction with an embodied agent controlled by ebica-based emotional interaction model. In Artificial General Intelligence: 15th International Conference, AGI 2022, Seattle, WA, USA, pp. 374-383. https://doi.org/10.1007/978-3-031-19907-3_36

[3] Kalatzis, A., Girishan Prabhu, V., Rahman, S., Wittie, M., Stanley, L. (2022). Emotions matter: Towards personalizing human-system interactions using a two-layer multimodal approach. In Proceedings of the 2022 International Conference on Multimodal Interaction, Bengaluru, India, pp. 63-72. https://doi.org/10.1145/3536221.3556582

[4] Samyoun, S., Mondol, A.S., Stankovic, J. (2022). A multimodal framework for robustly distinguishing among similar emotions using wearable sensors. In 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, pp. 4668-4671. https://doi.org/10.1109/EMBC48229.2022.9871229

[5] Caldeira, F., Lourenço, J., Tavares Silva, N., Chambel, T. (2022). Towards multimodal search and visualization of movies based on emotions. In ACM International Conference on Interactive Media Experiences, Aveiro JB, Portugal, pp. 349-356. https://doi.org/10.1145/3505284.3532987

[6] Teo, C. L., Ong, A.K.K., Lee, A.V.Y. (2022). Exploring students' epistemic emotions in knowledge building using multimodal data. In Proceedings of the 15th Computer-Supported Collaborative Learning, Hiroshima, Japan, pp. 266-273.

[7] Chen, W., Wu, G. (2022). A multimodal convolutional neural network model for the analysis of music genre on children's emotions influence intelligence. Computational Intelligence and Neuroscience, 2022: 5611456. https://doi.org/10.1155/2022/5611456

[8] Muszynski, M., Tian, L., Lai, C., et al. (2019). Recognizing induced emotions of movie audiences from multimodal information. IEEE Transactions on Affective Computing, 12(1): 36-52. https://doi.org/10.1109/TAFFC.2019.2902091

[9] Catharin, L.G., Ribeiro, R.P., Silla, C.N., Costa, Y.M., & Feltrim, V.D. (2020). Multimodal classification of emotions in latin music. In 2020 IEEE International Symposium on Multimedia (ISM), Naples, Italy, pp. 173-180. https://doi.org/10.1109/ISM.2020.00038

[10] Bollini, L., Fazia, I. D. (2020). Situated emotions. The role of the soundscape in a geo-based multimodal application in the field of cultural heritage. In Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part III 20, Cagliari, Italy,pp. 805-819. https://doi.org/10.1007/978-3-030-58808-3_58

[11] Ordonez-Bolanos, O.A., Gomez-Lara, J.F., Becerra, M.A., Peluffo-Ordóñez, D.H., Duque-Mejia, C.M., Medrano-David, D., Mejia-Arboleda, C. (2019). Recognition of emotions using ICEEMD-based characterization of multimodal physiological signals. In 2019 IEEE 10th Latin American Symposium on Circuits & Systems (LASCAS), Armenia, Colombia, pp. 113-116. https://doi.org/10.1109/LASCAS.2019.8667585

[12] Zheng, W.L., Liu, W., Lu, Y., Lu, B.L., Cichocki, A. (2018). Emotionmeter: A multimodal framework for recognizing human emotions. IEEE Transactions on Cybernetics, 49(3): 1110-1122. https://doi.org/10.1109/TCYB.2018.2797176

[13] Villegas-Ch, W., García-Ortiz, J., Sánchez-Viteri, S. (2023). Identification of emotions from facial gestures in a teaching environment with the use of machine learning techniques. IEEE Access, 11: 38010-38022. https://doi.org/10.1109/ACCESS.2023.3267007

[14] Hartikainen, S., Pylväs, L., Nokelainen, P. (2022). Engineering students' perceptions of teaching: Teacher-created atmosphere and teaching procedures as triggers of student emotions. European Journal of Engineering Education, 47(5): 814-832. https://doi.org/10.1080/03043797.2022.2034750

[15] Sun, Y. (2022). Changing positive academic emotions of art students utilizing computer information technology based on the perspective of teaching. Applied Bionics and Biomechanics, 2022: 7184274. https://doi.org/10.1155/2022/7184274

[16] Rehmat, A.P., Diefes-Dux, H.A., Panther, G. (2021). Engineering instructors' self-reported emotions during emergency remote teaching. In 2021 IEEE Frontiers in Education Conference (FIE), Lincoln, NE, USA, pp. 1-6. https://doi.org/10.1109/FIE49875.2021.9637440

[17] Ramos-Aguiar, L.R.,Álvarez-Rodríguez, F.J. (2021). Teaching emotions in children with autism spectrum disorder through a computer program with tangible interfaces. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 16(4): 365-371. https://doi.org/10.1109/RITA.2021.3125901

[18] Liang, Y.C., Lin, K.H.C., Li, C.T. (2021). Employing STEAM 6E teaching methods to analyze the academic emotions of the digital video practice course. In Innovative Technologies and Learning: 4th International Conference, ICITL 2021, Virtual Event, pp. 584-592. https://doi.org/10.1007/978-3-030-91540-7_60

[19] Pham, P., Wang, J. (2018). Predicting learners' emotions in mobile MOOC learning via a multimodal intelligent tutor. In Intelligent Tutoring Systems: 14th International Conference, ITS 2018, Montreal, QC, Canada, pp. 150-159. https://doi.org/10.1007/978-3-319-91464-0_15

[20] Barron-Estrada, M.L., Zatarain-Cabada, R., Aispuro-Gallegos, C.G. (2018). Multimodal recognition of emotions with application to mobile learning. In 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), Mumbai, India, pp. 416-418. https://doi.org/10.1109/ICALT.2018.00104

[21] Hu, A., Flaxman, S. (2018). Multimodal sentiment analysis to explore the structure of emotions. In proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining, London, United Kingdom, pp. 350-358. https://doi.org/10.1145/3219819.3219853

[22] Yin, Z., Zhao, M., Wang, Y., Yang, J., Zhang, J. (2017). Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. Computer Methods and Programs in Biomedicine, 140: 93-110. https://doi.org/10.1016/j.cmpb.2016.12.005