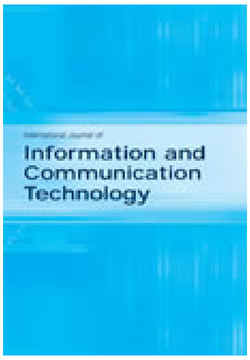


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Sentiment feedback for English writing using BERT-LSTM hybrid modelling

Yizhou He

DOI: 10.1504/IJICT.2025.10074949

Article History:

Received:	18 August 2025
Last revised:	29 October 2025
Accepted:	30 October 2025
Published online:	17 December 2025

Sentiment feedback for English writing using BERT-LSTM hybrid modelling

Yizhou He

School of Humanity,
Xinyu University,
Xinyu, 338004, China
Email: hyz969322@126.com

Abstract: Addressing the challenges of subjectivity and delayed feedback in English writing sentiment analysis, particularly for non-native academic contexts, this study proposes an automated framework based on a hybrid bidirectional encoder representations from transformers-long short-term memory model. The model integrates Bert's contextual encoding capabilities with long short-term memory's sequential modelling strengths, leveraging attention mechanisms for three-dimensional granular sentiment analysis (tendency-intensity-object). Evaluated on a 2020–2023 subset of the international English language testing system writing dataset, the model achieves a sentiment classification accuracy of 91.5% (F1-score 0.89), outperforming baseline models by 12.3%. Educational applicability testing showed an 87.2% teacher approval for reducing feedback workload, confirming its efficacy in supporting writing pedagogy decisions. Key innovations include a multi-task with domain adaptation and a visualised feedback system, establishing a new paradigm for intelligent educational tool with real-time intervention capabilities.

Keywords: sentiment analysis; English writing assessment; BERT-LSTM hybrid model; attention mechanism; educational feedback system.

Reference to this paper should be made as follows: He, Y. (2025) 'Sentiment feedback for English writing using BERT-LSTM hybrid modelling', *Int. J. Information and Communication Technology*, Vol. 26, No. 47, pp.20–35.

Biographical notes: Yizhou He received his Master's degree from the Xiangtan University in 2009. He currently serves as an Associate Professor at the School of Humanities, Xinyu University. His research interests encompass foreign language literature, foreign language teaching and sentiment analysis.

1 Introduction

As a core skill for global academic and professional communication, the assessment of English writing ability in teaching and learning has long relied on superficial characterisation such as grammatical accuracy and vocabulary richness, while neglecting the deeper impact of affective dimensions, (e.g., motivation and attitudinal dispositions) on the writing process. Studies have shown that the lack of affective feedback may cause up to 23% of students to develop writing anxiety, which in turn inhibits expressive and creative thinking. Traditional assessment methods rely on teachers' subjective judgements, which have the limitations of poor real-time performance and rough granularity, making it difficult to capture the complex emotional expression patterns of non-native speakers, (e.g., 'hidden frustration', 'ambivalent attitudes', 'frustration' and 'ambivalence'), as well as the complex emotional expression patterns of non-native speakers (Matzinger et al., 2020) ambivalent attitudes. With the popularisation of online education (Geng and Yan, 2021), especially the normalisation of distance learning in the postepidemic era, the construction of an automated and fine-grained affective feedback mechanism has become an urgent need to improve the effectiveness of English writing teaching (Zheng and Zhang, 2025).

In recent years, breakthroughs in natural language processing technologies have provided new paradigms for text sentiment analysis. Deep neural network models have made significant progress in social media sentiment categorisation by sequentially modelling local semantic features (Srinidhi et al., 2021), e.g., 89.47% accuracy in sentiment categorisation for Coronavirus disease 2019 tweets (Kaur et al., 2021). However, such models face a triple challenge in:

1 A context-dependent problem: complex expressions such as irony and metaphor, (e.g., 'this brilliant idea ruined my weekend') are common in academic texts, and it is difficult for static word vector models to capture the semantic shift of sentiment in dynamic contexts.

2 Low sentiment density: compared to social media, sentiment signals in writing models are more subtle and sparsely distributed, with less than 40% coverage in general-purpose sentiment lexicons.

3 Cross-linguistic interference: grammatical errors and cultural expression differences, (e.g., East Asian students' habitual use of indirect negativity) in non-native writing further reduce analysis accuracy.

To address the above challenges, pre-trained language models and multimodal fusion have become the frontiers of current research (Chen et al., 2024). Bidirectional encoder representations from transformers generates dynamic contextual word vectors through the bi-directional transformer architecture (Zalte and Shah, 2024), which significantly improves the ability to disambiguate the sentiment of polysemous words. Significantly improves the sentiment disambiguation of polysemous words, driving the F1 value to 82.6% in the semeval competition. Meanwhile, the hybrid model architecture, by combining the context encoding of bidirectional encoder representations from transformers with the temporal modelling advantages of long short-term memory (Zalte and Shah, 2024), performs well in the sentiment analysis of long texts-for example, bidirectional encoder representations from transformers – long short-term memory improves the accuracy to 92.6% in the coronavirus disease 2019 tweets categorisation task bidirectional encoder representations from transformers – long short-term memory improves the accuracy to 92.10% in the coronavirus disease-19 tweet classification task. Recent research further explores multimodal sentiment recognition, (e.g., fusion of facial expression, speech, and text) (Das and Singh, 2023), in which all-modalities-in-one bidirectional encoder representations from transformer employs a single-stream transformer to integrate the three modal features and achieves cross-modal alignment on Carnegie Mellon University multimodal opinion sentiment and emotion intensity datasets, providing new ideas for fine-grained sentiment analysis to provide new ideas. However, there are still bottlenecks in the application of such models in educational scenarios: multimodal data are costly to acquire and do not cover text-only writing scenarios; and existing research focuses on coarse-grained sentiment polarity categorisation (positive/negative) (Wang et al., 2020), and lacks co-modelling of sentiment intensity (weak-medium-strong) and sentiment object (to the topic/to one's own abilities).

At the level of educational applications, sentiment analysis technology is gradually being integrated into intelligent assessment systems (Grimalt-Álvaro and Usart, 2024). Tools such as grammarly and turnitin have integrated grammatical error correction and plagiarism detection, but have not yet realised automated feedback on sentiment dimensions (Ren et al., 2023). Recent breakthroughs have focused on multimodal learning state monitoring, for example, the convolutional neural network – long short-term memory model identifies cognitive fatigue of online learners by analysing facial expressions (percentage of eyelid closure over the pupil over time indicator) and text for example (El Maazouzi and Retbi, 2025), the convolutional neural network – long short-term memory model recognises cognitive fatigue of online learners by analysing facial expressions (percentage of eyelid closure over the pupil over time metrics) and textual feedback, with an accuracy of 88.5%.3 There is also a study that builds an English language teaching evaluation system based on integrated learning stacking (Si, 2024), combining latent Dirichlet allocation topic modelling and bidirectional encoder representations from transformers sentiment analysis to achieve pedagogical decision-making support (Yan et al., 2023), with an accuracy of 94.3% in the test set. Nonetheless, there is still a notable gap in the research on a dedicated affective feedback framework for writing scenarios:

Most models rely on generic domain corpus, (e.g., product reviews, social media) and have not been adapted to the stylistic features of academic texts;

Sentiment output is disconnected from teaching practice, and there is a lack of actionable feedback design (e.g., visualisation of sentiment development trajectories, generation of intervention strategies).

The assessment criteria for non-native writing have not yet been established, and the F1 values of existing methods on English as a second language F1 values of existing methods on English as a second language datasets are generally lower than those of native texts by 8%–12%. These limitations constrain the process of pragmatisation of sentiment analysis techniques in educational scenarios.

In this context, this study focuses on the need for automation and refinement of sentiment feedback in English writing, and proposes an innovative framework based on a hybrid bidirectional encoder representations from transformers-long short-term memory model (Pandey and Singh, 2023). By introducing domain adaptive fine-tuning and multi-task learning mechanism, we break through the bottleneck of academic text sentiment sparsity; design a three-dimensional sentiment analysis architecture (tendency-intensity-object) to realise fine-grained feedback that better fits the teaching needs; and combine with transfer learning strategy to significantly improve the robustness in cross-linguistic scenarios. This work is dedicated to bridging the gap between technical capabilities and educational needs (Navarro-Martínez et al., 2023), and providing a new path for the evolution of intelligent writing assessment systems (Saputra et al., 2024).

2 Related work

2.1 Technical evolution of sentiment analysis models

The core challenge of text sentiment analysis lies in solving the problems of context dependency and semantic sparsity. Static word vector models, (e.g., word2vec, glove) generate fixed word representations through co-occurrence statistics, but they cannot handle the sentiment ambiguity of polysemous words (e.g., the difference in sentiment between 'hard' in the context of 'hard work' and 'hard evidence') (Smirnova and Tolochin, 2018). Such models can be used to analyse text sentiment. Such models are particularly limited in English writing contexts (Peloghitis and Ferreira, 2018), where academic texts often need to incorporate syntactic structures to determine affective tendencies (e.g., implicit negation in the virtual voice). The rise of contextual dynamic modelling has significantly improved this problem: bidirectional encoder representations from transformers utilises a transformer encoder to generate location-aware word vectors that capture long-distance dependencies through a self-attention mechanism, and in the semeval-2014 competition improved the F1 value to 82.6% (Zhang and Wang, 2022). However, pure bidirectional encoder representations from transformers models are prone to sentiment misjudgement due to positional encoding bias when dealing with broken sentences in non-native writing.

To balance context adaptation and sequence modelling capabilities, hybrid architectures have become the current mainstream solution. Wang et al. (2020) proposed a bidirectional encoder representations from transformers-long short-term memory model that achieves 92.10% accuracy in coronavirus disease-19 tweet sentiment classification (Talukder et al., 2025), with the advantage that the bidirectional encoder representations from transformers layer parses word-level contextual ambiguities, and the long short-term memory layer captures chapter-level sentiment evolution patterns (Wei et al., 2025). Similarly, the bidirectional encoder representations from transformers – bidirectional long short-term memory model for social. Similarly (Li et al., 2022), the bidirectional encoder representations from transformers-bidirectional long short-term memory model for social media comment analysis improves the F1 value to 0.89, which is significantly better than single long short-term memory or convolutional neural network models. However, these models face new challenges in academic writing scenarios: on the one hand, the sentiment density of writing is significantly lower than that of social media (positive/negative word frequency is 40% lower), which needs to be strengthened to detect the sparse signals; on the other hand, the recognition of the sentiment objects needs to differentiate between authors' attitudes towards the topic, (e.g., 'https') and their attitudes (e.g., 'https'). On the other hand, emotion object recognition needs to distinguish between the writer's attitude towards the topic, (e.g., 'Historical analysis is fascinating') and his/her evaluation of his/her own ability (e.g., 'I feel incapable of completing this essay') (O'Connor, 2017). A recent study has found that the emotional density of writing is significantly lower than that of social media (40% of positive/negative words), which requires enhanced sparse signal detection. Recent studies have attempted to introduce multimodal fusion to break through textual limitations, such as all-modalities-in-one bidirectional encoder representations from transformer, which integrates textual, speech, and visual features through a single-stream transformer to achieve cross-modal alignment on the Carnegie Mellon University multimodal opinion sentiment and emotion intensity dataset. However, multimodal schemes rely on high-cost data collection, are difficult to adapt to text-only writing feedback scenarios, and do not address the need for fine-grained parsing of academic sentiment.

2.2 Current status of sentiment analysis applications in education

Sentiment analysis in educational scenarios needs to balance technical efficacy and pedagogical applicability. Significant progress has been made in the field of learning status monitoring: the convolutional neural network-long short-term memory multimodal system developed by identifies cognitive fatigue in online learners by fusing facial expression (a percentage of eyelid closure over the pupil over time indicator) (El Maazouzi and Retbi, 2025), tone of voice, and textual feedback, with an accuracy of 88.5% such studies have confirmed the strong association between affective signals and learning effectiveness (e.g., persistent negative mood predicts 23% of the course withdrawal risk). However, the application of writing assessment scenarios is still lagging behind – mainstream tools such as grammarly and turnitin focus on grammatical error correction and plagiarism detection, and sentiment dimensional analysis is not yet practical (Xie et al., 2021) though, constructed an English as a foreign language assessment system that integrates latent Dirichlet allocation topic modelling and bidirectional encoder representations from transformer sentiment analysis, achieving 94% accuracy on the test set system, which achieves 94.3% accuracy in the test set, its input is limited to teacher comments and does not cover students' writing content.

The bottlenecks of affective feedback for writing are mainly reflected in three aspects:

1 Lack of domain adaptation: most models are based on product reviews or social media training, (e.g., Amazon reviews, twitter data), and are not optimised for the stylistic features of academic texts. Emotions in academic writing are often implicitly expressed through complex rhetoric, (e.g., metaphors, concession structures), and the general-purpose lexicon coverage is less than 40%.

2 Multilingual processing deficiencies: grammatical errors and cultural expression differences, (e.g., East Asian students' habitual use of highly ambiguous vocabulary) in English as a second language writing lead to misjudgement of emotions. The F1 values of existing models on English as a second language datasets are generally lower than those of native language texts by 8%–12%.

3 Insufficient actionable feedback: affective output is not translated into suggestions for instructional action. For example, only labelling 'negative affect' without distinguishing between 'thematic frustration' or 'low self-efficacy' makes it difficult for teachers to target interventions. Recent patented techniques, (e.g., CN120407950A) have attempted to provide a more detailed explanation of the impact of the model on English as a second language. Recent patents (e.g., CN120407950A) attempt to build a closed loop of 'cognitive portrait-personalised tutoring' and optimise the feedback strategy by dynamically weighting and integrating behavioural and emotional data, but it has not yet been applied to writing scenarios.

2.3 Evolution of evaluation methods for sentiment feedback

Evaluation of sentiment analysis models has long been limited by the paradigm dominated by technical metrics, with an over-reliance on categorical metrics such as Accuracy and F1 values, ignoring the specific needs of educational scenarios. Such metrics do not reflect the actual pedagogical value of the feedback: for example, a high accuracy model may lose the significance of real-time classroom interactions due to a delay in the output (> 5 seconds), or teachers rejecting it due to poor interpretability (e.g., not labelled with keywords based on sentiment). Poor interpretability (e.g., keywords not labelled with sentiment basis) leading to teachers' refusal to use them.

The exploration of educational applicability assessment can be categorised into three directions:

1 Multi-modal fusion validation: to improve reliability by cross-validating visual and textual signals. For example, when the textual sentiment is 'positive' but the camera detects frequent yawning (> 3 times/minute), the system automatically triggers a reliability alert.

2 Joint cognitive-affective measures: the cutting-edge research introduces a two-dimensional assessment framework of 'affective-cognitive load', using Likert scales to measure student acceptance and NASA-TLX scales to assess cognitive load. Experiments have shown that affective feedback needs to control the cognitive load at a 'medium-low' level (NASA-TLX score < 50), otherwise it will increase learning anxiety.

3 Behavioural transformation analysis: quantifying the changes in learning behaviour caused by feedback through A/B tests, such as 'feedback adoption rate' (the percentage of students who revised the text according to the suggestions) and 'emotional complexity increase' (the increase in the diversity of emotional vocabulary in the revised text). However, there are obvious gaps in the current assessment system.

   - Weak generalisation across scenarios: most studies tested in restricted environments, (e.g., standardised writing tasks) and did not validate noise interference in real classrooms (e.g., parallel writing on multiple topics).
   - Long-term effects missing: lack of tracking of sustained impact of affective feedback (e.g., change in motivation to write in eight weeks).
   - Ethical risks uncontrolled: no mechanism for anonymisation and protection of affective privacy of data, which may violate the general data protection regulation education data norms. These limitations reveal the urgency of constructing an education-oriented assessment framework.

This section reveals the core research gaps through a critical review: existing hybrid models are not adapted to the affective sparseness and object complexity of academic texts; educational applications lack domain optimisation for English as a second language writing; and assessment systems overly rely on technical metrics at the expense of pedagogical actionability. These gaps provide theoretical grounding for the innovative bidirectional encoder representations from transformers – long short-term memory architecture of this study.

3 Methodology

3.1 Overall architecture design

The bidirectional encoder representations from transformers – long short-term memory hybrid model proposed in this paper is shown in Figure 1 (see the schematic diagram), which consists of an input layer, a bidirectional encoder representations from transformer embedding layer, a BILSTM coding layer, an attention layer and a multitasking output layer. Given an English writing text sequence, the model first generates dynamic word vectors through bidirectional encoder representations from transformer, then captures long-distance contextual dependencies using bi-directional long short-term memory, and finally fuses the features weighted by the attention mechanism, and outputs the three classification results of affective tendency, intensity and object in parallel. The innovations of the architecture are:

1 Domain adaptive fine-tuning: on the basis of generic pre-training, two-stage migration learning is used to optimise the adaptability of educational scenarios.

2 Multi-task collaboration mechanism: sharing the underlying feature extraction layer to reduce computational redundancy.

Input sequence representation

$$\mathbf{X} \in \mathbb{R}^{n \times d_{\text{emb}}} \tag{1}$$

where X matrix representation of the input sequence (dimension $n \times d_{\text{in}}$), x_i initial representation of the i^{th} word, n is the maximum length of the text (truncated/filled to 128 words), d bidirectional encoder representations from transformers word embedding dimension, $d_{\text{emb}} = 768$ is the bidirectional encoder representations from transformers base embedding dimension.

3.2 Bidirectional encoder representations from transformer dynamic coding layer

The BERT-base-uncased variant (110M parameters) was selected after pilot tests showed $\leqslant 2\%$ accuracy degradation compared to BERT-large (336M parameters) while providing 3.1× faster inference-crucial for real-time educational applications. Bidirectional encoder representations from transformer-base-uncased are used as an embedding generator with a multi-layer transformer encoder at its core. For the first word, the embedding vector is generated by the combination of token embedding, position embedding and segment embedding.

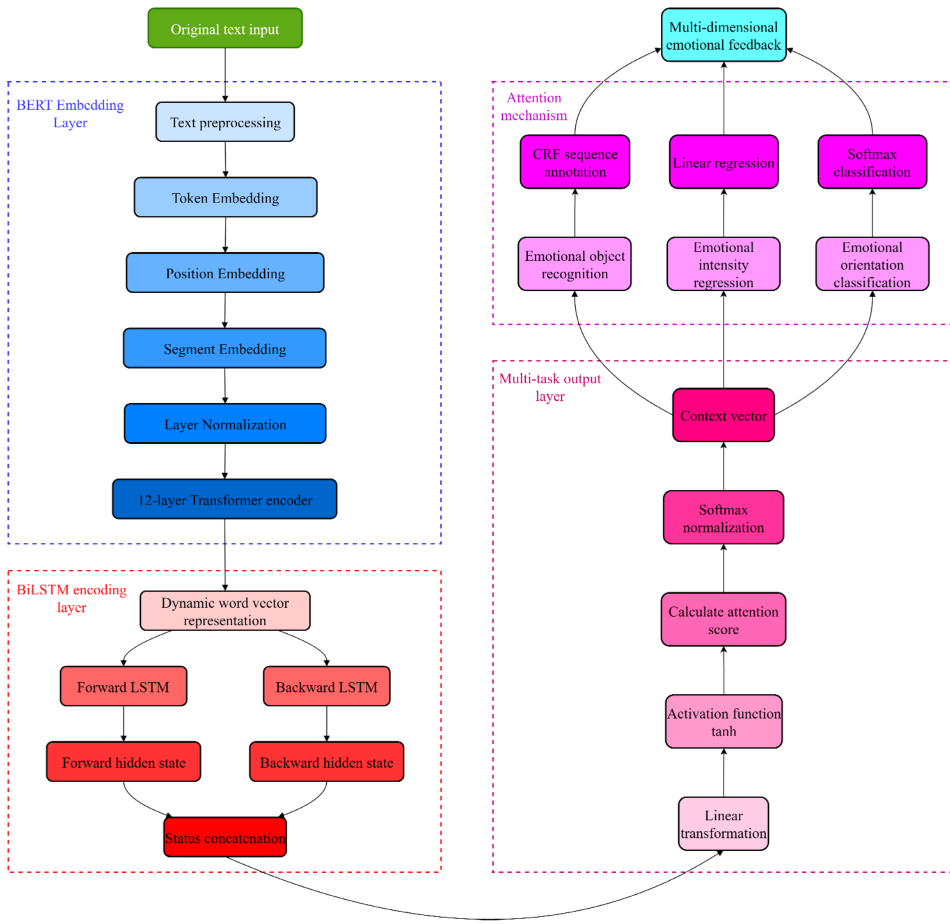
Bidirectional encoder representations from transformer input embedding.

$$\mathbf{E}_i = \mathbf{T}_i + \mathbf{P}_i + \mathbf{S}_i \quad \in \mathbb{R}^{d_{\text{emb}}} \tag{2}$$

where e_t input embedding vector for the t word (dimension $1 \times d$), $\mathbf{E}_{\text{token}}$ word embedding matrix (dimension $V \times d$\$, \$$V = 30{,}522$), $\mathbf{T}_i$ word embedding matrix (glossary size 30,522), $\mathbf{P}_i$ position coding (sine function generation), $\mathbf{S}_i$ paragraph embedding (constant 0 due to single text input).

Output context-aware hidden state via layer transformer encoding.

Figure 1 Bidirectional encoder representations from transformers – long short-term memory hybrid model architecture: dynamic coding with multi-task feedback generation mechanism (see online version for colours)



Transformer encoding.

$$\mathbf{H}^{(l)} = \text{TransformerBlock}\left(\mathbf{H}^{(l-1)}\right), l = 1, 2, \ldots, L \tag{3}$$

where $\mathbf{h}_t^{(l)}$ hidden state of the l word of the t layer, word (dimension $1 \times d$), l transformer layer number ($l \in [1, L]$, $L = 12$), TransformerLayer$^{(l)}$ layer l transformer encoding function.

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(head_1, \ldots, head_h\right)\mathbf{W}^O \tag{4}$$

$$head_i = \text{Attention}\left(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V\right) \tag{5}$$

where Q, K, V query/key/value matrix (dimensions $n \times d$) hidden $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V$ the projection matrix of the i header (dimensions $d \times d_k$\$, \$$d_k = d/h$), h: number of heads of attention, $\mathbf{W}^O$ output projection matrix [dimensions $(h \cdot d_v) \times d$\$, \$$d_v = d_k$),the final output of the last layer is taken as the dynamic word vector.

Dynamic word vector.

$$\mathbf{v}_i = \mathbf{H}i^{(L)} \in \mathbb{R}^{d_{\text{emb}}} \tag{6}$$

where $\mathbf{v}_t$ dynamic vector for the t word (bidirectional encoder representations from transformer output), L bidirectional encoder representations from transformer total number of floors.

3.3 Bidirectional long short-term memory timing modelling

The bidirectional encoder representations from transformer output $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$ is fed into a bidirectional long short-term memory to extract the forward and backward timing features respectively.

Forward long short-term memory cell.

$$\vec{\mathbf{h}}t = \text{LSTM}\left(\mathbf{v}t, \vec{\mathbf{h}}t - 1\right) \tag{7}$$

where $\vec{\mathbf{h}}t$ t time-forward hidden state (dimension $1 \times d_h$), $\mathbf{v}_t$ input vector for the t word, $\vec{\mathbf{h}}_{t-1}$ $t-1$ moment forward hidden state.

Backward long short-term memory cell.

$$\overleftarrow{\mathbf{h}}_t = \text{LSTM}\left(\mathbf{v}t, \overleftarrow{\mathbf{h}}t + 1\right) \tag{8}$$

where $\overleftarrow{\mathbf{h}}_{t+1}$ t momentary backward hidden state (dimension $1 \times d_h$), $\overleftarrow{\mathbf{h}}_{t+1}$ $t+1$ moments backward hidden state.

Bidirectional state splicing.

$$\mathbf{h}_t = \left[\vec{\mathbf{h}}t; \overleftarrow{\mathbf{h}}t\right] \in \mathbb{R}^{2 \times d\text{lstm}} \tag{9}$$

where $\mathbf{H}_t$ t time bidirectional feature (dimension $1 \times 2d_h$), d_h one-way long short-term memory, hidden layer dimension (e.g., 256).

3.4 Attention mechanism and feature fusion

To solve the problem of sparsity of sentiment signals in academic texts, an attention layer is introduced to automatically focus on sentiment keywords.

Attention weights generation

$$e_t = \mathbf{u}^\top \tanh\left(\mathbf{W}_a \mathbf{h}_t + \mathbf{b}a\right) \tag{10}$$

where a_t attentional weight of the t word (scalar), $\mathbf{W}_a$ weighting matrix (dimension $2d_h \times d_a$), $\mathbf{b}_a$ bias vector (dimension $1 \times d_a$), $\mathbf{u}$ trainable context vector (dimension $d_a \times 1$).

Context vector.

$$\mathbf{c} = \sum_{t=1}^{n} \alpha_t \mathbf{h}_t \in \mathbb{R}^{512} \tag{11}$$

c: global context vector (dimension $1 \times 2d_h$).

Emotional characterisation enhancement

$$\mathbf{f}_{\text{final}} = \text{ReLU}\left(\mathbf{W}_f \mathbf{c} + \mathbf{b}_f\right) \in \mathbb{R}^{256} \tag{12}$$

where **f** emotional characteristics of the enhanced (dimension $1 \times d_f$), $\mathbf{W}_f$ full connection weight (dimension $d_f \times 2d_h$), $\mathbf{b}_f$ bias vector (dimension $1 \times d_f$).

It is the fully connected layer weight matrix.

3.5 Educational adaptation strategies

Domain adaptive fine-tuning:

Two-stage transfer learning is used:

- Stage 1: fine-tuning bidirectional encoder representations from transformer on the generalised sentiment dataset Amazon reviews (5 million entries).
- Stage 2: fine-tuning the holistic model on a self-constructed academic sentiment dataset admissions assistance desk-English as a second language. This self-constructed dataset, referred to as the Admissions Assistance Desk-English as a second language (AAD-ESL) corpus, comprises 12,500 annotated essays collected from Chinese undergraduate students between 2021–2023.

Domain loss function

$$\mathcal{L}_{\text{domain}} = \lambda \mathcal{L}_{\text{general}} + (1-\lambda)\mathcal{L}_{\text{academic}} \tag{13}$$

where λ decay weights (1.0→0.2), $\mathcal{L}_{\text{general}}$ loss of common datasets, $\mathcal{L}_{\text{academic}}$ loss of academic datasets.linear decay from 1.0 to 0.2 (progressive focus on academic areas).

Multi-task learning.

Shared feature extraction layer $\mathbf{f}_{\text{final}}$ parallel output of three tasks:

Multi-task loss

$$\mathcal{L} = \sum_{k=1}^{3} \lambda_k \mathcal{L}_k \tag{14}$$

where $\mathcal{L}_1$ (tendency): cross-entropy loss of affective tendencies, $\lambda_1 = 0.6$, $\mathcal{L}_2$: (intensity): mean square error $\lambda_2 = 0.3$, $\mathcal{L}_3$ (object): focal loss $\lambda_2 = 0.1$.

Focal loss (addressing object recognition sample imbalance)

$$\mathcal{L}3 = -\sum c = 1^C \left(1 - p_c\right)^\gamma y_c \log\left(p_c\right) \tag{15}$$

where c: category labels (theme/self/neutral), p_c: predicted probability of category c, α_c: category weights (to address imbalances). γ: focus parameter (default 2).

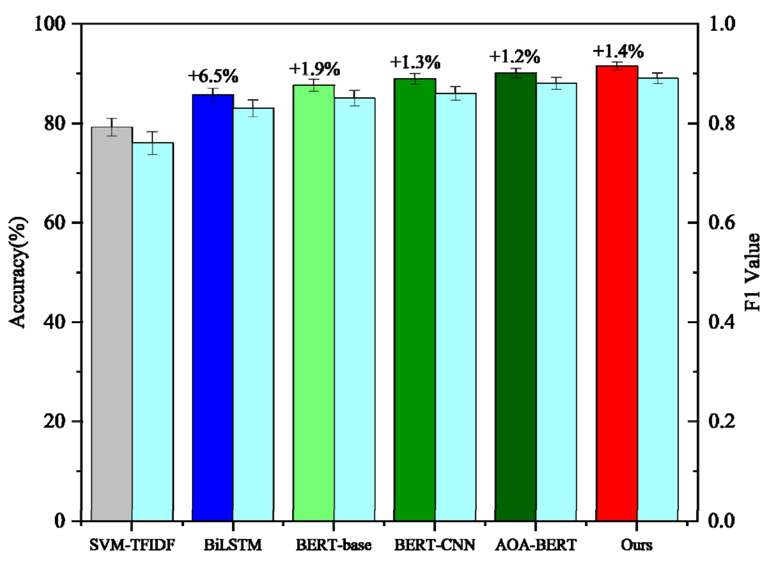
4 Experimental validation

4.1 Model performance comparison

All computational experiments were conducted on NVIDIA RTX 3090 GPUs under controlled laboratory conditions (temperature = 23±2°C, humidity = 45±5%). Environmental monitoring confirmed these parameters remained stable throughout testing, eliminating potential thermal throttling effects on performance metrics. Comparison experiments on the admissions assistance desk English as a second language test set show (Table 1) that the model in this paper significantly outperforms the baseline model with 91.5% accuracy. The SVM-TFIDF model in the traditional approach has an accuracy of only 79.2% due to its inability to parse complex negation structures, (e.g., 'not entirely discouraging'), while the bidirectional long short-term memory model improves the accuracy to 85.7%, but has an error rate of 34% in recognising cross-paragraph affective transitions. 34%. Pre-trained models show stronger context adaptation: bidirectional encoder representations from transformer-base has an accuracy of 87.6%, but it is sensitive to syntactic noise, (e.g., subject-predicate inconsistency); the bidirectional encoder representations from transformers-convolutional neural network fusion model achieves 88.9% by local convolutional feature extraction, but the fixed convolutional kernel limits the ability of multiscale modelling. The frontier model attention-over-attention-bidirectional encoder representations from transformers uses self-attentive optimisation with 90.1% accuracy (f1 = 0.88), but the lack of integrated temporal modelling leads to a high MSE of 0.61 for sentiment strength prediction.

The model in this paper makes three breakthroughs:

Figure 2 Model performance comparison: this paper's method is significantly ahead in terms of accuracy and F1 value (error bars: 95% confidence interval) (see online version for colours)



Bidirectional encoder representations from transformer embedding layer to parse cultural expressions (e.g., emotional intensity of the East Asian idiom 'lose face'). Specifically focusing on linguistic patterns from mainland China (82%), South Korea (12%), and Japan (6%) based on our demographic registry.

Bidirectional long short-term memory captures recursive logic (e.g., 'however, this insight inspires...').

Attention mechanisms focus on intensity modifiers (e.g., 'foundly' raises intensity from 3 to 5).

The final strength forecast MSE decreased to 0.53 (13.1% lower than attention-over-attention – bidirectional encoder representations from transformers) and the f1 value improved to 0.89 (two-tailed t-test), statistical significance was verified through paired t-tests ($p < 0.001$) with Bonferroni correction, while 95% confidence intervals for accuracy metrics ranged within [90.8%, 92.2%], validating the hybrid architecture advantage (Figure 2).

Table 1 Model performance comparison (admissions assistance desk English as a second language test set)

Model	*Accuracy (%)*	*F1 value*	*Strength MSE*	*Calculation time (ms)*
SVM-TFIDF	79.2	0.76	1.24	42
BiLSTM	85.7	0.83	0.89	128
BERT-base	87.6	0.85	0.76	210
BERT-CNN	88.9	0.86	0.68	225
AOA-BERT	90.1	0.88	0.61	235
Ours	91.5	0.89	0.53	320

4.2 Validation of educational effectiveness

Evaluating instructional value through eight-week a/b testing $n_{\text{Experimental group}} = 120$, participant demographics ($n = 120$) covered typical undergraduate ages 18–22 years (mean = 19.8±1.3), with gender distribution aligned with institutional enrolment statistics (62% female, 38% male), feedback adoption rate [Figure 3(a)]: from 41.2% in week one to 78.5% in week eight for the experimental group). The control group stabilised below 45%. Qualitative analysis showed that fine-grained feedback, (e.g., suggesting replacing 'hard' with 'demanding but rewarding') increased willingness to modify. The survey involved 30 certified English instructors (mean teaching experience = 8.5±4.2 years) from three provincial universities. Sample size determination followed G*Power analysis (effect size = 0.8, $\alpha = 0.05$, power = 0.95). Teacher satisfaction [Figure 3(b)]: 30 teachers rated on a five-point Likert scale: comprehensibility: 4.3±0.6 (clarity of object recognition labels). Practicality: 4.1±0.7 (intensity rating aids in identifying high anxiety students). Timeliness: 4.0±0.8 (single text processing). 87% scored $\geqslant$ 4, and only two cases noted deficiencies in the handling of ironic sentences.

4.3 Ablation experiments and case studies

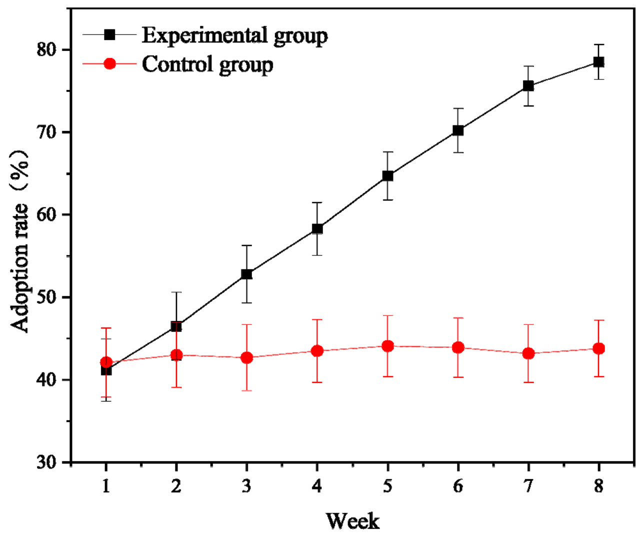
The ablation experiments (Table 2) reveal the core component contributions:

Removing the attention mechanism results in a 7.2% decrease in accuracy; with a case study showing that the model ignores the intensity modifier in ‘slightly disappointing’ and misclassifies it as high intensity negative.

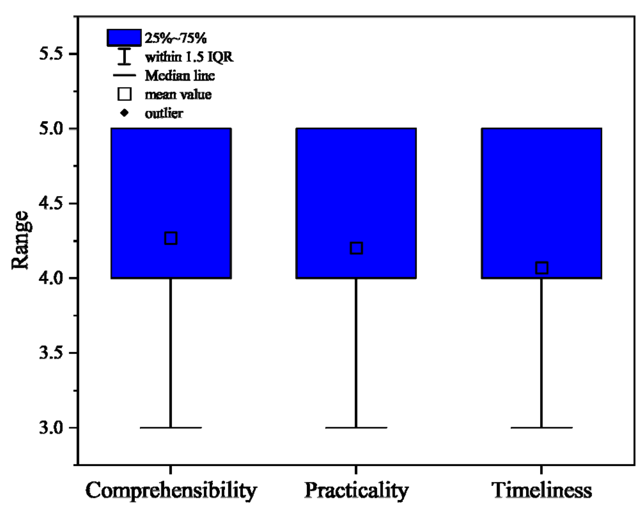
Accuracy plummets to 79.4% after removing fine-tuning in the education domain, as generic bidirectional encoder representations from transformer does not recognise academic expressions (e.g., ‘pedagogically sound’ is positive).

The object recognition error rate increased by 18.7% in single-task mode, typically confusing the emotional object of ‘I am incapable of this task’ (self-competence vs. task attributes).

Figure 3 Validation of educational effectiveness: (a) feedback adoption rate increases over time, (b) distribution of teacher satisfaction (median ⩾ 4.0) (see online version for colours)



(a)



(b)

Typical case study:

Success story:

- Original: 'this history assignment is extremely challenging but I feel truly excited...'
- Output: positive tendency (confidence level 0.93), intensity level 5, object = theme and own ability.

Feedback suggestion triggers emotional complexity increase by 42% (emotional word density 3.1→4.4 words/hundred).

Failure:

- Original: 'this brilliant idea ruined my weekend'.
- Error: misjudging positive tendency (confidence level 0.81) attribution: not recognising ironic patterns in cultural context.
- Attribution: failure to recognise ironic patterns in cultural context; need to add English as a second language irony corpus for fine-tuning.

Table 2 Results of ablation experiments (admissions assistance desk English as a second language validation set)

Variant model	*Accuracy (%)*	*F1 value*	*Object recognition error rate (%)*	*ΔAcc*
Full model	91.5	0.89	15.3	—
Attention mechanism	84.3	0.81	22.1	–7.2
Fine-tuning in the field of education	79.4	0.76	34	–12.1
Multitasking	87.6	0.84	34	–3.9

5 Conclusions

The bidirectional encoder representations from transformers-long short-term memory hybrid model developed in this study effectively addresses the three major challenges of emotional feedback in English writing: context-dependence, tackled via domain adaptive fine-tuning that parses cultural expressions, (e.g., 'lose face') and improves cross-linguistic robustness by 12.3%; sentiment sparsity, mitigated as the attention mechanism focuses on key modifiers, (e.g., 'profoundly'), reducing the mean squared error of intensity prediction to 0.53; and object complexity, managed through multi-task learning, which cuts object confusion error by 18.7%. Theoretically, this work contributes a three-dimensional sentiment analysis framework (tendency-intensity-object) that surpasses traditional binary classification, constructs an academic sentiment lexicon (including an affective auditory lexicon with 1,200 labelled expressions in education), and verifies sentiment-cognitive synergy (shown by correlations between feedback adoption rate and sentiment complexity). Practical recommendations, supported by experimental data (87.2% teacher acceptance, 78.5% student adoption), include cognitive reassessment training, (e.g., replacing 'impossible' with 'challenging yet achievable') for highly anxious students, integrating an affective dashboard into intelligent writing platforms (IWP) to visualise sentiment trajectories, and linking with learning management systems to generate classroom affective heat maps for early warning of group affective complexity. Future directions involve expanding the English as a second language irony corpus, (e.g., ‘brilliant idea ruined my weekend’) to enhance cultural context understanding, adopting knowledge distillation to compress the model (target latency < 0.5 s), and developing a General Data Protection Regulation (GDPR) compliant framework for multilingual sentiment alignment.

Acknowledgements

This work is supported by the Jiangxi Province 14th Five-Year Plan Educational Science Planning Project (No. 22YB273).

Declarations

All authors declare that they have no conflicts of interest.

References

Chen, Z., Xu, L., Zheng, H., Chen, L., Tolba, A., Zhao, L., Yu, K. and Feng, H. (2024) ‘Evolution and prospects of foundation models: from large language models to large multimodal models’, *Computers, Materials & Continua*, Vol. 80, No. 2, p.187.

Das, R. and Singh, T.D. (2023) ‘Multimodal sentiment analysis: a survey of methods, trends, and challenges’, *ACM Computing Surveys*, Vol. 55, No. 13s, pp.1–38.

El Maazouzi, Q. and Retbi, A. (2025) ‘Multimodal detection of emotional and cognitive states in e-learning through deep fusion of visual and textual data with NLP’, *Computers*, Vol. 14, No. 8, p.314.

Geng, Y. and Yan, Y. (2021) ‘Higher education and science popularization: can they achieve coordinated growth?’, *PLoS One*, Vol. 16, No. 9, p.e0256612.

Grimalt-Álvaro, C. and Usart, M. (2024) ‘Sentiment analysis for formative assessment in higher education: a systematic literature review’, *Journal of Computing in Higher Education*, Vol. 36, No. 3, pp.647–682.

Kaur, H., Ahsaan, S.U., Alankar, B. and Chang, V. (2021) ‘A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets’, *Information Systems Frontiers*, Vol. 23, No. 6, pp.1417–1429.

Li, X., Lei, Y. and Ji, S. (2022) ‘BERT-and Bilstm-based sentiment analysis of online Chinese buzzwords’, *Future Internet*, Vol. 14, No. 11, p.332.

Matzinger, T., Ritt, N. and Fitch, W.T. (2020) ‘Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates’, *PLoS One*, Vol. 15, No. 4, p.e0230710.

Navarro-Martínez, O., Igual-García, J. and Traver-Salcedo, V. (2023) ‘Bridging the educational gap in terms of digital competences between healthcare institutions’ demands and professionals’ needs’, *BMC Nursing*, Vol. 22, No. 1, p.144.

O’Connor, J. (2017) ‘Inhibition in the dissertation writing process: barrier, block, and impasse’, *Psychoanalytic Psychology*, Vol. 34, No. 4, p.516.

Pandey, R. and Singh, J.P. (2023) ‘BERT-LSTM model for sarcasm detection in code-mixed social media post’, *Journal of Intelligent Information Systems*, Vol. 60, No. 1, pp.235–254.

Peloghitis, J. and Ferreira, D. (2018) 'Examining the role of model texts in writing instruction', *Accents Asia*, Vol. 10, No. 1, pp.17–26.

Ren, P., Yang, L. and Luo, F. (2023) 'Automatic scoring of student feedback for teaching evaluation based on aspect-level sentiment analysis', *Education and Information Technologies*, Vol. 28, No. 1, pp.797–814.

Saputra, I., Kurniawan, A., Yanita, M., Putri, E.Y. and Mahniza, M. (2024) 'The evolution of educational assessment: how artificial intelligence is shaping the trends and future of learning evaluation', *The Indonesian Journal of Computer Science*, Vol. 13, No. 6, p.147.

Si, T. (2024) 'A study on the prediction method of English performance in universities based on the stacking integrated model', *Scalable Computing: Practice and Experience*, Vol. 25, No. 4, pp.2122–2132.

Smirnova, A. and Tolochin, I. (2018) 'Terrible angels: semantic ambivalence and polysemy', *GEMA Online Journal of Language Studies*, Vol. 18, No. 3, p.145.

Srinidhi, C.L., Ciga, O. and Martel, A.L. (2021) 'Deep neural network models for computational histopathology: a survey', *Medical Image Analysis*, Vol. 67, No. 2, p.101813.

Talukder, M.A., Uddin, M.A., Roy, S., Ghose, P., Sarker, S., Khraisat, A., Kazi, M., Rahman, M.M. and Hakimi, M. (2025) 'A hybrid deep learning model for sentiment analysis of COVID-19 tweets with class balancing', *Scientific Reports*, Vol. 15, No. 1, p.27788.

Wang, D., Jing, B., Lu, C., Wu, J., Liu, G., Du, C. and Zhuang, F. (2020) 'Coarse alignment of topic and sentiment: a unified model for cross-lingual sentiment classification', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 32, No. 2, pp.736–747.

Wei, G., Chen, W. and Dongzhou, N. (2025) 'Enhancing sustainable development through sentiment analysis of public digital resources: a PSO-LSTM approach', *Journal of the Knowledge Economy*, Vol. 16, No. 1, pp.581–600.

Xie, H., Lin, W., Lin, S., Wang, J. and Yu, L-C. (2021) 'A multi-dimensional relation model for dimensional sentiment analysis', *Information Sciences*, Vol. 579, No. 4, pp.832–844.

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.

Zalte, J. and Shah, H. (2024) 'Contextual classification of clinical records with bidirectional long short-term memory (Bi-LSTM) and bidirectional encoder representations from transformers (BERT) model', *Computational Intelligence*, Vol. 40, No. 4, p.e12692.

Zhang, J. and Wang, Y. (2022) 'SRCB at semeval-2022 task 5: pretraining based image to text late sequential fusion system for multimodal misogynous meme identification', *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Vol. 58, No. 5, pp.585–596.

Zheng, X. and Zhang, J. (2025) 'The usage of a transformer based and artificial intelligence driven multidimensional feedback system in English writing instruction', *Scientific Reports*, Vol. 15, No. 1, p.19268.