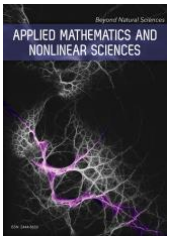
sciendo Applied Mathematics and Nonlinear Sciences, 9(1) (2024) 1-14



Applied Mathematics and Nonlinear Sciences

https://www.sciendo.com

The Construction of Cultural Identity in Multimodal Discourse Analysis under the Perspective of Media Convergence

Junfang Chang[1,†]

1. School of Foreign Languages, Huanghe Science and Technology University, Zhengzhou, Henan, 450000, China.

Submission Info

Communicated by Z. Sabir
Received November 15, 2022
Accepted April 14, 2023
Available online October 4, 2023

Abstract

This paper constructs a model of cultural identity by combining the analysis of multimodal discourse in the media fusion perspective and explores how to better construct cultural identity. Firstly, we construct a multimodal cultural identity model in the media convergence perspective from three modalities: text, image, and video combined with cultural identity, so that we can analyze people's cultural identity through emotion. Secondly, multimodal cross-sectional comparison experiments and dissimilarity experiments are conducted based on two widely used public datasets in the field, CMU-MOSI and CMU-MOSEI. The best experimental results were obtained in the baseline model, where Acc-2 reached 84.8%, F1 value reached 84.5, and MAE of the regression task dropped to 0.548. In the robustness study, when the missing rate reached 44%, Acc-2 was only 65.2%, and the gap between the results and the dichotomous classification accuracy of 76.5% achieved by the model in this paper reached 11.3 percentage points. Finally, through affective analysis, cultural identity is the affirmative embodiment of what is most meaningful to the nation formed by people living together in a national community for a long time, and its core is the identification with the basic values of a nation, which is the spiritual bond that unites this national community.

Keywords: Media convergence; Multimodal discourse; Comparison experiment; Dissolution experiment; Robustness.
AMS 2010 codes: 01A13

†Corresponding author.
Email address: jizhao107687008@163.com

ISSN 2444-8656
https://doi.org/10.2478/amns.2023.2.00478

1 Introduction

The essence of cultural identity belongs to the concept of caste, "the identification of all members of a people with the most basic and meaningful physical objects and values of the people [1-2]. Cultural identity is the subject's conceptual, psychological and behavioral recognition and acceptance of a particular culture. The most fundamental is the identification of the basic values of a certain nation [3]. Cultural identity is a community culture based on the idea and value of identity, a spiritual bond that allows the team to coalesce and develop, as well as the subject's identification with the basic values of the nation [4-5]. Cultural identity is a dynamic process in which the subject recognizes and acknowledges a certain cultural value and also a certain cognitive psychology from it [6-7]. Regarding the research on the strategic position of cultural identity, it is generally accepted in the academic community that cultural identity is both the foundation of the common spiritual home of the nation and the basis of national identity and national identity, as well as a strategic choice of the nation in the face of the globalization of culture, etc. [8-10]. There are also more and more scholars who analyze multimodal discourse from the media integration perspective to construct cultural identity and reduce internal spiritual conflict [11].

The perceived sense of biculturalism in Australia is examined in the literature [12], which explores the differences in cultural pluralism through the perceived specification of cultural pluralism and the assimilation of cultural pluralism. The literature [13] proposes that schools are an important environment for fostering cultural identity and explores how students' cultural identity can flourish through the means by which schools deal with different cultural differences. Also, the literature [14] proposes that culturally diverse schools are beneficial for adolescent development and explores the relationship between culture and groups by measuring the range of cultural diversity.

Also, a more precise ecological niche of culture as a species origin was constructed in the literature [15] by collecting cultural competence in five cultural domains to construct a cultural model. In the literature [16], the relationship between children's cultural identity and parents' cultural upbringing was analyzed in different countries, and it was found that children in different countries differed significantly in their expression of cultural emotions. The literature [17] argues that integrating cultural identity into learning is a great challenge, that the cultural background of scholars is important, and that teachers study the meaning of students' learning in terms of their cultural identity, thus improving learning outcomes.

In this paper, in order to be able to better analyze cultural identity, multimodality is incorporated in three domains of text, image and video in the media perspective, and a multimodal cultural identity model in the media convergence perspective is constructed in conjunction with cultural identity. The focus is on multimodal cross-sectional comparison experiments based on two widely used public datasets in the field, CMU-MOSI and CMU-MOSEI. The model in this paper achieved a performance improvement of 4 and 10.6 percentage points in the CMU-MOSI dataset for the sentiment dichotomy task, with Acc taking 36 and a reduction in MAE of up to 0.212 in the regression task. The best experimental results were achieved in the MU-MOSEI dataset with Acc-2 reaching 84.8% and F1 value reaching 84.5, and the regression task with MAE also decreased to 0.548, with different magnitudes of performance improvement. Secondly, in the abasement experiments, the gap between the effect of 76.5% of the dichotomous classification accuracy achieved by the model of this paper in the robustness study was obtained to reach 11.3 percentage points. Finally, the feasibility of constructing cultural identity under multimodal utterance analysis from the media perspective is demonstrated by sentiment analysis.

2 Multimodal discourse analysis cultural identity model in the context of media convergence

Multimodal discourse analysis interprets the representational meanings of text, images, and videos from multiple perspectives and explores the synergy and interaction between the various modalities in expressing meaning. The multimodal discourse application of mainstream media for cultural identity construction relies on the existing media matrix, based on the current reading habits of viewers, and uses a "re-mediated" practice - smartphone - to build an interactive space with viewers. The module consists of three parts, namely text embedding, image embedding and video embedding.

2.1 Text embedding

2.1.1 Relationship embedding

Text embedding is divided into relational embedding and attribute embedding. The medium is combined with multimodal utterance analysis by means of text. Given a relational triple (e_h, rel, e_t), where e_h is the head entity, e_t is the tail entity, and rel is the relationship between the head and tail entities. Suppose h is the feature embedding vector of the head entity e_h, r is the feature embedding vector of the relation rel, and t is the relation embedding vector of the tail entity e_t. The basic idea of TransE is that if a triple (e_h, rel, e_t) is a true triple, then the corresponding vector in the vector space needs to match $h + r \approx t$. The scoring function of TransE is defined as shown in Eq:

$$S_{rel}(e_h, rel, e_t) = \|h + r - t\|_2^2 \tag{1}$$

where $\|\cdot\|_2$ is the L_2-parameter number. In this paper, a marginal-based loss function is used to distinguish between positive and negative cases in order to expect the lower value of S_{rel} to be better. The marginal-based loss function is shown in Eq:

$$L_{rel}(h,r,t) = \sum_{(e_h, rel, e_t) \in X} \sum_{(e_h', rel, e_t') \in X'} \left[\gamma + S_{rel}(e_h, rel, e_t) - S_{rel}(e_{h'}', rel, e_t')\right]_+ \tag{2}$$

Where γ is the hyperparameter, $[x]_+ = \max(0, x)$ denotes the maximum value between 0 and x, and X and X' denote the set of positive examples and the set of negative examples, respectively. It is worth noting that in this paper, the set of negative examples is generated by randomly replacing the head entity, tail entity or relation of the triple in the set of positive examples, i.e.:

$$X' = \left\{(e_h', rel, e_t) \mid e_h' \in E\right\} \cup \left\{(e_h, rel, e_t') \mid e_t' \in E\right\} \cup \left\{(e_h, rel', e_t) \mid rel' \in R\right\} \tag{3}$$

2.1.2 Attribute embedding

The attribute triad is denoted as (e, a, n), where e denotes an entity, a denotes an attribute and n denotes an entity e corresponding to the attribute value of a certain attribute. In this paper, we use e to denote the feature embedding vector of the entity e, a to denote the feature embedding vector of the attribute a, and n to denote the feature embedding vector of the attribute value n to describe a specific entity by attribute and corresponding attribute value. Unlike relational triples,

attribute triples are difficult to be represented by the form of relational triples, but with the help of attribute values in attribute triples, the information between entities can be supplemented to achieve the goal of improving the entity alignment performance of knowledge graphs. Since attributes contain more numerical data, the Gaussian kernel function can be used to dimensionalize the data so that the original linearly indistinguishable data becomes linearly distinguishable, and the specific application of the Gaussian kernel function is shown in Eq:

$$f\left(\mu_{(e,a)}\right)=\exp\left(\frac{-\left\|\mu_{(e,a)}-c_i\right\|^2}{\sigma_i^2}\right) \tag{4}$$

where $\mu_{(e,a)}$ represents a specific attribute value of the entity, c_i represents the center of the kernel function, and σ_i is a bandwidth parameter to control the radial range of the function. The specific numerical information is transferred to the above Gaussian kernel function to obtain the numerical vector n. The obtained numerical vector n is stitched with the embedding attribute vector a, and assuming that each vector has dimension m, a new matrix $C=[a,n]^T$ of $2*m$ is generated after stitching. C' is obtained after convolving the matrix C and passing through a fully connected layer, C' is represented as shown in Eq:

$$C'=\tanh(\mathit{flat}(\mathit{Conv}(\tanh(C)))D) \tag{5}$$

Where $\mathit{flat}(\cdot)$ is the vector flattening operation, $\mathit{Conv}(\cdot)$ denotes the vector convolution operation, $\tanh(\cdot)$ is the activation function, and D denotes the fully connected layer weight vector, which is used to enhance the generalization of the model. Based on this, the scoring function and loss function of the attributes are shown in Eq:

$$S_{attr}(e,a,n)=\left\|e-C'\right\|_2 \tag{6}$$

$$L_{attr}=\sum_{(e,a,n)\in Y}\log\left(1+e^{S_{attr}^2(e,a,n)}\right) \tag{7}$$

2.2 Image embedding

The integration of media pictures into multimodal utterance analysis is achieved by convolutional neural networks, which are closely related to multimodal discourse analysis.

2.2.1 Convolutional neural network

The convolutional neural network is one of the representative algorithms of deep learning, which directly takes the original information of the image as input, thus reducing the time of the algorithm.

1) Convolutional layer

The convolutional layer is the core infrastructure in the convolutional neural network. The role of convolution is to extract a local feature of the input data, and each convolution kernel in the convolution layer is a feature extractor, so that features can be fully extracted.

We assume that the input signal sequence x of a one-dimensional convolutional layer is $x_1, x_2, ..., x_t$

and the convolutional kernel or filter ω of the convolutional layer is $\omega_1, \omega_2, ..., \omega_m$. Then the expression for the convolution of the filter and the input signal sequence is calculated as

$$y_t = \sum_{k=1}^{m} \omega_k \cdot x_{t-k+1} \tag{8}$$

2) Pooling layer

The specific role of the pooling layer is to filter and compress the feature mapping so that the number of features can be reduced to achieve the effect of reducing the number of parameters. The pooling operation is to divide the feature mapping input to the pooling layer into regions, and the divided regions can overlap, and the more common pooling operations are average pooling and maximum pooling. The output of average pooling is the sum of all values in a region of the input feature mapping, while the output of maximum pooling is the maximum value of a region.

3) Fully-connected layer

The fully-connected layer is generally used as a classification layer, which integrates the feature information obtained from the previous convolution layer, pooling layer and activation function and maps them in the sample space. The fraction range of the input image in the sample space is $[-\infty, +\infty]$, which needs to be normalized by the classifier to achieve the final classification.

For the multiclassification problem, the expression for calculating the conditional probability that an input sample x, *Softmax* is predicted to belong to a category c is:

$$p(y = c \mid x) = \frac{e^{\omega_c^T x}}{\sum_{j=1}^{C} e^{\omega_j^T x}} \tag{9}$$

Where C represents the total number of categories, ω_c represents the weight vector of the category c, and p represents the conditional probability that the category belongs to the category c. The conditional probability obtained after *Softmax* has a value between 0 and 1, and the sum of the probability values of all categories should be 1.

2.2.2 Convolutional neural network based image embedding

While some existing works use ResNet to learn the vector representation of images, this paper embeds images based on the DenseNet model proposed by Huang et al. In the embedding process, the last fully connected layer of DenseNet is removed, and each image is processed into a 1024-dimensional vector. The most important feature of DenseNet compared with ResNet is that it uses a dense connection mechanism, which results in more efficiency and feature reusability than ResNet. The scoring function for image data processing is shown in Eq:

$$i = DenseNet(i) \tag{10}$$

$$S_{img}(e, i) = \left\| e - \tanh(flat(i)) \right\|_2 \tag{11}$$

Where e denotes the feature embedding of entity e, i denotes the feature vector representation obtained after passing the picture i through DenseNet. $DenseNet(\cdot)$ is a DenseNet layer.

According to the above scoring function, the loss function of the picture part is shown in Eq:

$$L_{img} = \sum_{(e,i)\in P_i} \log\left(1+\exp\left(S^2_{img}(e,i)\right)\right) \tag{12}$$

2.3 Video embedding

Media video embedding is paramount in multimodal utterance analysis, which allows detailed analysis of emotions and, thus, intuitive construction of cultural identity.

2.3.1 Media convergence video embedding

For the processing of video data, most of the existing works use the traditional 3DCNN model, while this paper uses the Timesformer model proposed by Bertasius et al. This is mainly because:

1) The average duration of video data in the Douban-Baidu dataset is 2 minutes.

2) Timesformer's temporal attention mechanism allows it to model long video frame sequences, while the traditional 3DCNN model can only handle short video data of a few seconds at most. Therefore, the Timesformer model is chosen in this paper. Specifically, firstly, each video is first keyframed extracted. Then, these keyframes are processed into a vector of dimensions by Timesformer. The scoring function for video data processing using Timesformer is shown in Eq:

$$v = Timesformer(v) \tag{13}$$

$$S_{vid}(e,v) = \left\| e - \tanh\left(proj(v)\right)\right\| 2 \tag{14}$$

Where e denotes the feature embedding vector of the entity e, v denotes the feature embedding vector obtained after the video v is passed through Timesformer, $Timesformer(\cdot)$ is a Timesformer layer, and $proj(\cdot)$ denotes the projection operation, which aims to map the high-dimensional embedding to the low-dimensional space. Based on the above scoring function, the loss function of the video data part is defined as shown in Eq:

$$L_{vid} = \sum_{(e,v)\in P_v} \log\left(1+\exp\left(S^2_{vid}(e,v)\right)\right) \tag{15}$$

2.3.2 Entity alignment loss function

In terms of multimodal data fusion, existing methods do not sufficiently consider the interaction of different modal information and do not handle the noise in multimodal information well, so there is much room for improvement in multimodal knowledge fusion in existing methods. For example, the model MMEA simply splices the knowledge representations from multiple modalities, and the interactions among the modalities are not sufficient, so the fusion effect is not particularly satisfactory. To address the previous shortcomings, a new multimodal interaction fusion module is proposed to further fuse the feature embedding vectors obtained from the multimodal knowledge embedding module. In detail, in the multimodal interaction fusion module, an attention mechanism is designed to further improve the performance of the multimodal knowledge map entity alignment task by not only learning the interaction information between any pair of modalities but also considering the interaction information between any modal pair. The steps are as follows: first, the initial interaction

information is obtained by multiplying each modality two by two, after which it is normalized by a *softmax* -layer; then, the interaction information between different modal pairs is stored with the help of a memory unit, and finally, inspired by the successful application of residual networks in picture embedding, this paper incorporates the weighted information of the original paired modal information in the extracted information and obtains the final entity embedding representation. The formal representation of the multimodal knowledge interaction fusion module is as follows.

Given any platform, the initial interaction information is first obtained by multiplying the individual modal information in the platform two by two, as shown in Eq:

$$E'_{mn} = softmax\left(E_m \cdot E_n^T\right) \tag{16}$$

Where T represents the transpose of the matrix, the values of m and n depend on which two modal embedding information is used for the interaction, and $m,n \in \{R,I,A,V\}$, E_m and E_n represent the embedding representation of any two modalities, respectively. Immediately after, input *softmax* is normalized. To further improve the fusion effect, this paper stores the interaction information between different modal pairs with the help of a shared memory unit W. Finally, the final entity embedding representation E'_{mn} is obtained by incorporating the weighted information of the original paired modal information in the extracted information using the residual operation, as shown in Eq:

$$E_{mn} = \frac{\left(E_m + E_n\right)}{2} + W \cdot E'_{mn} \tag{17}$$

Among them, the weighted information of the original paired modal information E_m and E_n is obtained by summing to find the mean value.

The multimodal knowledge interaction fusion module not only learns the interaction information between any pair of modalities but also considers the interaction information between any pair of modalities, thus enabling the multimodal features to fully interact, enhancing the complementarity of multiple modalities, reducing the noise caused by a single modality, and improving the performance of the multimodal knowledge mapping entity alignment task. Meanwhile, the module maps individual modal information from their respective spaces to the common space based on the multimodal knowledge interaction fusion loss function. The formal definition is shown in Eq:

$$L_{in}\left(E,E_R,E_A,E_I,E_V\right) = \sum_{m,n\in\{R,I,A,V\}} \omega_{mn} \cdot \left\|E - E_{mn}\right\|_2^2 \tag{18}$$

where ω_{mn} is the hyperparameter. In the experiment, ω_{RA}, ω_{RI}, and ω_{RV} are set to 0.1, and ω_{IV}, ω_{IA}, and ω_{AV} are set to 0.01.

For data on different knowledge graphs, the entities aligned in the annotated dataset point to the same object in the real world. Therefore, in the common space, these two entities should be closer in theory. Based on this, this paper defines the entity alignment loss function L_{final} as shown in Eq:

$$L_{final}\left(E_1,E_2\right) = \left\|E_1 - E_2\right\|_2^2 \tag{19}$$

Where E_1 and E_2 represent the embedded representations of the entities in Z_1 and Z_2, respectively. It is worth stating that Z_1 and Z_2 are defined as $Z_1 = \{e_1 \mid e_1 \in KG_1 \cap (e_1, e_2) \in A\}$, $Z_2 = \{e_2 \mid e_2 \in KG_2 \cap (e_1, e_2) \in A\}$, respectively.

3 Affective analysis of model-based cultural identity construction

Most current approaches focus on integrating data from different modalities into a unified feature representation and exploiting the complementary nature of multimodal data to enhance recognition. To address the problem that multimodal data of text, speech and images have uncertainties such as noise and redundancy and lack reliable assessment of prediction results, we propose to analyze the consistency of cultural identity sentiment based on uncertainty estimation of multimodal utterances to construct cultural identity.

3.1 Experimental preparation

3.1.1 Data collection

Two widely used public datasets in the field, CMU-MOSI and CMU-MOSEI, were used.

The CMU-MOSEI dataset is shown in Table 1.

Table 1. shows the data set of CMU-MOSI

Total number of subjective opinion fragments	2300
Total video count	101
Number of speakers	92
Average opinion segment	22.8
Average segment length	4.8s
Average number of words per opinion segment	14
Total number of words in opinion segment	27523
Opinion snippets do not repeat the number of words	3200
The number of words that appear more than 10 times	3200

The CMU-MOSEI dataset is shown in Table 2.

Table 2. shows the situation of CMU-MOSEI data set

Total number of subjective opinion fragments	24567
Total video count	3300
Number of speakers	1200
Average segment length	268
Average opinion segment	8.4
Average number of words per opinion segment	8.12
Total number of words in opinion segment	456786
Opinion snippets do not repeat the number of words	23258
The number of words that appear more than 10 times	3528
The number of words that appear more than 20 times	1980

The number of words that appear more than 50 times	868

For each video clip, it was divided into 7 categories according to the emotional intensity from Strongly Positive to Strongly Negative; strongly Positive, Positive, Weakly Positive, Neural, Weakly Negative, Negative, and Strongly Negative were labeled as 3, 2, 1, 0, -1, -2, and -3 respectively.

3.1.2 Evaluation indicators

For CMU-MOSI and CMU-MOSEI, the two datasets are linearly labeled according to the degree of affective states from negative to positive, so it can be regarded as a regression task. Neural, Weakly Negative, Negative, Strongly Negative and Positive. Therefore, it can also be regarded as a classification problem. Therefore, we introduced the mean average absolute error MAE seven classification accuracy Acc-7, dichotomous classification accuracy Acc-2, and F1 values as evaluation metrics.

3.2 Multi-modal lateral comparison experimental results and analysis

In order to verify the emotion recognition effect of the multi-level cross-modal-aware emotion recognition network proposed in this paper, we selected currently available excellent multimodal emotion recognition algorithmic models as a baseline method for cross-sectional comparison with the algorithmic models proposed in this paper, including:

1) EF-LSTM network: an early fusion strategy is used to splice multimodal feature vectors in the feature layer and then feed them into an LSTM network for prediction after learning.

2) LF-LSTM network: using a late fusion strategy, three different LSTM networks are used to learn the information of each modality, and the output is spliced for prediction.

3) RAVEN network: captures the dynamic nature of nonverbal intent by changing word representations based on concomitant nonverbal behavior and dynamically transforms word representations based on nonverbal cues.

4) MulT network lt3: It focuses on the interaction between multimodal sequences at different time steps and potentially adapts the data stream from one modality to another; and is one of the most advanced methods available.

We conducted comparison experiments on the baseline model and the model proposed in this paper to test the sentiment recognition effect. The test results models of different models on the CMU-MOSI dataset are shown in Table 3.

Table 3. shows the test results of different models in the CMU-MOSI dataset

Model	Acc-7	Acc-2	F1	MAE
EF-LSTM	32.1	74.2	75.6	1.138
LF-LSTM	34.8	78.2	78.9	0.899
RAVEN	32.7	73.1	74.8	1.108
MulT	39.8	82.3	82.3	0.876
The Model	41.2	84.6	84.2	0.848

The test result models of different models on the CMU-MOSEI dataset are shown in Table 4.

Table 4. shows the test results of different models in the CMU-MOSEI dataset

Model	Acc-7	Acc-2	F1	MAE
EF-LSTM	48.3	75.2	76.3	0.688
LF-LSTM	50.8	78.1	79.2	0.628
RAVEN	47.7	78.4	75.8	0.638
MulT	50.2	82.9	82.9	0.588
The Model	51.1	84.8	84.5	0.548

As can be seen from Table 3, compared to the EF-LSTM method and LF-LSTM method, the model layer fusion approach in this paper can learn the interaction of cross-modal elements at the model level using the cross-modal Transformer mechanism and achieves a 6.4 to 10.4 percentage point improvement in the dichotomous classification accuracy and a maximum reduction of 0.223 in the MAE metric in the regression task. MulT and RAVEN models, the model in this paper achieved 4 and 10.6 percentage point performance gains in Acc36 for the sentiment classification task on the CMU-MOSI dataset and a maximum reduction of 0.212 in MAE in the regression task.

The experiments conducted on the CMU-MOSEI dataset can be obtained from Table 4, still achieving the best experimental results in the baseline model, with Acc-2 reaching 84.8%, F1 values reaching 84.5, and MAE for the regression task dropping to 0.548, with different magnitudes of performance improvement.

3.3 Ablation experiments and emotional robustness analysis

3.3.1 Ablation experiments

To test the contribution of each component of the model, i.e., the cross-modal interaction layer, the unimodal enhancement layer, and the EmbraceNet fusion layer, ablation experiments were conducted on our CMU-MOSI dataset. Among them, the model without a cross-modal interaction layer is denoted by w/oc, the model without a unimodal enhancement layer is denoted by w/os, and the model without an EmbraceNet fusion layer but with simple vector stitching is denoted by w/oe. Figure 1 shows the results of the four models in the ablation experiment.

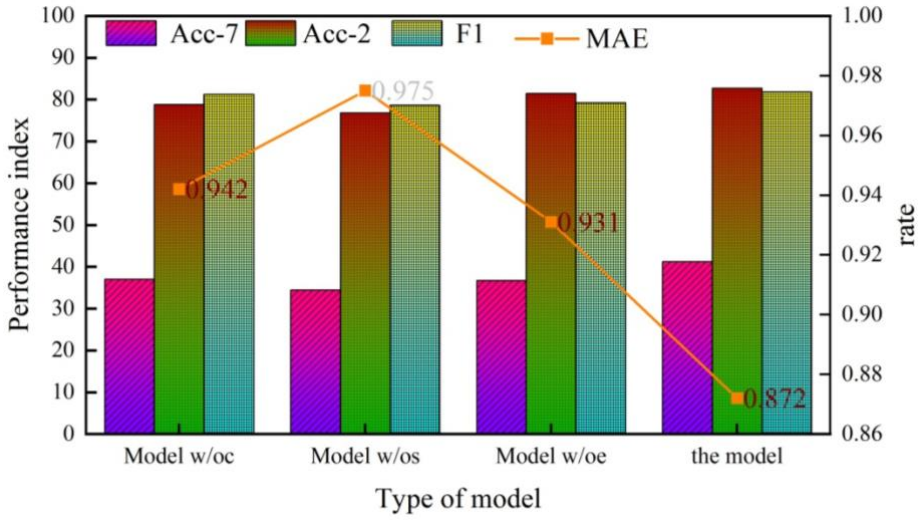


Figure 1. shows the results of the four models in the ablation experiment

As we can see in Figure 1, removing any module from the model leads to a decrease in model performance. Specifically, removing the unimodal enhancement layer has the greatest impact on model performance, resulting in a 5.2% decrease in binary classification accuracy, as well as a decrease of about 10% in absolute error from the peak, which indicates that emotions are expressed more through the global semantic expression of the context, and the gating mechanism of the gating loop unit can effectively highlight the key information on the whole temporal sequence. After discarding the cross-modal interaction layer, although the complexity of the model operation is reduced, a channel for information communication between modalities is also lost, leading to the degradation of model recognition performance. Regarding the fusion of the final features, after discarding the EmbraceNet fusion layer, although the fusion information is richer, the accuracy and F1 values in the experimental results are lower than the effect of the original model due to the lack of a regularization mechanism like EmbraceNet.

3.3.2 Robustness evaluation and analysis

To evaluate whether our model can resist perturbations and still maintain the robustness of sentiment recognition in the presence of missing data, we conducted the following experiments on the CMU-MOSI dataset. In the time series data of each modality, data on time steps of a certain length are randomly selected and deleted, and the deleted part of the data is filled with zero values. Where the missing rate is equal to the number of missing time steps divided by the total length of the data, we use 20% to 84% with an interval of 8% as the target missing rate.

Figure 2 shows the change in Acc-2 metrics when some data are missing.

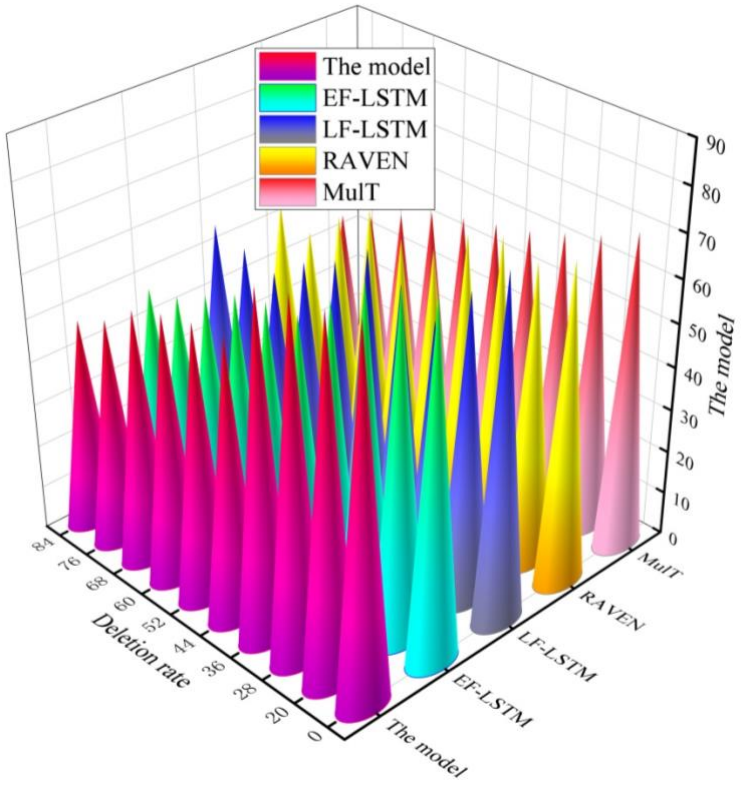


Figure 2. shows the changes of AC-2 index when partial data is missing

Figure 3 shows the change in Acc-7 indicators when some data are missing.

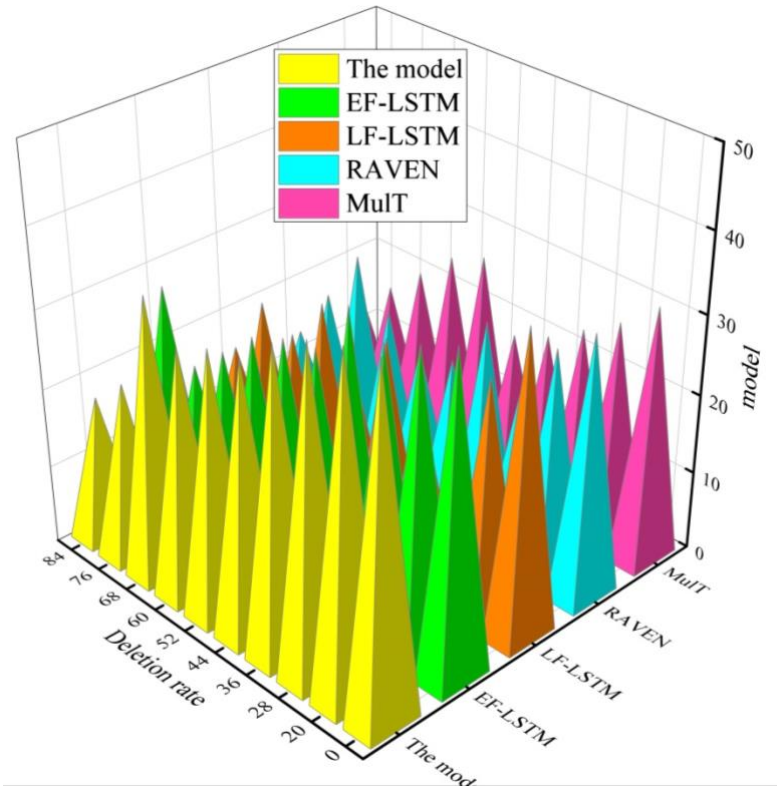


Figure 3. shows the change of AC-7 index when partial data is missing

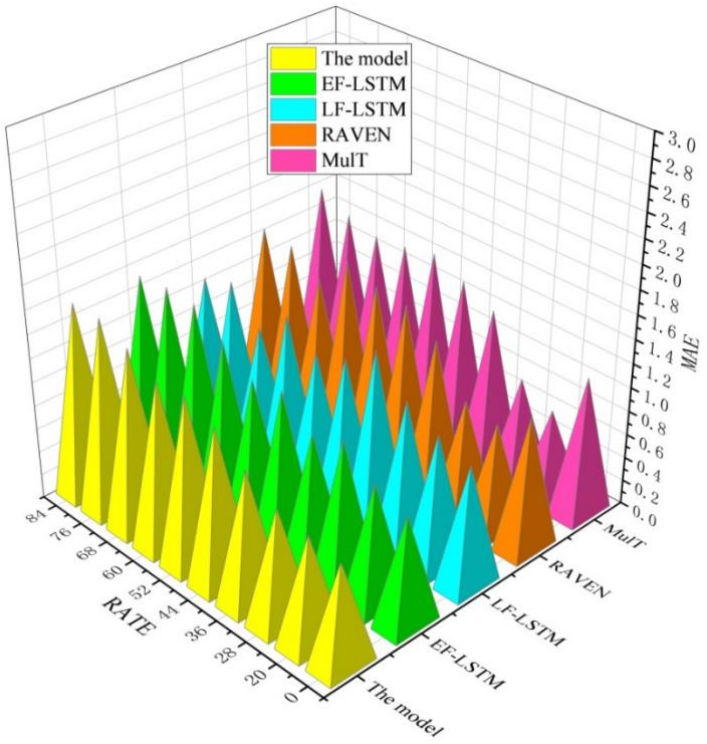


Figure 4. shows the change of MAE index when partial data is missing

It can be obtained that the performance of the network model tends to decrease when different degrees of missing data are introduced due to the reduction of information. On the MOSI dataset, the baseline approach is exceeded on most of the evaluated metrics; for example, in the presence of 44% missing rate, the Acc-2, Acc-7, and MAE values are 79.5%, 38.6%, and 1.468, respectively, which are significantly better than the performance metrics of other baseline models in the presence of the same missing rate. In particular, in contrast to the MulT model, which in the absence of missing data, is equipped with a similar performance to the model proposed in this paper, however, when the missing rate reaches 44%, Acc-2 is only 65.2%, and the difference between the effect of 76.5% dichotomous accuracy achieved by the model in this paper reaches 11.3 percentage points. These results confirm that our model is effective in ensuring robust performance despite incomplete data.

3.4 Affective analysis of model-based cultural identity construction

The experiments in 3.2 and 3.3 show that in the multimodal discourse analysis in the media fusion perspective, it is possible to express emotions steadily, combine the model with the construction of cultural identity, and build cultural identity from people's emotions by fusing multimodal discourse analysis, and cultural identity is output through the fusion of text, pictures and videos with emotions. From the emotional analysis, "cultural identity" is the affirmative recognition of the most meaningful things of the nation formed by people living together for a long time in a national community, the core of which is the identification with the basic values of a nation, the spiritual bond that unites the national community, and the spiritual foundation for the continuation of the life of the national community.

4 Conclusion

In this paper, the following conclusions were obtained based on the construction of cultural identity in the model of sentiment analysis in multimodal discourse analysis from the media convergence perspective:

1) This paper achieved the best experimental results in the baseline model, with Acc-2 reaching 84.8%, F1 value reaching 84.5, and MAE of the regression task dropping to 0.548, with different magnitudes of performance improvement, which can construct cultural identity well.

2) When the missing rate reaches 44%, Acc-2 is only 65.2%, and the gap between the effect of 76.5% dichotomous accuracy achieved by the model in this paper reaches 11.3 percentage points. These results confirm that our model effectively ensures a robust performance despite incomplete data, indicating that the model in this paper is able to construct cultural identity stably.

3) From the emotional analysis, cultural identity is the affirmative recognition of the most meaningful things of the nation formed by people living together in a national community for a long time, and its core is the identification with the basic values of a nation, the spiritual bond that unites this national community, and the spiritual foundation for the continuation of the life of this national community.

References

[1] Okladnikova, Y. A. (2020). PEDAGOGIC CONDITIONS FOR FORMING A CULTURAL IDENTITY OF RURAL SCHOOLCHILDREN OF LENINGRAD REGION WITH THE EYES OF TEACHERS. Vestnik Kostroma State University Series Pedagogy Psychology Sociokinetics, (1), 41-45.

[2] Trung, N. S., Van, V. H. (2020). Vietnamese Cultural Identity in the Process of International Integration. Journal of Advances in Education and Philosophy, 4(6), 220-225.

[3] Meca, A., Sabet, R. F., Farrelly, C. M.,et al. (2017). Personal and Cultural Identity Development in Recently Immigrated Hispanic Adolescents: Links With Psychosocial Functioning. Cultural Diversity & Ethnic Minority Psychology, in press(3).

[4] Ozer, S., Bertelsen, P., Singla, R., et al. (2017). "Grab Your Culture and Walk With the Global": Ladakhi Students' Negotiation of Cultural Identity in the Context of Globalization-Based Acculturation. Journal of Cross-Cultural Psychology, 48(3). 002202211668739.

[5] G. Szabó., Ward, C. (2022). Development and Validation of a Short Form Version of the Ethno-Cultural Identity Conflict Scale (EICS-SF):. Assessment, 29(5), 1020-1032.

[6] Pirri, Valentini, A. (2021). Translating the concept of "cultural identity" in public policies: Between the international and the national, and the tangible and intangible dimension. International Journal of Constitutional Law, 19(5), 1756-1777.

[7] Caust, J. (2019). Cultural rights as human rights and the impact on the expression of arts practices. Journal of Citizenship and Globalisation Studies, 3(1), 17-30.

[8] Vecchi, A., Silva, E. S., Angel, L. (2020) Nation branding, cultural identity and political polarization – an exploratory framework. International Marketing Review, ahead-of-print(ahead-of-print).

[9] Yim, H. (2002) Cultural identity and cultural policy in South Korea. International journal of cultural policy, 8(1), 37-48.

[10] Albawardi, A., Jones, R. H. (2020). Vernacular mobile literacies: Multimodality, creativity and cultural identity. Applied Linguistics Review, 11(4), 649-676.

[11] Yuan, H., Tang, Y., Xu, W., & Lau, R. Y. K (2021). Exploring the influence of multimodal social media data on stock performance: an empirical perspective and analysis. Internet Research, ahead-of-print(ahead-of-print).

[12] CN, Tseung-Wong, Dandy, J., Lane, M. (2022). Perceived diversity norms, cultural identity styles and bicultural identity consolidation in two bicultural groups in Australia. International Journal of Psychology, 2022, 57(3):363-371.

[13] Schachner, M. K. (2019). From equality and inclusion to cultural pluralism - Evolution and effects of cultural diversity perspectives in schools. European journal of developmental psychology, 16(1).

[14] Aral, T., Schachner, M., Juang, L., et al. (2021). Cultural diversity approaches in schools and adolescents' willingness to support refugee youth.. The British journal of educational psychology, e12458.

[15] Dressler, W. W. (2020). The construction of the cultural niche: A biocultural model. American Journal of Human Biology, 32(4).

[16] Sugimura, K., Crocetti, E., Hatano, K., et al. (2018). A cross-cultural perspective on the relationships between emotional separation,parental trust, and identity in adolescents. Journal of Youth and Adolescence, 47, 749-759.

[17] Altugan, A. S. (2015) The relationship between cultural identity and learning. Procedia-Social and Behavioral Sciences, 186, 1159-1162.