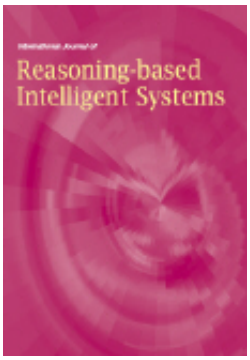
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556
https://www.inderscience.com/ijris

Objective evaluation of English teaching in colleges and universities based on textual analysis in multi-element perspective

Xiaohua Chen

DOI: 10.1504/IJRIS.2025.10071973

Article History:

Received:	06 May 2025
Last revised:	24 May 2025
Accepted:	24 May 2025
Published online:	10 July 2025

Objective evaluation of English teaching in colleges and universities based on textual analysis in multi-element perspective

Xiaohua Chen

School of Foreign Languages,
Zhanjiang University of Science and Technology,
Zhanjiang 524094, China
Email: chenxiaohua202504@163.com

Abstract: Focusing on the lack of objectivity in college English teaching assessment due to the neglect of multi-modal elements' features in current studies, this paper first designed the cross-modal attention mechanism (CMAM) to learn relevant features among modal elements in the text, and used the gating mechanism to realise the adaptive fusion of multi-modal important information. Then use Transformer model to integrate multi-modal features and modal importance information to realise text sentiment analysis. The analysis results were fused into the evaluation method, and weighted naive Bayes (WNB) algorithm was used to determine the weights of indicators. Finally, a more objective evaluation result is obtained by adjusting the contribution weight of the text to the evaluation result according to students' emotional tendency. The experimental outcome indicates that the F1 of the proposed approach is improved by at least 5.14%, which verifies the effectiveness of the proposed method.

Keywords: English teaching assessment; text analysis; cross-modal attention mechanism; multimodal element; weighted naive Bayes.

Reference to this paper should be made as follows: Chen, X. (2025) 'Objective evaluation of English teaching in colleges and universities based on textual analysis in multi-element perspective', *Int. J. Reasoning-based Intelligent Systems*, Vol. 17, No. 8, pp.11–20.

Biographical notes: Xiaohua Chen received her Master's degree from the University of Leeds, UK in 2024. She currently serves as a Lecturer at the Zhanjiang University of Science and Technology. Her research interests include English language teaching (ELT) and artificial general intelligent control.

1 Introduction

As English teaching reform continues to be intensified, people are increasingly concerned about the educational quality of English programs. The quality and impact of teaching directly determine the quality of talent formation in colleges (Jia and Zhang, 2020). In the academic affairs system of universities, a large amount of English teaching evaluation text information is stored each semester. This text information contains multimodal elements such as text and images. Since it is difficult to analyse this evaluation text quantitatively, it is not being used to its full value, and reading through the text one item at a time requires a lot of time and effort (Okoye et al., 2020). However, the current English teaching evaluation system in universities still has some limitations, such as over-reliance on subjective questionnaire surveys or single test scores, and a lack of objective analysis of the multimodal elements of the teaching process (Rockoff and Speroni, 2010). Therefore, the analysis of text in classroom teaching evaluation can not only promote the quality of teaching in colleges, but also play a positive part in driving the practical utilisation of academic research.

Jiang et al. (2018) used SVM to classify teaching evaluation texts and introduced a fuzzy neural network to obtain objective evaluation results. Li (2022) proposed a method for evaluating teachers' teaching performance based on the K-means clustering algorithm, but the accuracy was not high. Hou (2021) used SVM and decision trees to classify the sentiment of evaluation texts in the teaching administration system. The classification results were combined with various indicators that have an impact on the quality of teachers' instruction, and the evaluation results were obtained by assigning different weights to the indicators. Li (2021) used a combination of analytic hierarchy process and decision trees to evaluate teaching quality, but the objectivity of the evaluation was insufficient. Zhang et al. (2022) combined the features of English teaching, fully considered the characteristics of emotional words, the position weight of words and the modification of mutual relationships, proposed an SVM for dependency syntax analysis, and classified the evaluation texts, nevertheless, the accuracy rate of classification was unsatisfactory. Huang (2021) analysed the impact of different stop word deletion methods on text classification results, and then used the TF-IDF function to calculate

feature weights based on sentiment-labelled texts, and SVM technology to implement teaching evaluation.

The above-mentioned machine learning-based text communication feature expression capability is very weak, and manual feature extraction is required, which is very costly. Deep learning greatly improves the efficiency of text analysis by automatically extracting features and reducing manual intervention. Geng (2021) used a CNN to classify text. New short text vector representations were obtained by adding words and related concepts to the pre-trained vectors, and the classification effect was good. Qin and Irshad (2024) used the TF-IDF algorithm (Qaiser and Ali, 2018) to weight word vectors and input text vectors into a CNN to automatically extract text features. The text features and various teaching indicator features were quantified to obtain more reasonable evaluation results. Yang (2023) used an LSTM to classify teaching evaluation texts, improving the accuracy of text classification. Chen and Aleem (2024) used an LDA topic model and an LSTM network to automate the extraction and classification of teaching evaluation texts, improving the accuracy of teaching evaluations. Qi et al. (2024) used transformer to perform feature enhancement and classification on evaluation texts, achieving good classification results.

In addition to text, the text evaluation data also includes multimodal elements such as images and audio. Therefore, these multimodal elements need to be considered during text analysis. Aljaloud et al. (2022) fused the features of the audio and image elements in the teaching text through an attention mechanism, and used CNN and LSTM to extract the audio and image features respectively. Su and Peng (2023) proposed a multi-layer attention network model that integrates speech and text annotation, aiming to extract sequence-level features from speech and combine them with the analysis of annotated text to obtain the final classification result. Yan et al. (2023) used an LSTM and an attention mechanism to fuse the features of multimodal elements in the evaluation text. Appropriate weights were generated based on the text context to combine features from different modalities. Sebbaq and El Faddouli (2022) introduced a gated recurrent unit (GRU) to learn the nonlinear combination of multimodal elements in the teaching evaluation text to calculate the displacement vector, but the feature distribution between the modes is very different, resulting in poor feature fusion of multiple elements.

A comprehensive analysis of the current research status shows that existing English teaching evaluation methods ignore the characteristics of multimodal elements in the text, resulting in poor objective evaluation of teaching. Thus, this article suggests an objective assessment approach for college English teaching based on text analysis from a multi-element perspective. First, the three modal elements of text, audio, and image are represented by features. Guided by the text mode, CMAM is designed to learn the correlation features between text and non-text modes. At the same time, a gating scheme is used to gain the singular features of the three modal elements, thereby achieving hierarchical and adaptive fusion of important multimodal information. Then, the transformer model is used to fuse the features of multimodal elements and modal importance information to achieve text sentiment analysis. The quantified text analysis results are integrated into the objective evaluation of English teaching. Based on the influence degree of different indicators on the evaluation outcomes, a weighted naïve Bayes algorithm is employed to determine the weight values of each evaluation metric. Finally, the contribution weight of text segments annotated with abnormal emotions to the overall English teaching score is adjusted based on students' sentiment tendencies. Through a weighted summation approach, all evaluation dimension scores across the course are aggregated, ultimately yielding a more objective assessment result of English teaching effectiveness. The experimental outcome implies that the proposed approach has an F1 of 90.81%, which is 5.14–18.66% higher than the benchmark method. It holds significant theoretical and practical implications for enhancing the teaching evaluation system.

2 Relevant technologies

2.1 LSTM network

A LSTM network is a modified version of an RNN designed to overcome the gradient explosion problem inherent in traditional RNNs when handling extended input sequences (Farzad et al., 2019). GRU is another variant of RNN, but GRU's simplified gating structure may lose some of the detail information in long sequences, and it may not perform as well as LSTM in text analysis tasks. LSTM introduces an internal state c is employed to retain long-term information. A gating mechanism is incorporated to regulate the flow of information. The 'gate' refers to a fully connected layer whose input is a vector and whose output is a real vector scaled to the interval (0, 1).

The internal calculation process of the LSTM network at time t is implied in Figure 1. The LSTM model employs three gates: the forget gate f_t, the input gate i_t, and the output gate o_t. f_t regulates which information is to be removed from the prior cell state at the present moment. i_t regulates which information from input x_t should be retained in cell state c_t at the present moment. o_t controls what information from c_t needs to be output at the present moment. i_t, f_t, o_t, the external state h_t and c_t are calculated as follows:

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, w_{it}] + b_i\right) \tag{1}$$

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, w_{it}] + b_f\right) \tag{2}$$

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, w_{it}] + b_o\right) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh\left(W_c \cdot [h_{t-1}, w_{it}] + b_c\right) \tag{4}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{5}$$

where W_i, W_f, W_o and W_c are the weights of the input gate, forget gate, output gate, and internal state, respectively, and σ is the activation function.

2.2 Attention mechanism

The attention mechanism in deep learning models effectively tackles information overload, simplifies model structure, and boosts performance. Attention mechanisms have seen widespread adoption across various deep learning domains, including natural language processing, picture detection, yielding exceptional outcomes (Soydaner, 2022). The attention mechanism enables models to focus on task-relevant information by amplifying its importance and diminishing the impact of irrelevant data. Suppose the contents of the memory consist of an address *key* and a value *value*, and a task-related query vector q is given. The probability of selecting each value is represented by calculating the similarity between q and *key*, that is, the attention distribution. Finally, the result is obtained by summing the weighted values. The general form of an attention mechanism is as follows:

$$s_i = score(Key_i, q) \tag{6}$$

$$a_i = Softmax(s_i) = \frac{\exp(s_i)}{\sum_i \exp(s_i)} \tag{7}$$

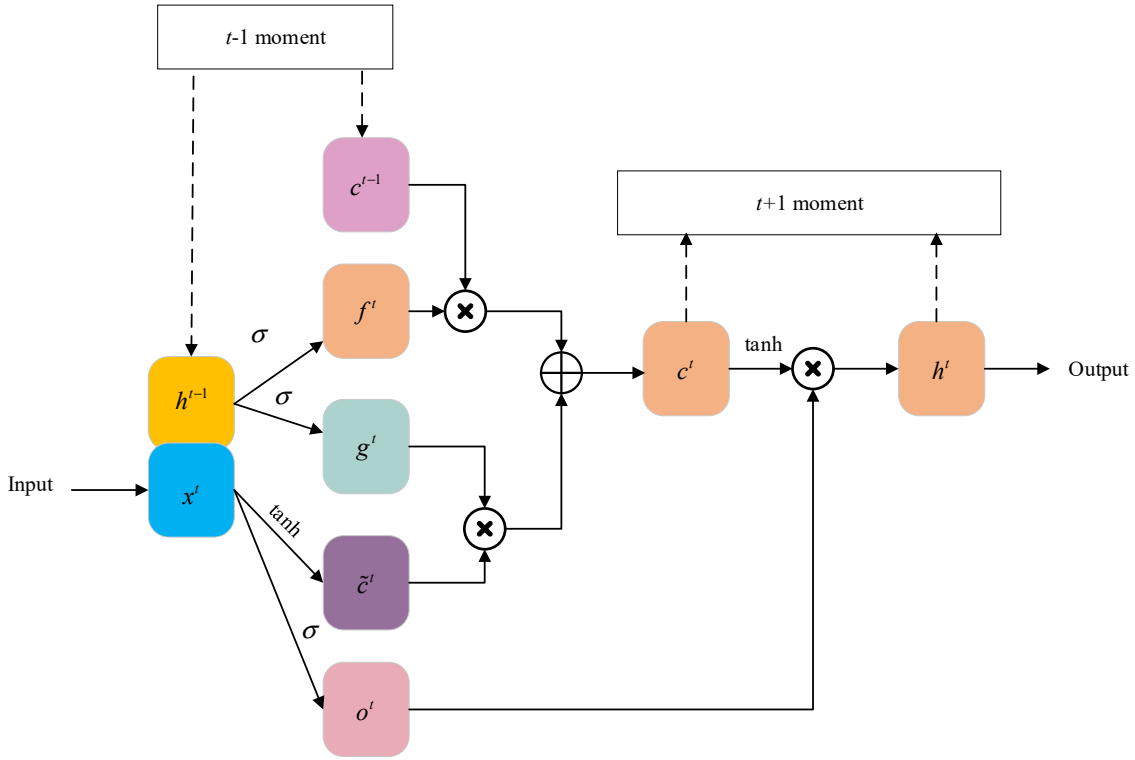
where a_i is the attention weight, which represents the importance of the i^{th} message in the form of a probability. The $Softmax(\cdot)$ function is also called the exponential normalisation operation, this function maps a K-dimensional vector to another with values in (0, 1).

3 Text categorisation for English teaching evaluation based on multimodal element fusion

3.1 Representation of the characteristics of multimodal elements and local interactions in teaching evaluation texts

To address the issue that the existing English teaching evaluation methods ignore the characteristics of multimodal elements in the evaluation text, resulting in poor classification results, this paper designs a text classification model for English teaching evaluation relied on the integration of multimodal elements. The model structure is shown in Figure 1. First, text, image, and audio elements are adaptively selected and integrated based on textual modal information. Then, based on CMAM, important information is represented between two modalities. A multimodal adaptive gate control mechanism is designed to realise hierarchical adaptive integration relied on multimodal vital data. Finally, the text sentiment classification is realised by comprehensively considering the features of multimodal elements in the text and the modal importance information.

Figure 1 Internal calculation process of the LSTM model (see online version for colours)



The English teaching evaluation text mainly involves three modal elements: text, audio, and images. The input sequence is defined as $F_i \in R^{l_i}$, $i \in \{t, a, v\}$, where $l_{\{t,a,v\}}$ indicates the sequence length of each modal element. Three independent subnetworks are used to obtain the feature representations of the three modal elements. For the textual modality, a pre-trained 12-layer BERT (Aum and Choe, 2021) is used to extract the sentence representation, and the first word vector in the last layer is used as the representation of the entire sentence. The feature representation obtained by BERT for evaluating the textual modality is as follows:

$$H_t = BERT\left(F_t, \theta^{bert}\right) \tag{8}$$

where $H_t \in R^{l_t \times d_t}$ indicates the text modal feature; l_t is the sequence length of the text mode; d_t is the feature dimension of the text mode; and θ^{bert} is the network parameter of the BERT model.

For the speech and image modalities, a unidirectional LSTM is used to obtain the temporal features corresponding to the two modalities, and the hidden moment state of the last layer is used as the representation of the entire sequence. F_a and F_v respectively obtain the feature representations of the speech and image modal elements through a unidirectional LSTM, as shown in equations (9) and (10), respectively.

$$H_a = LSTM\left(F_a, \theta_a^{lstm}\right) \tag{9}$$

$$H_v = LSTM\left(F_v, \theta_v^{lstm}\right) \tag{10}$$

where $H_a \in R^{l_a \times d_a}$ is the feature of the speech element, $H_v \in R^{l_v \times d_v}$ is the feature of the picture element; l_a and l_v stand for the sequence length of the speech and picture elements, individually; d_a and d_v is the feature dimension of the speech and picture elements, individually; and θ^{lstm} is the network parameter of the LSTM model.

Guided by the text modality, a CMAM is designed to learn the relevant features between the text modality and non-text modalities. Taking the interaction between text and image modality elements as an example, as shown in Figure 3. When there are two elements, image modality (V) and text modality (T), and the features are denoted as H_v and H_t, the cross-modal attention from text modality to image modality is represented as follows:

$$\begin{aligned} CM_{t\to v}\left(H_v, H_t\right) &= \text{softmax}\left(\frac{Q_v K_t^T}{\sqrt{d_k}}\right)V \\ &= \text{softmax}\left(\frac{H_v W_{Q_v} W_{K_t}^T H_t^T}{\sqrt{d_k}}\right)H_t W_{V_t} \end{aligned} \tag{11}$$

where $W_{Q_v} \in R^{d_v \times d_k}$, $W_{K_t} \in R^{d_t \times d_k}$ and $W_{V_t} \in R^{d_t \times d_v}$ are the linear transformation weight matrices. This paper uses two CMAM modules to obtain two sets of modal interaction features for text-to-speech and ext-to-image. At this time, H_t provides the K and V vectors, and H_a and H_v provide the Q vectors, as represented below:

$$H_T^A = CM_{T\to A}\left(H_T, H_A\right) \tag{12}$$

$$H_T^V = CM_{T\to V}\left(H_T, H_V\right) \tag{13}$$

Figure 2 The text categorisation model for English teaching evaluation (see online version for colours)

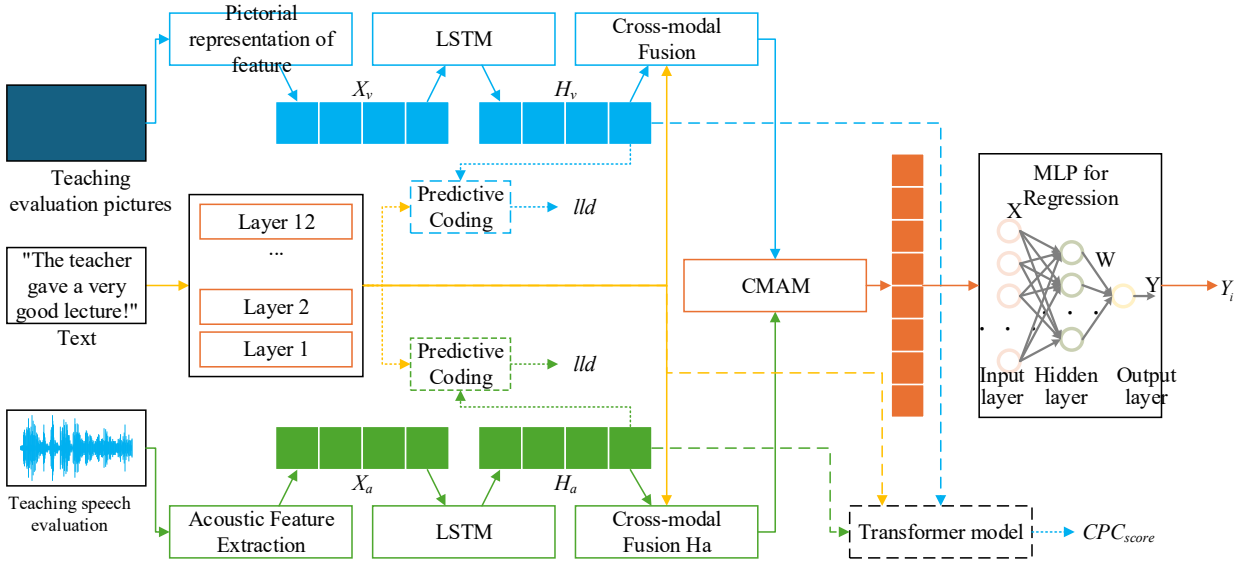
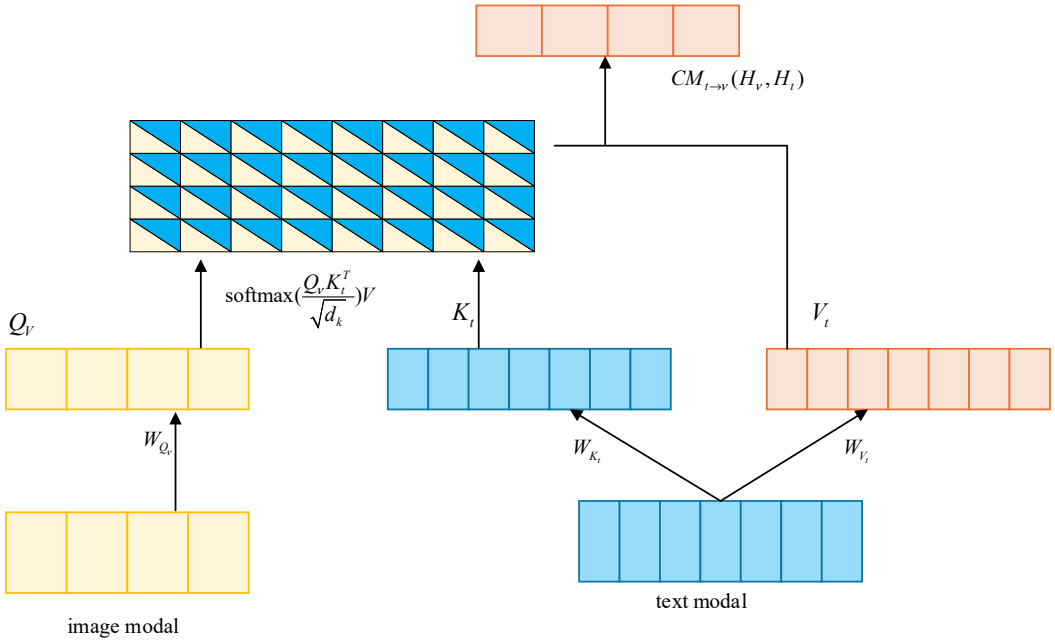


Figure 3 Cross-modal attention mechanism (see online version for colours)



Then, the text feature H_t, the text-speech interaction feature H_T^A, and the text-picture interaction feature H_T^V are connected and mapped to a low-dimensional space. The process is as below:

$$H_m = \mathrm{ReLU}\left(W_{l1}^{mT}\left[H_t; H_T^A; H_T^V\right] + b_{l1}^m\right) \tag{14}$$

where $[\cdot; \cdot; \cdot]$ indicates a connection operation, W_{l1}^m is the weight, ReLU is the activation function, and H_m is the correlation feature of the three modal elements.

3.2 Teaching and assessment texts based on the gate control unit multimodal elements global interaction

A multimodal characteristic interaction module on a global scale is constructed through the use of a gate control unit, intended to extract the distinctive features of various modalities. This module is directed by the pertinent features of the text mode, and employs a gate control mechanism to acquire the distinctive characteristics of elements across the three modals. In this module, the output modal correlation characteristic H_m of the local modal interaction module and the speech modal characteristic H_a of the feature representation module are input into two different linear layers. The output of the two linear levels is adopted as the gating unit's unit. The multimodal correlation feature is used to filter out the unique features of the single modal elements. Taking the speech modal as an example, the entire process is as below:

$$\lambda_a = sigmoid\left(W_m H_m + W_a H_a\right) \tag{15}$$

$$H_a^* = (1-\lambda_a) * H_a \tag{16}$$

where λ_a denotes the similarity weight among the characteristics associated with multiple modalities and speech characteristics, W_m and W_a are defined as parameter matrices, and $H_\alpha^* \in R^{l_a \times d_a}$ is the characteristic specific to the speech modal. Through the repetition of the previously mentioned steps, the unique characteristics of text mode and picture mode can be derived, symbolised by $H_t^* \in R^{l_a \times d_a}$ and $H_v^* \in R^{l_a \times d_a}$, respectively.

Then, the text-specific feature H_t^*, the speech-specific feature H_α^*, and the picture-specific feature H_v^* are connected and mapped to a low-dimensional space R^{d_m}. The process is as below:

$$H_m^* = \mathrm{ReLU}\left(W_{l2}^{mT}\left[H_t^*; H_a^*, H_v^*\right] + b_{l1}^m\right) \tag{17}$$

where $W_{l2}^{mT} \in R^{(d_t+d_a+d_v)\times d_m}$ and ReLU are activation functions; H_m^* is a characteristic feature of different modes.

3.3 Feature fusion and affective tendency output based on the transformer model

After the modal-specific characteristics H_m^* are gained adopting the related characteristics guided through the text mode, in this paper, a local-global characteristic integration module is constructed based on the transformer model (Guerra and Mota, 2006) to achieve text sentiment analysis by synthesising multi-modal element features and modal importance information. First, the characteristics associated with modalities and the characteristics unique to each modality are superimposed onto matrix $M = [H_m, H_m^*] \in R^{2\times d_m}$; then matrix M is used as the input to the transformer, and based on the multi-head self-attention mechanism, each vector learns to represent

other cross-modalities. Comprehensive global multimodal features are used to achieve comprehensive sentiment analysis of texts.

In the context of the self-attention mechanism, matrix $Q = K = V = M \in R^{2\times d_m}$ is defined, and based on this, the transformer model yields a novel matrix $Q = K = V = M \in R^{2\times d_m}$, the computation process is as below:

$$Attention(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_m}}\right)V \tag{18}$$

$$head_i = Attention\left(QW_i^q, KW_i^k, VW_i^v\right) \tag{19}$$

$$\begin{aligned}\bar{M} &= MultiHead\left(M;\theta^{att}\right)\\ &= \left(head_1 \oplus \ldots \oplus head_n\right)W^o\end{aligned} \tag{20}$$

where $W_i^{q,k,v} \in R^{d_m\times d_m}$ and W^o are the linear transformation weight matrices, $\oplus$ represents splicing, and $\theta^{att} = \{W^q, W^k, W^v, W^o\}$.

Finally, the output of the transformer is obtained, the output vectors are spliced, and they are sent to the linear layer for final prediction. The structure of the linear layer is a multi-level perceptron making up of an obscured level and an output layer with z nodes, where z is the number of categories. The prediction process is as follows:

$$\tilde{Y}_i = relu\left(w_1\left(F_i^G\right)\right) \tag{21}$$

where $relu(\cdot)$ is the linear rectified activation function used in the two hidden levels, w_1 is the parameter to be trained in the hidden layer, and the softmax is used as the activation operation in the output level to obtain the model prediction result, as shown below, and w_z is the parameter of the output level.

$$Y_i = \frac{\exp\left(w_z\left(\tilde{Y}_i\right)\right)}{\sum_i \exp\left(w_z\left(\tilde{Y}_i\right)\right)} \tag{22}$$

4 Objective evaluation of English teaching from a multi-element perspective by integrating the results of text sentiment classification

4.1 Selection of evaluation indicators and weight allocation for English teaching courses

English teaching evaluation has long relied on subjective questionnaires and test scores, lacking quantitative analysis of the dynamic emotional factors in the classroom. To this end, this paper integrates the above-mentioned quantitative text sentiment analysis results into the objective evaluation of English teaching, the overall process is shown in Figure 4. First, English teaching course evaluation indicators are selected from five aspects: student satisfaction, teacher satisfaction, peer evaluation, teaching platform data, and expert evaluation. In terms of the degree of impact of different indicators on the objective evaluation results of English teaching, a method is proposed to decide the weight of each assessment indicator adopting the correlation probability of attribute-like properties, and the corresponding weights are set for each evaluation indicator. Then, the contribution weight of the text with abnormal emotional annotations to the overall English teaching score is adjusted according to the emotional tendencies of the students, and the scoring scores of all English teaching evaluation dimensions are summed up by weighted summation, so as to obtain a more objective English teaching course evaluation.

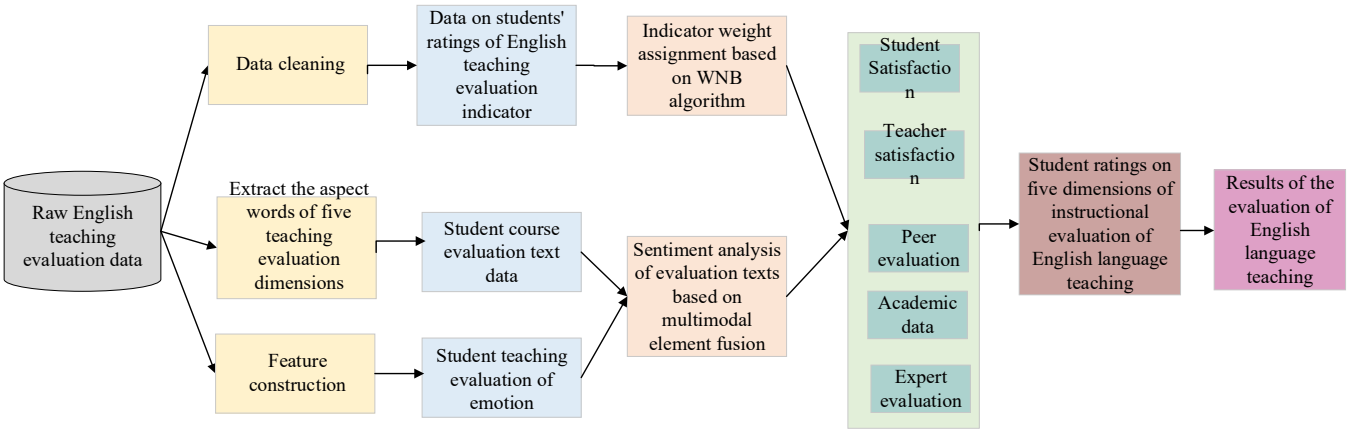
This study is based on the English teaching assessment system in existing research (Liu et al., 2021), combined with the present state of English instruction and the existing status of the evaluation metrics construction. Through summarisation and induction, the existing indicators are summarised, organised and analysed. 50 indicators such as teaching content, teaching methods, and teaching effectiveness are selected from five aspects: student satisfaction, teacher satisfaction, peer evaluation, data from the teaching platform, and expert evaluation as the evaluation indicators for English teaching courses, denoted as $x_1, x_2, \ldots, x_n$.

To reduce the calculation cost, it is assumed that each indicator is independent of each other. In this paper, WNB algorithm (Jiang et al., 2016) is adopted to allocate an appropriate weight to the indicator based on its level of contribution of the indicator to the evaluation result. It not only preserves the high-speed performance of the WNB algorithm but also mitigates the impact of the conditional independence hypothesis on the outcome. The computation equation is as below:

$$p\left(C_j \middle| X\right) = \arg\max_{C_j} p\left(C_j\right)\prod_{i=1}^{n} p\left(x_i \middle| C_j\right)^{w_i} \tag{23}$$

where w_i is the weight of x_i, it determines the degree of importance of different indicators. The larger the value of w_i, the more important x_i is. The pivotal issue in the context of WNB lies in how to ascertain the precise weight for each indicator.

Figure 4 Objective evaluation of English teaching that incorporates the results of text analysis (see online version for colours)



The correlation analysis between individual evaluation indices and the comprehensive evaluation value derived from teaching evaluation data reveals that the magnitude of influence exerted by each index on the evaluation outcome varies. Therefore, this article offers an approach of determining the weight of every assessment index through the correlation probability. x_i may be with K various values, and a_k implies a specific value, where $k \in K$. When x_i is equal to a_k, the relevant probability $p(x_i|rel)$ and the irrelevant probability $p(x_i|norel)$ of x_i regarding C_j for the evaluation result C_j are calculated as follows:

$$p\left(x_i \middle| rel\right) = \frac{count\left(x_i = a_k \wedge C_j\right)}{count\left(x_i = a_k\right)} \tag{24}$$

$$p\left(x_i \middle| norel\right) = 1 - p\left(x_i \middle| rel\right) \tag{25}$$

where *count* represents the statistical number, the weight of a_k is calculated as follows:

$$w\left(x_i, a_k, j\right) = \frac{p\left(x_i \middle| rel\right)}{p\left(x_i \middle| norel\right)} \tag{26}$$

4.2 Objective evaluation of English teaching by integrating the results of text sentiment classification

The objective evaluation method of English teaching courses from a multi-element perspective uses the results of the sentiment classification of multimodal evaluation texts to adjust the evaluation scores of students for the teaching evaluation indicators. The designed text sentiment analysis model converts student evaluations of texts into sentiment scores for English teaching evaluation indicators, defined as $B_{i,j} = \{b_{1,1}, b_{1,2}, \ldots, b_{i,j}\}$. The sentiment score for each teaching evaluation indicator ranges from 0 to 4, where 0 indicates that the text is not related to the teaching evaluation indicator, and 1–4 indicates the student's sentiment, including the four dimensions of very dissatisfied, basically satisfied, satisfied, and very satisfied. Since the designed objective evaluation method for English teaching fully considers the role of both the real student teaching score $S_{i,j}$ and the students' emotional tendencies towards different teaching dimensions in the course evaluation text, the average of the two is taken to obtain a comprehensive teaching evaluation score.

$$R_{i,j} = \begin{cases} \dfrac{S_{i,j} + b_{i,j}}{2}, & \text{if } b_{i,j} \neq 0 \\ S_{i,j}, & \text{if } b_{i,j} = 0 \end{cases} \tag{27}$$

The above equation can be used to obtain the overall score of student j for English course i in the m teaching evaluation dimensions, which is defined as $R_{i,j} = \{r_{1,1}, r_{1,2}, \ldots, r_{i,j}\}$. However, since student j's English teaching evaluation text is marked as abnormal, it has not been biased by the text sentiment analysis model to correct the actual score of English teaching, resulting in their score for the English course still being equal to $S_{i,j}$.

After calculating the weight $w(x_i, a_k, j)$ of each indicator based on Section 4.1 and the text sentiment classification model in Section 3 to obtain the comprehensive score of n indicators for the English teaching course, the scores of all students are integrated using m teaching evaluation dimensions. The final objective evaluation score of English teaching course i is as follows, where I_i is the set of all students j in course i.

$$Sum_i = \frac{\sum_{j \in I_i} R_{i,j} * w\left(x_i, a_k, j\right)}{\left|I_i\right|} \tag{28}$$

4.3 Privacy and accountability for evaluation results

In the process of building a teacher evaluation system, the dual challenges of privacy protection and bias prevention must be faced squarely. If teachers' personal data, such as teaching style, family status, and health information, are improperly collected or leaked during the evaluation process, it not only violates their basic rights, but also may lead to a crisis of trust, weakening the credibility of the evaluation process. At the same time, subjective bias in the evaluation process, whether it stems from the inherent cognitive and emotional tendencies of the evaluators or the flaws in the system design, may lead to evaluation results

that deviate from the objective facts, discourage teachers' motivation, and impede the healthy development of the education cause. Therefore, it is crucial to establish an automated evaluation system with a high degree of transparency and accountability.

The system needs to clarify the scope of data collection and use norms, and ensure that teachers can clearly understand the basis of evaluation by publicising the logic of the algorithm and the evaluation criteria; at the same time, it should build a sound accountability mechanism to pursue responsibility for the misuse of data and unfair evaluation behaviours, so as to make the evaluation process and results stand up to the supervision and inspection, effectively protect the rights and interests of teachers, and enhance the scientificity and fairness of the evaluation system.

5 Experimental results and analyses

The experimental data in this paper comes from the teaching management system of a school and contains the English classroom teaching evaluation texts of students from 2020 to 2023, a total of eight semesters. The English teaching evaluation information from the eight semesters, about 150,000 data records, is used as the initial dataset for the experiment. Labelling methods with supervised learning are used, and the data are labelled by professionals according to the criteria for assessing the quality of teaching and learning. The labelling process should follow clear standards and specifications to ensure the accuracy and consistency of the labels. For multimodal data such as audio and video, temporal alignment is achieved through timestamps or synchronisation signals to ensure that data from different modalities are synchronised in time. The processed data is randomly divided into a test set and a training set using a random split method, with the ratio of the test set and training set being 20% and 80% respectively. The simulation tool is based on pyTorch version 3.6. When training the model, the word embedding dimension of the English evaluation text is 768, the learning rate of the model is 0.01, and the batch size and epoch value during iteration are set to 64 and 30 respectively. To prevent overfitting, the regularisation penalty coefficient of the network is set to 0.01, and the initialisation parameters of the network conform to a normal distribution $U(-0.05, 0.05)$.

To more intuitively observe the changes in the quality of English teaching at the school, the dates of the observations are divided into months, and the quantitative average scores of teacher and student evaluations for each month from 2020 to 2023 are obtained, as shown in Figure 5. The quantitative average scores of teacher and student evaluations for each year are shown in Table 1. As can be seen in Figure 5, there was a sharp decline in the overall score from the end of 2022 to the beginning of 2023. As can be seen in Table 1, over time, the overall quality of English teaching at the school has also shown a downward trend. The most obvious change began in 2023, when online teaching methods were probably adopted, and the quality of English teaching has changed to a certain extent.

The proposed method TCETE is compared with the baseline methods ETLSTM (Chen and Aleem, 2024), CNN-LSTM (Aljaloud et al., 2022), SHAM (Su and Peng, 2023), LSTM-AM (Yan et al., 2023), and MSGRU (Sebbaq and El Faddouli, 2022) in terms of mean average precision (mAP) and F1 score in the five aspects of teaching attitude, teaching content, teaching methods, teaching effectiveness, and expert evaluation, as shown in Figure 6. The TCETE method has achieved better classification results than existing methods, especially a more granular affective analysis of different teaching aspects, which fully demonstrates the superiority of integrating multimodal element features in English teaching evaluation. The mAP and F1 of TCETE are 91.66% and 90.81% respectively, which is 5.14–18.66% higher than the baseline method. This shows that the TCETE method uses text modalities as a guide to achieve interaction between the three modalities, and uses text features as a guide during the interaction process. It can effectively explore the relationships within and between modalities, improve the classification accuracy, and thus improve the effectiveness of objective evaluation of English teaching.

Figure 5 Overall score of English teaching evaluation

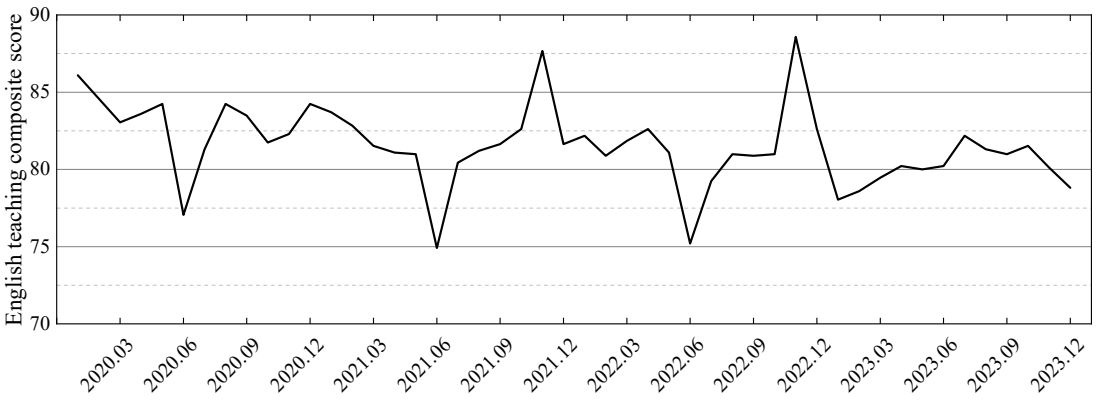


Table 1 Changes in English teaching scores from 2020 to 2023

Year	*2020*	*2021*	*2022*	*2023*
English teaching comprehensive score	80.27	79.04	78.41	76.73

Figure 6 Comparison of the performance of different teaching dimension evaluation categories (see online version for colours)

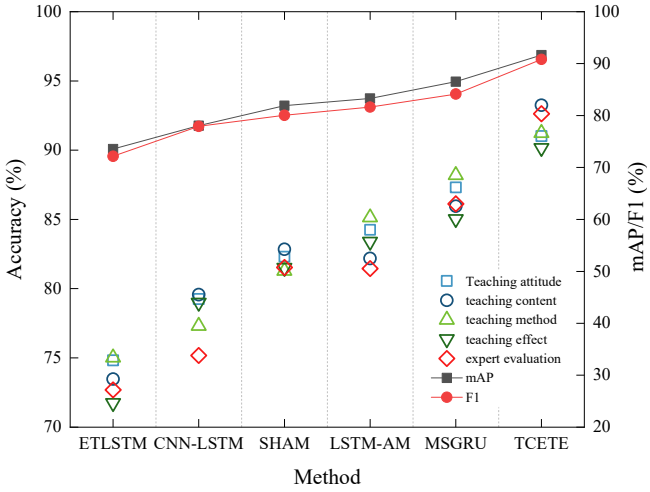


Table 2 Experimental results of ablation of each component in the TCETE method

CMAM	*GU*	*RFF*	*SFF*	*Accuracy/%*	*F1/%*
×	√	√	√	82.42	83.61
√	×	√	√	80.21	81.31
√	√	×	√	81.59	83.43
√	√	√	×	82.9	84.5
√	√	√	√	93.29	90.81

To explore in depth the key role of each component in the TCETE method, this study designed four ablation experiments to assess the specific influence of each module on the overall performance of the model. The evaluation indicators are accuracy and F1, as shown in Table 2. To make the performance indicators of the model more intuitive, this article chooses to present the data in Table 2 in the form of a bar chart. Specific results are shown in Figure 7.

1 Cross-modal attention removal (-CMAM): remove the local cross-modal interaction component from the full-fledged model.

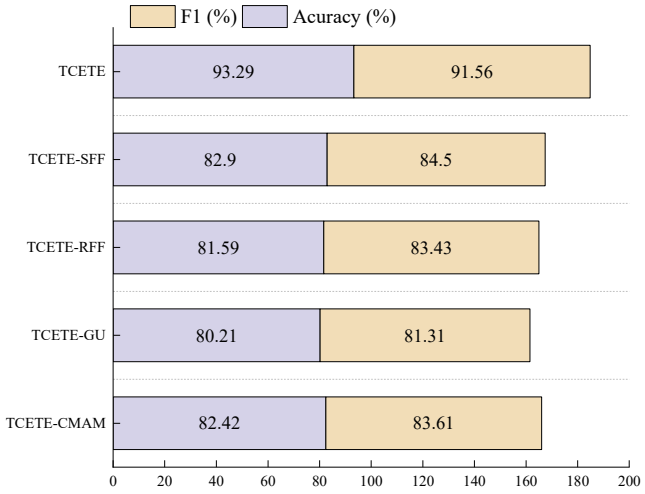
2 Remove the door control unit (-GU): deletes the global multimodal interaction component from the full-fledged model.

3 Remove related feature fusion (-RFF): within the local-global feature fusion module, modal-specific characteristics are excluded, and solely modal-related features are employed.

4 Remove specific feature fusion (-SFF): within the local-global feature fusion module, mode-dependent characteristics are excluded, and solely mode-specific features are employed.

The experimental results reveal that each component of TCETE has a vital influence on the classification performance. The accuracy of TCETE is improved by 10.87%, 13.08%, 11.7% and 10.39% compared with -CMAM, -GU, -RFF and -SFF, respectively. This shows that each part is essential to the overall performance of the model. This visualisation not only makes it easy to understand the contribution of each component of the model to the final performance, but also further confirms the rationality and effectiveness of the text classification model design suggested in this article.

Figure 7 Comparison of ablation test results (see online version for colours)



6 Conclusions

With the development of internationalisation of education, the scientific assessment of English teaching quality in higher education has emerged as a key focus in educational research. To address the issue that current English teaching evaluation methods overlook the characteristics of multimodal elements in evaluation texts, leading to inaccurate teaching evaluation results, this paper designs an objective evaluation method for college English teaching based on text analysis from a multi-element perspective. First, the three modal elements of text, audio, and image in the English teaching evaluation text are represented. Then, guided by the text mode, a CMAM is designed to represent the important information between the two modes. Key information from multiple modalities undergoes hierarchical fusion via an adaptive gating process that automatically regulates information flow. Then, based on the transformer model, the sentiment of the text is analysed by comprehensively considering the features of multimodal elements and modal importance information. The results of the analysis are integrated into an objective assessment of English teaching, and the weights of the individual

evaluation indicators are determined using the WNB algorithm. Finally, the contribution weights of emotion-annotated text segments to the overall English teaching score are adjusted based on student sentiment analysis. A weighted summation of all instructional evaluation dimension scores is then computed to derive a more objective assessment of teaching effectiveness. The experimental outcome implies that the assessment accuracy of the proposed method is 93.29%, which can achieve a more accurate evaluation of English teaching quality. In future research, this study will not further explore the semantic interaction among text mode and non-text mode in English teaching evaluation text analysis, and the dual recognition of emotion and emotion. Experiments are conducted on more English teaching evaluation datasets to validate the efficacy of the proposed approach.

Declarations

This work is supported by the Higher Education Research Project for the 14th Five-Year Plan (2024) of Guangdong Higher Education Association named: Research on the Evaluation and Enhancement of Digital Literacy among Foreign Language Teachers in Universities of Western Guangdong under the Background of Artificial Intelligence (No. 24GYB114).

The author declares that she has no conflicts of interest.

References

Aljaloud, A.S., Uliyan, D.M., Alkhalil, A., Abd Elrhman, M., Alogali, A.F.M., Altameemi, Y.M., Altamimi, M. and Kwan, P. (2022) 'A deep learning model to predict Student learning outcomes in LMS using CNN and LSTM', *IEEE Access*, Vol. 10, pp.85255–85265.

Aum, S. and Choe, S. (2021) 'srBERT: automatic article classification model for systematic review using BERT', *Systematic Reviews*, Vol. 10, pp.1–8.

Chen, C. and Aleem, M. (2024) 'A model for new media data mining and analysis in online English teaching using long short-term memory (LSTM) network', *PeerJ Computer Science*, Vol. 10, p.e1869.

Farzad, A., Mashayekhi, H. and Hassanpour, H. (2019) 'A comparative performance analysis of different activation functions in LSTM networks for classification', *Neural Computing and Applications*, Vol. 31, pp.2507–2521.

Geng, L. (2021) 'Evaluation model of college English multimedia teaching effect based on deep convolutional neural networks', *Mobile Information Systems*, Vol. 10, No. 2, pp.18–27.

Guerra, F.d.C.F. and Mota, W.S. (2006) 'Current transformer model', *IEEE Transactions on Power Delivery*, Vol. 22, No. 1, pp.187–194.

Hou, J. (2021) 'Online teaching quality evaluation model based on support vector machine and decision tree', *Journal of Intelligent & Fuzzy Systems*, Vol. 40, No. 2, pp.2193–2203.

Huang, W. (2021) 'Simulation of English teaching quality evaluation model based on Gaussian process machine learning', *Journal of Intelligent & Fuzzy Systems*, Vol. 40, No. 2, pp.2373–2383.

Jia, W. and Zhang, S. (2020) 'The unique study of talent cultivation for English major in private universities of China', *Amazonia Investiga*, Vol. 9, No. 32, pp.108–116.

Jiang, L., Li, C., Wang, S. and Zhang, L. (2016) 'Deep feature weighting for naive Bayes and its application to text classification', *Engineering Applications of Artificial Intelligence*, Vol. 52, pp.26–39.

Jiang, Y., Zhang, J. and Chen, C. (2018) 'Research on a new teaching quality evaluation method based on improved fuzzy neural network for college English', *International Journal of Continuing Engineering Education and Life Long Learning*, Vol. 28, No. 3, pp.293–309.

Li, H. (2022) 'Application of fuzzy K-Means clustering algorithm in the innovation of english teaching evaluation method', *Wireless Communications and Mobile Computing*, Vol. 22, No. 1, p.7711386.

Li, N. (2021) 'A fuzzy evaluation model of college English teaching quality based on analytic hierarchy process', *International Journal of Emerging Technologies in Learning (iJET)*, Vol. 16, No. 2, pp.17–30.

Liu, H., Chen, R., Cao, S. and Lv, H. (2021) 'Evaluation of college English teaching quality based on grey clustering analysis', *International Journal of Emerging Technologies in Learning (iJET)*, Vol. 16, No. 2, pp.173–187.

Okoye, K., Arrona-Palacios, A., Camacho-Zuñiga, C., Hammout, N., Nakamura, E.L., Escamilla, J. and Hosseini, S. (2020) 'Impact of students evaluation of teaching: a text analysis of the teachers qualities by gender', *International Journal of Educational Technology in Higher Education*, Vol. 17, No. 1, pp.1–27.

Qaiser, S. and Ali, R. (2018) 'Text mining: use of TF-IDF to examine the relevance of words to documents', *International Journal of Computer Applications*, Vol. 11, No. 1, pp.25–29.

Qi, F., Gao, Y., Wang, M., Jiang, T. and Li, Z. (2024) 'Data mining of online teaching evaluation based on deep learning', *Mathematics*, Vol. 12, No. 17, p.2692.

Qin, Y. and Irshad, A. (2024) 'Research on the evaluation method of English textbook readability based on the TextCNN model and its application in teaching design', *PeerJ Computer Science*, Vol. 10, p.e1895.

Rockoff, J.E. and Speroni, C. (2010) 'Subjective and objective evaluations of teacher effectiveness', *American Economic Review*, Vol. 100, No. 2, pp.261–266.

Sebbaq, H. and El Faddouli, N-E. (2022) 'An explainable attention-based bidirectional GRU model for pedagogical classification of MOOCs', *Interactive Technology and Smart Education*, Vol. 19, No. 4, pp.396–421.

Soydaner, D. (2022) 'Attention mechanism in neural networks: where it comes and where it goes', *Neural Computing and Applications*, Vol. 34, No. 16, pp.13371–13385.

Su, B. and Peng, J. (2023) 'Sentiment analysis of comment texts on online courses based on hierarchical attention mechanism', *Applied Sciences*, Vol. 13, No. 7, p.4204.

Yan, C., Liu, J., Liu, W. and Liu, X. (2023) 'Sentiment analysis and topic mining using a novel deep attention-based parallel dual-channel model for online course reviews', *Cognitive Computation*, Vol. 15, No. 1, pp.304–322.

Yang, Y. (2023) 'Machine learning for English teaching: a novel evaluation method', *International Journal of Computer Applications in Technology*, Vol. 71, No. 3, pp.258–264.

Zhang, Z., Gao, Q. and Chen, F. (2022) 'Evaluating English language teaching quality in classrooms using OLAP and SVM algorithms', *Mobile Information Systems*, Vol. 20, No. 11, p.9327669.