Journal of Network Intelligence ©2024 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **9**, Number **3**, August **2024**

BERT and Multilevel Data Enhancement-Based Aspect-Level Sentiment Analysis in English Language Teaching and Learning

Chun-Yan Peng*

Zhejiang Yuexiu University, Shaoxing, 312000, P. R. China
Graduate School
Lyceum of the Philippines University, Intramuros, Manila 1002, Philippines
20242031@zyufl.edu.cn

Xiang-Zhen Zhou

Zhejiang Yuexiu University, Shaoxing, 312000, P. R. China
Faculty Information Science and Technology
National University of Malaysia, Selangor 43600, Malaysia
100618@shengda.edu.cn

*Corresponding author: Chun-Yan Peng
Received February 22, 2024, revised May 29, 2024, accepted July 4, 2024.

Abstract. *Intending to the current aspect-level sentiment analysis models that cannot well characterize the deep-level word vector information and cannot address the issue of poor classification effect due to the existence of insufficient number of training samples and feature differences, this article suggests an aspect-level sentiment analysis method for English language teaching based on BERT and multilevel data enhancement. Firstly, the BERT model is enhanced adopting the contrastive learning algorithm, which strengthens the capability of word vector representation to capture semantics. Then the improved BERT model is fused as a coding layer to process the generated word vectors. Secondly, the dependency feature of the link between aspect words and sentiment words in the English teaching evaluation text is extracted using the multi-attention mechanism to enhance the performance of the aspect-level sentiment analysis task. Finally, the training samples are expanded in terms of number and features through multilevel data enhancement methods, and the context-related information is integrated to construct a novel loss function to predict the probability of different sentiment polarities. The experimental results show that the accuracy, recall, and MF1 of the proposed method are 90.5%, 91.8%, and 89.5%, respectively, and it has good performance in sentiment classification.*
Keywords: Sentiment analysis, BERT, Comparative learning, Multilevel data augmentation, Attention mechanism

1. **Introduction.** As the reform of English teaching in higher schools deepening, people are increasingly concerned about the quality of English teaching. The quality of English teaching directly affects the quality of talent cultivation [1]. In the academic affairs system of universities, a large number of English teaching evaluation texts are stored in each semester, which are difficult to be analyzed quantitatively, and it takes a lot of effort to read these texts one by one [2, 3]. Therefore, aspect-level sentiment analysis of English teaching evaluation texts can not only promote the quality of English teaching, but also contribute to academic research in the practical field [4]. By analyzing the fine-grained affective tendencies of English teaching evaluation texts, teachers can objectively

recognize their own strengths and weaknesses, and at the same time can improve their teaching methods and approaches [5, 6].

Aspect-Based Sentiment Analysis (ABSA) is a branch of sentiment analysis, which aims to identify and extract the views and emotions of specific aspects in the text. ABSA plays an important role in various natural language processing tasks, such as comment analysis, product recommendation and social media analysis.

1.1. **Related Work.** Some recent research hotspots in ABSA include ABSA based on pre-training language model, cross-language ABSA, modal ABSA and real-time ABSA. The research direction of this work is ABSA based on pre-training language model. Ngoc et al. [7] found that weighting the input information can effectively improve aspectual feature extraction. Mikolov et al. [8] suggested a continuous bag-of-words model and a word-skipping model with negative sampling to accelerate text categorization for English teaching evaluation. Song et al. [9] offered a sentiment word embedding method for manually annotated sentiment corpora, which enhances the performance of sentiment classification. Xu et al. [10] suggested an aspect-oriented sentiment classification model based on multi-attention, however it needs to capture more semantic information. Chen et al. [11] utilized the recursive attention mechanism of memory networks to efficiently retrieve sentiment information that is far away from each other. Ameer et al. [12] utilized transfer learning to transfer document-level knowledge to the aspect-level sentiment classification task, but it requires extensive training on manually labeled datasets for training.

To enhance the efficiency of aspect-level sentiment classification, researchers embedded the attention mechanism [13] into BERT [14] and Transformer [15] and applied it to the ABSA domain, which greatly enhanced the model's effectiveness. Fu et al. [16] adopted LSTM with attention mechanism to model sentences and enhance the classification performance of English teaching text. Singh et al. [17] adopted a generative data augmentation approach with a Bert pre-trained language model to generate coherent text data on large-scale data, but it was not efficient. Pota et al. [18] suggested a BERT model for sentiment analysis in English language teaching and learning, but it did not extract enough semantic information, which led to a poor classification performance. Kardakis et al. [19] suggested an ABSA model in English language teaching that incorporates bilateral attention mechanisms, incorporating contextual word cascades for categorization, but it requires access to a large amount of labeled data, which is costly.

Data augmentation is an effective technique to alleviate data deficiencies and has been used in the domain of ABSA to augment existing labeled data and to improve deep network performance. Goldberg [20] investigated an approach to augmenting textual data for EFL assessment with vocabulary substitution, but with altered text labels. Li et al. [21] introduced successive word vectors to augment the text for a multi-categorization task, which guaranteed augmentation with data diversity but lose aspect words. Kang et al. [22] implement a knowledge-based data enhancement method to improve deep learning model performance. Bacco et al. [23] experimentally demonstrated that some popular data enhancement techniques do not consistently improve the performance of text categorization tasks based on Transformers pre-trained models and are used for feature differences between samples, resulting in low classification accuracy. Multi-level data enhancement is a data enhancement technology, which improves the generalization ability and robustness of the model by enhancing data at multiple levels. Multi-level data enhancement can be applied to various data types, including text, images, audio and video.

1.2. **Contribution.** Focusing on the issue that the present models do not have enough training samples and cannot comprehensively characterize the deep-level word vector information, which leads to low classification accuracy, this article designs an aspect-level sentiment analysis method based on BERT and multilevel data augmentation for English teaching. The basic idea of multilevel data enhancement is to perform different enhancement operations on data at different levels, thus generating more diverse and representative data samples. For example, in image data enhancement, images can be enhanced at pixel level, feature level and semantic level.

Firstly, to address the anisotropy problem of the pre-trained BERT model, the contrast learning algorithm is used to improve the BERT model (MD-IBERT), so that the processed word vectors are more uniformly distributed in the word representation space. Then the IBERT model is fused to generate word vectors.

Secondly, the dependency feature of the connection between aspect words and sentiment words in the English teaching evaluation text is extracted adopting the multi-attention mechanism to enhance the performance of the aspect-level sentiment analysis task. Finally, the training samples are enlarged in terms of number and features through multilevel data enhancement, and a novel loss function is constructed to predict the probability of different sentiment polarities by integrating context and target-related information. Multilevel data enhancement can effectively improve the generalization ability and robustness of the model. This is because multilevel data enhancement can generate more diverse and representative data samples, so that the model can better learn the inherent laws of data and reduce the risk of over-fitting. In addition, multi-level data enhancement can help the model deal with noise and outliers better.

2. Theoretical analysis.

2.1. **BERT model.** The BERT model mainly adopts the decoder of the bidirectional Transformer as the main model feature extraction structure [24], and its framework is indicated in Figure 1.

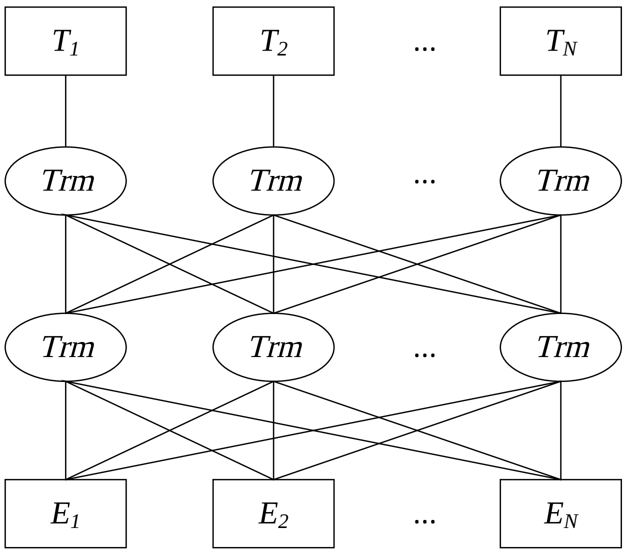

Figure 1. The framework of the BERT model

In Figure 1, E denotes the corresponding Token encoding in the text sequence, Trm denotes the Transformer encoder, and T denotes the encoding of the trained target Token. The BERT model divides the input part into Word Vector Embedding, Segment

Embedding, and Position Embedding. Among them, Word Vector Embedding is used to transform text data into structured data that can be learned by a machine; Segment Embedding mainly splits the context sentence into two or more segments, which is usually adopted to determine whether different words belong to the same sentence. Position Embedding is used to encode the position information of the whole tokens in the text sequence into a feature vector. The calculation of positional encoding is indicated in Equation (1) and Equation (2).

$$PE(pos, 2j) = \sin\left(\frac{pos}{10000^{\frac{2j}{d_{mod}}}}\right) \tag{1}$$

$$PE\left(\frac{pos}{pos, 2j+1}\right) = \cos\left(\frac{pos}{10000^{\frac{2j}{d_{mod}}}}\right) \tag{2}$$

where pos represents the position number and d_{mod} represents the input sequence length.

2.2. **Text multi-Level data enhancement method.** Easy Data Augmentation (EDA) is a multilevel data augmentation technique applied in text categorization tasks [25]. The method mainly realizes text data augmentation by introducing four atomic operations for word changes in statements, as bellow.

(1) Synonym Replacement (SR): N non-deactivated words are randomly chosen from the utterance and then synonym replacement is performed on these words.

(2) Random Insertion (RI): Randomly select a non-deactivated word from the statement and find its near-synonym, insert the near-synonym randomly into any position in the sentence, and repeat the operation N times.

(3) Random Swap (RS): Choose two words at random from the statement and exchange their positions, and repeat the operation N times.

(4) Random Deletion (RD): Randomly delete a word in a sentence with probability p.

Each of the above atomic operations can independently generate a new statement. To form as many novel features as possible without changing the original semantics, the amount of operations m required to generate a new statement is determined by the length of the sentence. Assuming that the length of the sentence is k and the probability of the operation for each statement is η, the amount m of atomic operations required for each generated sentence can be denoted as $m = \eta * k$.

3. **Improvement of the BERT model.** Focusing on the anisotropy issue of pre-trained BERT models, this article suggests a novel method (IBERT) to enhance pre-trained BERT models by adopting contrast learning algorithms [25], which strengthens the ability of word vector representations to capture the semantics, and improves the performance of various tasks efficiently. The IBERT model is divided into the BERT coding level, contrast loss level, and classification output level, as indicated in Figure 2.

(1) **BERT coding level**: Two BERT models are adopted and they share parameters during training. $BERT_b$ uses the [CLS] bit to output the vector representation of the whole sentence, and $BERT_a$ calculates the average of the output vectors of each token for the vector representation, and finally obtains the two representations s_j and s_j^+ of the same sample sentence C, which effectively avoids the problem of the gradual degradation of the training signals of BERT in the course of training.

Given n sentences in the text, denoted as $\{c_1, c_2, ..., c_n\}$, each sentence c_m is entered into BERT and the character-level hidden representation $H_{m,j} \in \mathbb{R}^{le(c_m)\times v}$ is computed as indicated in Equation (3).

$$[H_{m,0}, H_{m,1}, ..., H_{m,l}, ..., H_{m,k}] = BERT(c_m) \tag{3}$$

where $0 \leq l \leq k$, k is the amount of obscured level in the BERT model, $le(c_m)$ is the length of the sentence, and v is the dimension of the obscured level of the BERT model. $h_{m,j} = p(H_{m,j})$ is the sentence-level characteristic $h_{m,j} \in \mathbb{R}^r$ computed by applying the pooling function p to each level $H_{m,j}$. In this case, $BERT_b$ directly adopts the output vector of the [CLS] bits to represent the vector representation of the entire sentence, and $BERT_a$ adopts the average pooling strategy.

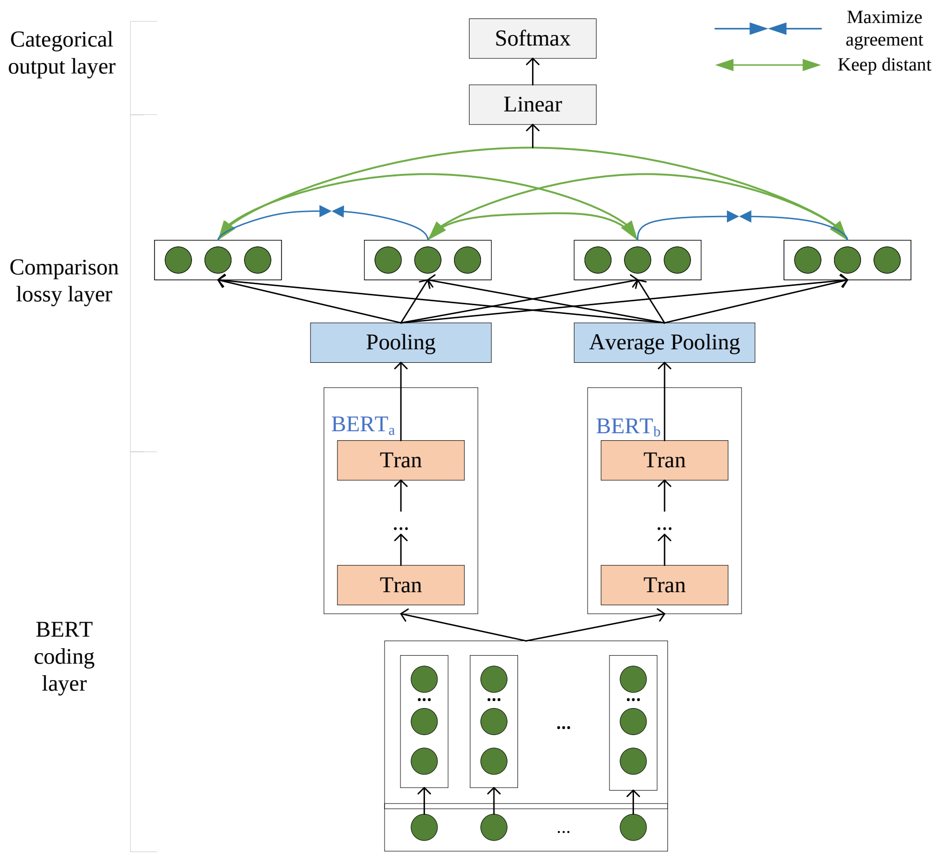

Figure 2. The structure of the IBERT model

(2) Contrastive loss level. The main strategy of the contrast learning construction is to design the average of the vector representation of the [CLS] bits of the same sample and the rest of the token embeddings of the same sample as positive samples, and the rest of the samples within the same text as negative samples, so as to bring the positive samples closer to the negative samples, and push them away from the negative samples. The objective function is indicated below.

$$L_j = -\log \frac{\exp\left(sim\left(s_j, s_j^+\right)/\sigma\right)}{\sum_{i=0}^{M} \exp\left(sim\left(s_j, s_i\right)/\sigma\right)} \tag{4}$$

where sim function is the cosine similarity function, σ is the temperature coefficient, and s is the corresponding vector representation. In Equation (4), the minimization of the guarantee L_j is to minimize the similarity between the positive and negative example pairs.

(3) Categorize the output level. Classify the sentiment polarity of the input text into two categories: positive and negative. Predictions are made using a fully connected layer,

and the layer is normalized using a softmax activation function to produce a sentiment probability distribution p.

$$p(a) = softmax(W_p s + b_p) \tag{5}$$

where W_p and b_p are learnable weights and biases, respectively.

Then the standard gradient descent algorithm is adopted for training, and the objective function adopted is cross entropy loss, as indicated below.

$$Loss = -\sum_{(b,a)\in C}\sum_{\hat{p}\in Q}\log p(a) + \mu\left\|\vartheta\right\|_2 \tag{6}$$

where C represents the set of sentence aspect pairs, Q represents the set of different emotional polarities, $\hat{p}$ represents different labels, μ represents regularization coefficients, and ϑ denotes the entire trainable model parameters.

4. **Multilevel data enhancement-based aspect- layer sentiment analysis in English language teaching and learning.** On the ground of IBERT, this article designs an aspect-level sentiment analysis model (MD-IBERT) relied on multilevel data enhancement. This paper constructs a feature extractor relied on multi-head self-attention and full connection of location correlation to extract the characteristics of aspect words and emotion words in English teaching evaluation texts, makes full use of the various weights obtained by the attention mechanism to analyze the interaction between aspects, and expands the number and features of training samples by using data enhancement methods to fully integrate sentence features and aspect words. The whole framework is indicated in Figure 3.

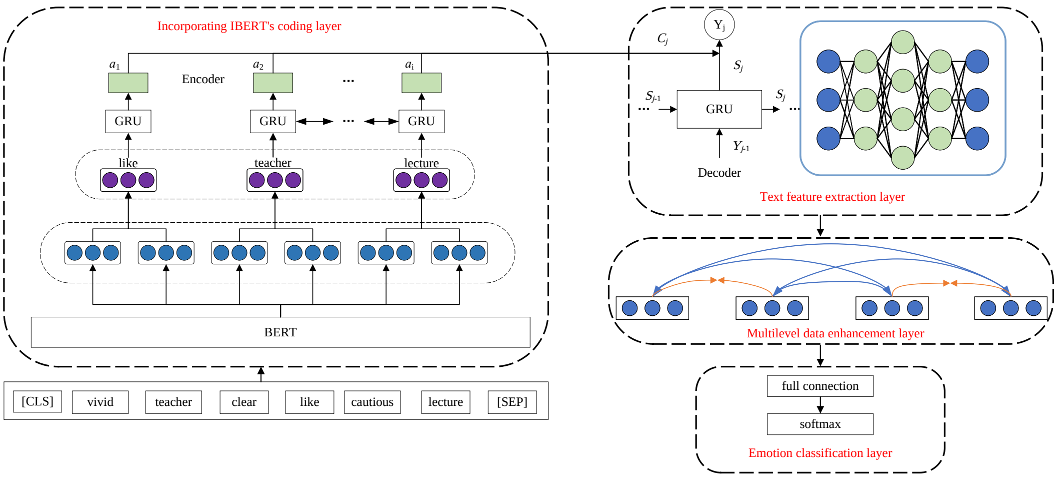

Figure 3. The whole framework of the suggested model

4.1. **Fusion IBERT coding layer.** Firstly, the English teaching evaluation text sequence is input into the IBERT model, and the word vector is output after being constructed by two-way Transformer. Then, the word vector is fused according to the predetermined word boundaries, and the fusion strategy is mainly relied on the "sum and average of all words", which is shown as bellow:

$$U_i = \frac{(x_1 + x_2 + \cdots + x_m)}{m} \tag{7}$$

where x_i represents each word vector that makes up the word vector, and U_i represents the word vector representation achieved from the sum average of all word vectors.

Secondly, the word vector output by IBERT is transferred to the bidirectional recurrent neural network as the input sequence of bidirectional gated cyclic unit (BiGRU), and the word vector is further learned. The fusion word vector trained by IBERT is represented as:

$U^i = (U_1, U_2, ..., U_m)$, where U^i represents the vector matrix of the i-th statement, U_i represents the vector sign of the i-th word, and m represents the maximum sentence length.

Then the IBERT word vectors of the sequence are mapped to the forward and backward obscured states through $\vec{h}_k = f_{GRU}(u_i, h_{i-1})$ and $\overleftarrow{h}_i = f_{GRU}(u_i, h_{i+1})$ respectively, and all the obscured states of the forward and backward GRUs are concatenated and combined as the output $h_{ik} = [\vec{h}_k; \overleftarrow{h}_i]$ of the GRU.

4.2. **Text feature extraction layer of English teaching evaluation.** To extract the dependencies between aspect words and sentiment words among sentences. First, a sequence parser is used to obtain the original sentence hidden state, and then the dependency features of the link between aspect words and sentiment words are extracted based on the attention mechanism to improve the performance of the aspect-level sentiment analysis task.

The encoder output h_k and the subject distribution ϑ from the variational autoencoder, the sequence decoder adopts a one-way GRU level, and resolved dependencies are as follows.

$$S_i = f_{GRU}([y_i; \vartheta], s_{i-1}) \tag{8}$$

where y_i is the input of the decoder at moment i, and s_{i-1} is the obscured state of the previous moment.

The achieved dependency relationship is represented by graph G with m nodes (words), where the edge of graph G represents the dependency relationship between each word, and the neighborhood nodes of node i are indicated in M_i. The neighborhood nodes are aggregated one by one using multi-head attention, and the iterative update of the representation of each node is finally realized. The calculation in detail is as follows.

$$h_{att_i}^{l+1} = \Big\|_{k=1}^{K} \sum_{j \in M_i} \beta_{ij}^{lk} V_k^l h_j^l \tag{9}$$

where h_{att}^{l+1} is the attention head of node i at level $l+1$, $\|_{k=1}^{K}$ is the vector concatenation operation from x_1 to x_k, $\beta_{ij}^{l,k}$ is the normalized attention coefficient calculated for the k-th attention of level l, and V_k^l is the input transformation matrix.

Because the neighboring nodes with different dependencies have various effects, additional relational headers are adopted to control the information flow of the neighboring nodes. First, the dependency relationship is mapped to a vector representation, and then a relationship header is calculated by Equation (10).

$$h_r^{l+1} = \Big\| \begin{matrix} N \\ n=1 \end{matrix} \quad \left(exp(g_{ij}^{lm}) \Big/ \sum\nolimits_{j=1}^{M_i} exp(g_{ij}^{ln})\right) V_m^l h_j^l \tag{10}$$

where $g_{ij}^{ln} = \vartheta(relu(s_{ij}V_{n1} + a_{n1})V_{n2} + b_{n2})$, s_{ij} is relational embeddings between nodes i and j, a represents aspect words, and the final representation of a single node is as follows.

$$h_i^{l+1} = relu(V_{l+1}(h_{att_i}^{l+1} \parallel h_{r_i}^{l+1}) + a_{l+1}) \tag{11}$$

Finally, the attention mechanism can be stacked to obtain the input text feature representation $\{\hat{h}_0, \hat{h}_1, \ldots, \hat{h}_m\}$, and pooled to achieve the final characteristics representation as $h_i = pool(h_0, h_1, \ldots, h_m)$.

4.3. **English teaching evaluation text multi-level data enhancement layer.** To extract characteristics without changing the semantics of the utterance, the training samples need to be expanded in terms of number and features through a multilevel data augmentation EDA approach as below.

Suppose a text contains a sentence $C = [w_0, w_1, \ldots, w_{m-1}]$ of m words, where w_q denotes the q-th word in the sentence, and w_q has an information gain value of $IG(w_q)$. Then the utterance can be represented as a weight vector W consisting of the information gain values of each word.

$$W = [IG(w_0), IG(w_1), \ldots, IG(w_{m-1})] \tag{12}$$

Since some of the high-frequency feature words with categorical attributes in the data are prone to be chosen repeatedly when performing operations such as synonymous substitution, it is necessary to perform smoothing on W. The new weights for each word in W are computed as indicated below.

$$W[q] = \sqrt{W[q]}, q \in \{0, 1, \ldots, m-1\} \tag{13}$$

In this article, the roulette algorithm is adopted to choose words to be synonymy replaced. The probability of word w_i being selected for synonymy replacement is as below.

$$P_{sr}(w_i) = \frac{W[i]}{\sum_{j=0}^{m-1} W[j]} \tag{14}$$

For deletion operation, the roulette algorithm [27] is also adopted to choose the words to be operated, but the probability of deletion of the words is negatively correlated with the information gain value, so the probability of w_i being chosen for deletion is as below.

$$P_{rd}(w_i) = \frac{\max(W) - W[i]}{\sum_{j=0}^{m-1}(\max(W) - W[j])} \tag{15}$$

The resulting new data set is denoted as below.

$$Sen = \lambda * Sen_1 + \mu * Sen_2, \lambda + \mu = 1 \tag{16}$$

where the statement set generated by word EDA is Sen_1 respectively, and the same number of statement sets generated by sentence EDA is Sen_2. Some data are randomly selected from each of the two sets to form a new data set in the proportion of λ and μ, and the enhancement feature is $h_i^* = sen \| pool(h_0, h_1, \ldots, h_m)$.

4.4. **English teaching aspect level emotion classification layer.** The classification output layer finally processes and outputs the enhanced English teaching evaluation text characteristic h_i^*. After applying attention mechanisms on aspect-oriented dependencies, the probability of passing through a fully connected softmax level and output different emotional polarities.

$$p(a) = softmax(W_p h_i^* + b_p) \tag{17}$$

Since softmax function is adopted in the output level, the loss function in this article is a multi-class cross-entropy loss function. The computation formula is as bellow.

$$L_2 = -\sum_i c_i \log(y_i) \tag{18}$$

where c_i denotes the real emotion category of the data, and y_i represents the emotion category predicted by softmax classifier. The model training purpose is to reduce losses and optimize results through learning.

The performance of the model depends on the relative weight between the losses of each task. This paper combines the loss of the score generated by aspect words and the loss of the score generated by emotion labels as the loss function of the joint task.

$$Loss = \rho_1 L_1 + \rho_2 L_2 \tag{19}$$

where ρ_1 and ρ_2 are the weight coefficients of the loss, and $\rho_1 + \rho_2 = 1$.

5. Performance testing and analysis.

5.1. **Experimental results and analysis.** This article adopts a combination of Accuracy (Acc), Recall (Rec) and Macro-F1 (MF1) [29] to assess the effectiveness of emotion categorization at the English language teaching aspect level. Table 1 demonstrates the comparative outcome of the various models for each metric.

Table 1. Performance comparison outcome for various models

Model	Acc	Rec	MF1
ME-BERT	0.841	0.829	0.836
SE-ABSA	0.694	0.713	0.704
FL-ABSA	0.782	0.791	0.784
MD-IBERT	0.905	0.918	0.895

The analysis yielded that the MD-IBERT model improved on all three evaluation metrics, Acc, Rec, and MF1, compared to the three aspect-level sentiment analysis-based models, ME-BERT, SE-ABSA, and FL-ABSA. It increased by 7.61%, 30.4% and 15.73% in Acc, 10.74%, 28.75% and 16.06% in Rec, and 7.06%, 27.13% and 14.16% in MF1. This model always outperforms the comparison model in three indicators, because MD-IBERT improves the BERT model, which can adaptively model words according to contextual words and effectively extract semantic information. In addition, MD-IBERT improves the classification ability of the model through multi-level data enhancement by separating some isolated utterances that are not easy to determine the sentiment polarity from other utterances that contain such content and adding them to the training set.

To validate the performance of different approaches to capture long-term dependencies, validation experiments were constructed in different lengths of English language teaching assessment texts. As indicated in Figure 4, when the text length is 50, the MF1 of ME-BERT, SE-ABSA, FL-ABSA, and MD-IBERT are 0.83, 0.72, 0.76, and 0.87, respectively, and MD-IBERT is enhanced in the comprehensive performance index MF1 compared with ME-BERT, SE-ABSA, and FL-ABSA models. This suggests that the transformer encoder learns the hidden relationships between contexts better and improves the performance of aspect-level sentiment classification by extracting the dependency features of the associations between aspect words and sentiment words relied on the attention mechanism.

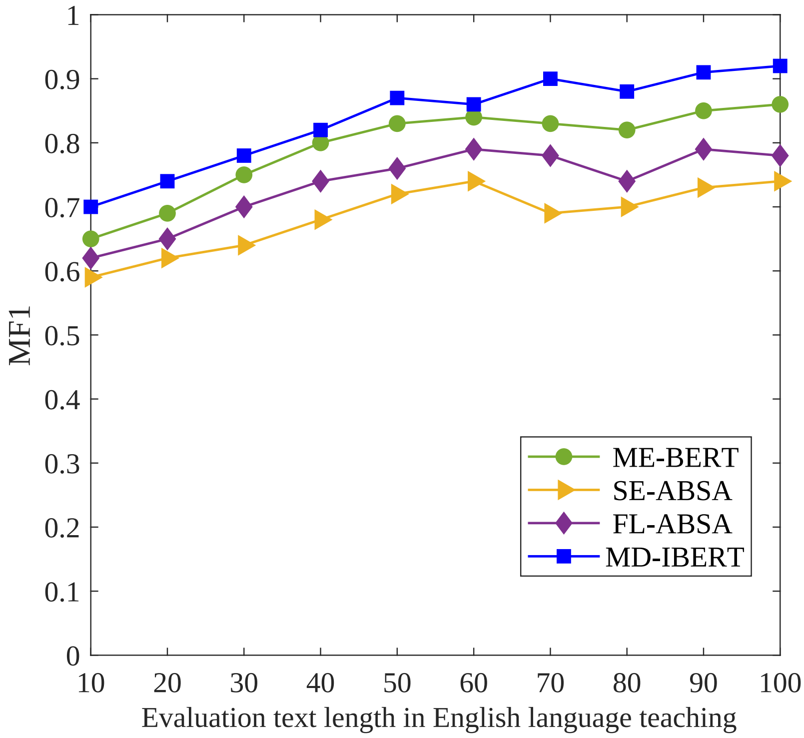

Figure 4. Comparison of MF1 across models for different lengths of evaluation texts

5.2. **Ablation experiment.** To better validate the impact of each component of the model MD-IBERT in this paper, ablation experiments were conducted on the dataset and four comparison models were designed for comparative analysis.

(1) The coding level in MD-IBERT is removed and the other layers remain unchanged, noted as MD-IBERT/CD.

(2) Remove the text characteristic extraction level from MD-IBERT and leave the other layers unchanged, denoted as MD-IBERT/FT.

(3) Remove the textual multilevel data enhancement level from MD-IBERT, leaving the other layers unchanged, denoted as MD-IBERT/DE.

(4) Remove the attention module of the emotion categorization level in MD-IBERT and leave the other levels unchanged, denoted MD-IBERT/EC.

The outcome of the model evaluation metrics is given in Table 2, and the outcome are plotted on a visual bar comparison chart, as indicated in Figure 5. Removing the coding layer CD reduces Acc, Rec and MF1 of MD-IBERT by 8.38%, 9.03% and 8.22% respectively. This indicates that the coding information between aspects and context words in a sentence is not negligible. Removing either the textual characteristic extraction level FT or the sentiment categorization level attention module EC leads to different degrees of accuracy degradation. In the case of removing the text multilevel data enhancement level DE, the Acc, Rec, and MF1 of MD-IBERT decreased by 7.61%, 9.94%, and 6.67%, respectively, which indicates that the multilevel data enhancement has a greater impact on the MD-IBERT model, and side by side reflects the necessity of data enhancement. Overall, MD-IBERT with all the modules integrated achieves the best performance.

6. **Conclusion.** Focusing on the current ABSA models that cannot well characterize the deep-level word vector information and cannot deal with the issue of low classification

Table 2. Remove experimental outcome for various modules

Model	Acc	Rec	MF1
MD-IBERT/CD	0.835	0.842	0.827
MD-IBERT/FT	0.874	0.861	0.852
MD-IBERT/DE	0.841	0.835	0.839
MD-IBERT/EC	0.882	0.876	0.861
MD-IBERT	0.905	0.918	0.895

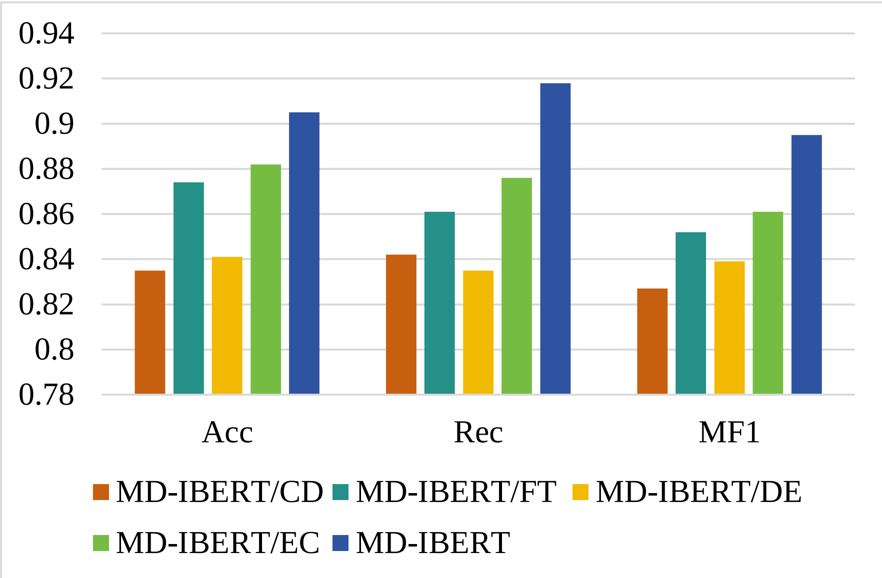

Figure 5. Comparison of classification accuracy of different models

accuracy owing to the insufficient amount of training samples, this article suggests an ABSA method for English teaching relied on BERT and multilevel data enhancement. Firstly, to address the anisotropy issue of the pre-trained BERT model, the BERT model is enhanced by adopting the comparative learning algorithm, which makes the processed word vector representations more isotropic. Then the word vectors are fused relied on the improved BERT model and processed to generate word vectors. Next, the word vectors are decoded using a sequence parser, and the dependency feature of the link between aspect words and sentiment words in the English teaching evaluation text is extracted using the multi-head attention mechanism to improve the performance of the aspect-level sentiment analysis task. Finally, the training samples are enlarged in terms of number and features through multilevel data enhancement, and a novel loss function is constructed to predict the probability of different sentiment polarities by integrating context- and target-related information. The experimental outcome indicates that the suggested method can effectively enhance the accuracy, recall and MF1 of aspectual sentiment analysis in English teaching.

REFERENCES

[1] D. R. Cole, "The power of emotional factors in English teaching," *Power and Education*, vol. 1, no. 1, pp. 57-70, 2009.

[2] M. Pawlak, J. Zawodniak, and M. Kruk, "The neglected emotion of boredom in teaching English to advanced learners," *International Journal of Applied Linguistics*, vol. 30, no. 3, pp. 497-509, 2020.

[3] L. Unsworth, and K. A. Mills, "English language teaching of attitude and emotion in digital multimodal composition," *Journal of Second Language Writing*, vol. 47, pp. 100712, 2020.

[4] A. C. Wardhani, O. T. Ena, and C. L. Mbato, "Constructing teacher identity: Teachers emotion in teaching English for pre-school students," *Language Circle: Journal of Language and Literature*, vol. 15, no. 2, pp. 193-200, 2021.

[5] J. C. Richards, "Exploring emotions in language teaching," *Relc Journal*, vol. 53, no. 1, pp. 225-239, 2022.

[6] F. Su, "Research on Integration of Emotion Analysis in English Modular Teaching Based on Natural Language Processing," *Frontiers in Psychology*, vol. 13, pp. 928883, 2022.

[7] P. V. Ngoc, C. V. T. Ngoc, T. V. T. Ngoc, and D. N. Duy, "A C4. 5 algorithm for English emotional classification," *Evolving Systems*, vol. 10, no. 3, pp. 425-451, 2019.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[9] M. Song, H. Park, and K.-S. Shin, "Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean," *Information Processing & Management*, vol. 56, no. 3, pp. 637-653, 2019.

[10] Q. Xu, L. Zhu, T. Dai, and C. Yan, "Aspect-based sentiment classification with multi-attention network," *Neurocomputing*, vol. 388, pp. 135-143, 2020.

[11] T. Chen, H. Yin, X. Yuan, Y. Gu, F. Ren, and X. Sun, "Emotion recognition based on fusion of long short-term memory networks and SVMs," *Digital Signal Processing*, vol. 117, pp. 103153, 2021.

[12] I. Ameer, N. Bölücü, M. H. F. Siddiqui, B. Can, G. Sidorov, and A. Gelbukh, "Multi-label emotion classification in texts using transfer learning," *Expert Systems with Applications*, vol. 213, pp. 118534, 2023.

[13] J.-W. Liu, and X.-l. LUO, "Research progress in attention mechanism in deep learning," *Chinese Journal of Engineering*, vol. 43, no. 11, pp. 1499-1511, 2021.

[14] P. Ganesh, Y. Chen, X. Lou, M. A. Khan, Y. Yang, H. Sajjad, P. Nakov, D. Chen, and M. Winslett, "Compressing large-scale transformer-based models: A case study on BERT," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1061-1080, 2021.

[15] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15908-15919, 2021.

[16] X. Fu, J. Yang, J. Li, M. Fang, and H. Wang, "Lexicon-enhanced LSTM with attention for general sentiment analysis," *IEEE Access*, vol. 6, pp. 71884-71891, 2018.

[17] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 33, 2021.

[18] M. Pota, M. Ventura, H. Fujita, and M. Esposito, "Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets," *Expert Systems with Applications*, vol. 181, pp. 115119, 2021.

[19] S. Kardakis, I. Perikos, F. Grivokostopoulou, and I. Hatzilygeroudis, "Examining attention mechanisms in deep learning models for sentiment analysis," *Applied Sciences*, vol. 11, no. 9, pp. 3883, 2021.

[20] Y. Goldberg, "Neural network methods for natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. 10, no. 1, pp. 1-309, 2017.

[21] Q. Li, X. Li, Y. Du, Y. Fan, and X. Chen, "A New Sentiment-Enhanced Word Embedding Method for Sentiment Analysis," *Applied Sciences*, vol. 12, no. 20, pp. 10236, 2022.

[22] T. Kang, A. Perotte, Y. Tang, C. Ta, and C. Weng, "UMLS-based data augmentation for natural language processing of clinical research literature," *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 812-823, 2021.

[23] L. Bacco, A. Cimino, F. Dell'Orletta, and M. Merone, "Explainable sentiment analysis: a hierarchical transformer-based extractive summarization approach," *Electronics*, vol. 10, no. 18, pp. 2195, 2021.

[24] A. S. Alammary, "BERT models for Arabic text classification: a systematic review," *Applied Sciences*, vol. 12, no. 11, 5720, 2022.

[25] D. A. Van Dyk, and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1-50, 2001.

[26] Y. G. Jang, H. Y. Kim, and M. Yi, "A color contrast algorithm for e-learning standard," *International Journal of Computer Science and Network Security*, vol. 7, no. 4, pp. 195-201, 2007.

[27] K. Asghari, M. Masdari, F. S. Gharehchopogh, and R. Saneifard, "Multi-swarm and chaotic whale-particle swarm optimization algorithm with a selection method based on roulette wheel," *Expert Systems*, vol. 38, no. 8, e12779, 2021.

[28] L. Li, "Emotion recognition of foreign language teachers in college English classroom teaching," *Frontiers in Psychology*, vol. 12, 788552, 2021.

[29] C. Li, Y. Qi, X. Ding, J. Zhao, T. Sang, and M. Lee, "A deep learning method approach for sleep stage classification with eeg spectrogram," *International Journal of Environmental Research and Public Health*, vol. 19, no. 10, 6322, 2022.