electronics

MDPI

Article

Short-Text Sentiment Classification Model Based on BERT and Dual-Stream Transformer Gated Attention Mechanism

Song Yang [1,2,*], Jiayao Xing [1], Zhaoxia Liu [1,2] and Yunhao Sun [1]

[1] School of Software, Dalian University of Foreign Languages, Dalian 116044, China; xingjiayao@dlufl.edu.cn (J.X.); liuzhaoxia@dlufl.edu.cn (Z.L.); sunyunhao@dlufl.edu.cn (Y.S.)
[2] Research Center for Networks Space Multi-Languages Big Data Intelligence Analysis, Dalian 116044, China
* Correspondence: ocean_epoch@163.com

Abstract

With the rapid development of social media, short-text data have become increasingly important in fields such as public opinion monitoring, user feedback analysis, and intelligent recommendation systems. However, existing short-text sentiment analysis models often suffer from limited cross-domain adaptability and poor generalization performance. To address these challenges, this study proposes a novel short-text sentiment classification model based on the Bidirectional Encoder Representations from Transformers (BERTs) and a dual-stream Transformer gated attention mechanism. This model first employs Bidirectional Encoder Representations from Transformers (BERTs) and the Chinese Robustly Optimized BERT Pretraining Approach (Chinese-RoBERTa) to achieve data augmentation and multilevel semantic mining, thereby expanding the training corpus and enhancing minority class coverage. Second, a dual-stream Transformer gated attention mechanism was developed to dynamically adjust feature fusion weights, enhancing adaptability to heterogeneous texts. Finally, the model integrates a Bidirectional Gated Recurrent Unit (BiGRU) with Multi-Head Self-Attention (MHSA) to strengthen sequence information modeling and global context capture, enabling the precise identification of key sentiment dependencies. The model's superior performance in handling data imbalance and complex textual sentiment logic scenarios is demonstrated by the experimental results, achieving significant improvements in accuracy and F1 score. The F1 score reached 92.4%, representing an average increase of 8.7% over the baseline models. This provides an effective solution for enhancing the performance and expanding the application scenarios of short-text sentiment analysis models.

Keywords: short text sentiment classification; BERT; dual-stream; gated attention mechanism; data enhancement; feature fusion; data imbalance

check for updates

Academic Editors: Jie Zou, Wanyu Chen and Shimin Cai

Received: 12 August 2025
Revised: 25 September 2025
Accepted: 27 September 2025
Published: 30 September 2025

Citation: Yang, S.; Xing, J.; Liu, Z.; Sun, Y. Short-Text Sentiment Classification Model Based on BERT and Dual-Stream Transformer Gated Attention Mechanism. *Electronics* **2025**, *14*, 3904. https://doi.org/10.3390/electronics14193904

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Information dissemination has become increasingly diverse in today's digital era, with short text becoming a predominant medium for daily communication and information exchange. From concise social media posts and product reviews on e-commerce platforms to rapid interactions in online forums, short texts now permeate virtually every dimension of modern life. Despite their brevity, such texts frequently convey subtle and complex sentimental cues, functioning as micro-level signals that collectively form a vast and intricate network of sentimental expression. However, this sentimental information is frequently buried in massive volumes of data, making effective capture and interpretation

challenging [1]. Furthermore, the inherent constraints of short texts—such as limited contextual information and ambiguous semantics—impede a comprehensive understanding of human sentimental communication patterns and hinder the development and practical implementation of related technologies. Research on short-text sentiment classification is of profound theoretical and practical significance.

Sentiment classification, as a key subfield of natural language processing (NLP), aims to automatically identify the sentimental tendencies embedded in text, determining whether a given text conveys positive, negative, or neutral sentiments. This technology has demonstrated substantial application value in areas such as public opinion monitoring, market research, personalized recommendation systems, customer feedback analysis, intelligent customer service, and social network analytics. It enables enterprises to gain a more accurate understanding of user needs and supports government agencies in monitoring public sentiments for informed decision-making. Moreover, sentiment analysis exhibits promising potential in emerging fields such as online education, healthcare, and financial market prediction.

To address these challenges, a short-text sentiment classification model based on BERT and dual-stream transformer gated attention mechanism is proposed. Masked language modeling (MLM) generation techniques are employed by incorporating BERT and Chinese-RoBERTa models to synergize data augmentation with feature extraction. This approach enriches the diversity of the training corpus, extracts multidimensional semantic information, and implicitly expands minority class samples, thereby mitigating bias induced by class imbalance. Subsequently, the dual-stream Transformer gated attention mechanism enables adaptive feature fusion adjustment. The semantic expression capabilities of the model are enhanced by dynamically allocating character-level and word-level feature weights based on text characteristics, and the recognition performance for minority samples is improved. Finally, integrating BiGRU with Multi-Head Self-Attention (MHSA) enhances sequential information processing, improving the model's ability to capture long-range dependencies. This enables precise sentiment analysis of short texts and achieves balanced recognition of majority and minority samples.

2. Related Work

Sentiment analysis has undergone significant methodological evolution as a pivotal research direction in the field of natural language processing (NLP), reflecting the development of artificial intelligence (AI). Early studies predominantly relied on lexicon-based approaches to determine sentimental polarity using domain-specific sentiment dictionaries. For instance, Jiang et al. [2] innovatively integrated an enhanced TF-IDF algorithm with BERT-based semantic representation. Using this method, a domain-specific sentiment lexicon was constructed, which achieved F1 score of 78.02% and 88.35% for sentiment analysis in the automotive and mobile phone domains, respectively. Building on this work, Wang et al. [3] further advanced the approach by integrating sentiment lexicons as prior knowledge into a SABLSTM model. Their experiments, conducted on a large-scale hotel review dataset (Tan Songbo's corpus), achieved an F1 score of 92.18%, demonstrating the effective fusion of symbolic prior knowledge with neural network architectures. Nevertheless, lexicon-based methods have inherent limitations, particularly due to the lag in updating static dictionaries and insufficient sensitivity to contextual nuances. Li et al. [4] reported that their sentiment lexicon for online bullet-screen comments achieved only 63% coverage for emerging terms and exhibited a 28% polarity misclassification rate in financial text analysis. These challenges have driven researchers toward more dynamically adaptive machine learning paradigms.

The emergence of supervised learning marked a shift toward feature engineering-driven sentiment classification. Researchers sought to enhance model performance by designing handcrafted text features and jointly optimizing classifiers. Tarımer et al. [5] demonstrated through a comparative study that support vector machines (SVM) weighted with TF-IDF achieved an accuracy of 78.5% in e-commerce review classification. However, their performance degraded by 12% when handling long text, highlighting the inherent deficiencies in semantic coherence processing. Dervenis et al. [6] proposed a random forest (RF) model enhanced with integrated n-gram features, achieving a macro F1 score of 81.3% on a student review dataset. Despite these improvements, feature engineering accounted for 62% of the total development cycle, highlighting bottlenecks in manual feature construction efficiency. Moreover, as feature dimensionality exceeds 5000, training times increase exponentially, and cross-domain generalization performance declines sharply. For instance, when a medical domain model was applied to financial text, classification accuracy dropped by 35%. These constraints catalyzed the adoption of deep learning techniques centered on automatic feature extraction.

Deep learning has opened a new chapter in sentiment analysis by surpassing traditional approaches' performance limitations through its hierarchical feature abstraction capabilities. Long et al. [7] designed a hybrid model integrating Self-Attention with TextCNN-BiLSTM, which leverages the synergy between global semantic perception and local feature extraction. Evaluated on the hotel review corpus curated and annotated by Professor Tan Songbo, the model achieved an F1 score of 77.75%. Huang et al. [8] proposed a dual-channel BiLSTM-CNN architecture tailored to the fragmented semantic nature of Weibo short texts. By incorporating both temporal modeling and spatial feature extraction, the model achieved an F1 score of 79.38% on the hotel review dataset curated by Tan Songbo. In the Chinese context, Sun et al. [9] proposed the HDCL model, which employs dynamic contextual representation to effectively alleviate lexical ambiguity. The incorporation of Transformer architecture represents a substantial technological breakthrough. The model was evaluated on three publicly available Chinese sentiment classification datasets: OCEMOTION, online_shopping_10_cats, and DMSC, achieving F1 scores of 0.582, 0.953, and 0.814, respectively. Zhang et al. [10] developed the TE-GRU model, which integrates transformers with GRU networks, achieving an F1 score of 89.3% for long text analysis, an improvement of 8.7% over traditional recurrent networks. Similarly, Hu et al. [11] developed a BERT-BGRU-Att model, which attained an F1 score of 92% on the publicly available online review dataset curated by Professor Tan Songbo. The 120 million parameter model's significant size resulted in an energy consumption of 320 kWh during training, and Tembhurne's multimodal model [12] encountered a 480 ms inference delay, highlighting the practical challenges of deploying large models. These findings have motivated a shift in research toward balancing accuracy and computational efficiency. A Weibo sentiment analysis model created by Li et al. [13] integrates BERT for semantic representation, sentiment lexicons to enhance feature extraction, BiLSTM for contextual modeling, and an attention mechanism for feature weighting, resulting in substantial improvements via collaborative processing stages.

In vertical domains, the evolution of sentiment analysis techniques has closely intertwined with application requirements. Multi-source data fusion and adversarial learning are emphasized for robust financial sector predictions. Sun et al. [14] created a model that combines LightGBM and LSTM by incorporating macroeconomic indicators, news sentiment indices, and search behaviour data, thus lowering the gold futures price prediction error to 2.14%, representing an 11.2% improvement over single sequence models. Xu et al. [15] presented the SA-BERT-LSTM framework, which achieved an accuracy of 87.6% in forecasting the CSI 300 index and maintained stock-specific prediction discrep-

ancies within a ±1.5% margin by incorporating sentiment features, thereby confirming the value of unstructured sentiment information. Duan et al. [16] proposed the FinBERT-RCNN-ATTACK model to counter adversarial examples specific to financial texts. By incorporating adversarial perturbations at the embedding layer, the model achieved an F1 score of 94.56% on the dataset from the "CCF Negative Financial Information and Entity Identification Competition".

In the education domain, deep learning models have outperformed traditional techniques. Huang et al. [17] introduced a BERT-BiLSTM-Attention model that visualizes attention weights to identify critical negative sentiment terms, such as "insufficient course interaction," improving the F1 score to 90.3%. However, when processing dialectal expressions, the accuracy of the model dropped to 61%, suggesting the need for fine-grained semantic understanding. Liang et al. [18] developed a BERT-BiGRU-LDA hybrid model that dynamically fuses features, achieving 92.3% accuracy in the educational review analysis, a 7.8% improvement over single models. Liyih et al. [19] combined a CNN-BiLSTM architecture with Word2Vec and SMOTE techniques to attain a classification accuracy of 95.73% in cross-lingual and cross-domain tasks, particularly for YouTube comments on contentious subjects, thus demonstrating the utility of transfer learning methods.

Sentiment analysis in the healthcare domain faces unique challenges due to dense technical terminology and privacy sensitivities, prompting tailored solutions. Guo et al. [20] constructed a domain-specific medical sentiment lexicon by combining an LDA topic model with the SO-PMI algorithm. Their study identified four prominent themes from 32,000 patient reviews. By evaluating the average performance across four classification models and three metrics, the Support Vector Machine (SVM) achieved the highest mean scores on all three evaluation criteria, thereby establishing a novel paradigm for the quantitative assessment of medical services. Wen et al. [21] proposed a CNN-BiLSTM-LDA fusion model that captures local semantic features and contextual dependencies, resulting in an F1 score of 94.43%—a 10.2% improvement over single models—and providing insight into the sentiment distribution of five key medical topics. Zhang et al. [22] designed a BERT-BiLSTM-CNN dual-channel architecture that achieved a macro F1 score of 91.2% on medical text binary classification tasks. An innovative fusion of features led to a 230 texts per second increase in GPU inference speed, paving the way for real-time medical sentiment monitoring systems. These developments collectively signify a move away from coarse-grained approaches and toward fine-grained, domain-adaptive techniques.

Systematic comparisons of technical paradigms have illuminated their respective boundaries and development potential. Mishev et al. [23] reported that traditional bag-of-words models achieved a macro F1 score of only 53.7% on 10,000 financial texts in low-resource scenarios, whereas a lightweight transformer model improved the accuracy to 85.7%. However, this came at the cost of a 15-fold increase in training energy consumption, underscoring the importance of green computing. On medium-scale datasets (50,000–100,000 samples), Yan et al. [24] proposed the DAPDM dual-channel model, which achieved 89.7% accuracy, outperforming SVM by 21.4%, and validating the superior feature abstraction capabilities of deep learning. For fine-grained sentiment analysis, Wang et al. [25] developed the MacBERT interactive attention mechanism, which dynamically associates aspect terms with contextual semantics, improving the F1score to 92.3%. Building on this, Gao et al. [26] combined a multiscale fusion network with soft attention, surpassing the 95% F1 score threshold and setting a new benchmark for complex semantic parsing.

Sentiment analysis has seen the emergence of three major trends. Firstly, data types have shifted from individual text analysis to the inclusion of various, multiple data sources, as illustrated by Sun et al.'s [14] integration of macroeconomic indicators and sentiment indices for cross-modal learning purposes. Second, model architectures have evolved

from generic frameworks to domain-specific designs, with FinBERT achieving an accuracy gain of 28% in financial sentiment tasks. Third, research priorities have transitioned from merely pursuing prediction accuracy to balancing interpretability and practical utility. Notably, Rizinski et al. [27] developed the Transformer-SHAP framework, which enhanced financial lexicon coverage by 38% and improved decision transparency through feature attribution visualization. Advances in sentiment analysis indicate a trend toward more sophisticated and practical applications. Overcoming key challenges, such as dynamic lexicon construction, low-resource transfer learning, and energy efficiency optimization, will be critical in propelling the field.

3. Methodology

The proposed model comprises five core components: a data augmentation module, a feature extraction and fusion module, a sequential information processing module, an aggregation and classification module, and an optimization and training module. Figure 1 illustrates the overall model architecture.

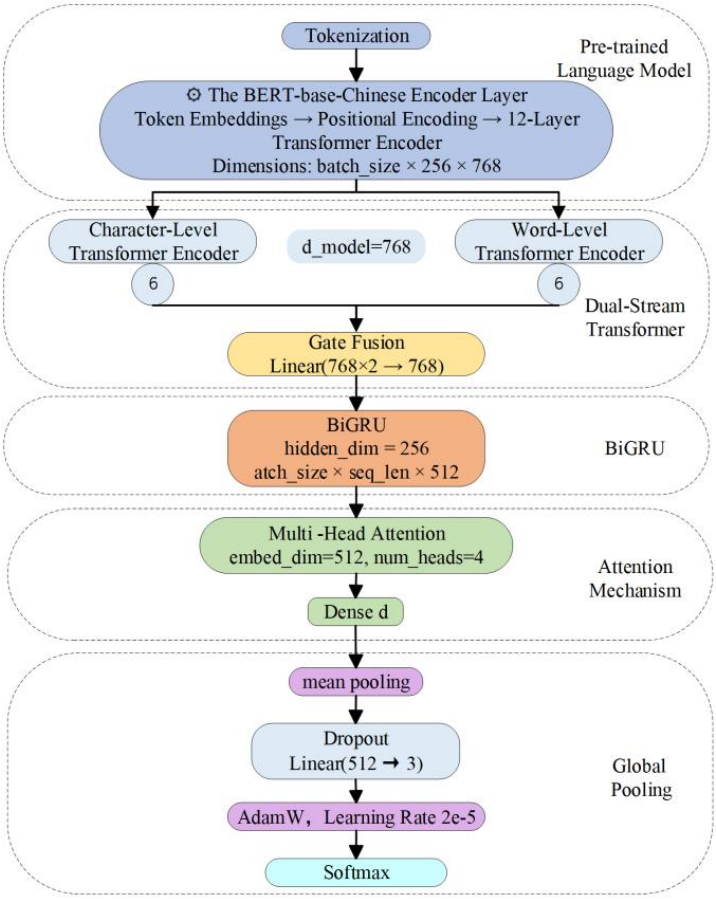


Figure 1. Model Structure Diagram.

This model employs a multi-module collaborative design for sentiment analysis tasks, enhancing overall semantic understanding while mitigating class imbalance issues. First, the MLM mechanism from the BERT pre-trained models is incorporated for data augmentation during the initial data phase. This mechanism implicitly increases minority class samples without altering the original semantics, exposing the model to more diverse semantic expressions during training and effectively mitigating the impact of data distribution imbalance on minority class recognition. Second, the Chinese-RoBERTa model is employed to extract character-level and word-level feature representations, enabling multilevel seman-

tic information mining from local to global contexts and providing foundational support for capturing key sentiment cues. Building upon this, a gated attention mechanism is introduced. It generates dynamic gate values through a fully connected layer and Sigmoid activation function, enabling weighted fusion of character-level and word-level features. Subsequently, the model integrates bidirectional contextual information via a BiGRU layer, enhancing its understanding of local sequences in short texts while capturing complete sentiment features by synthesizing contextual semantics in longer texts. Then, an MHSA layer is integrated. Multiple independent attention heads focus on critical text segments to capture global semantic information. The synergistic effect of BiGRU and MHSA not only improves the model's understanding of textual sequence information but also assigns higher weights to rare sentiment trigger words in minority classes, mitigating the sparse sample problem induced by category imbalance. Subsequently, the model employs an average pooling layer to aggregate MHSA outputs, simplify computations, strengthen global feature representation, and reduce model complexity. Finally, fully connected layers map pooled features to sentiment categories for precise classification.

Dropout prevents overfitting during training, and the AdamW optimizer performs regularization and parameter optimization, ensuring stable learning on minority samples and maintaining robust generalization capabilities.

3.1. The BERT-Driven Data Augmentation Module

To achieve data augmentation, this study incorporates the MLM mechanism from the BERT pre-trained model during the model's initial data phase. The Chinese BERT pre-trained model BERT-Base-Chinese from Hugging Face is employed to perform masking operations on the original text sequences. This involves randomly selecting and masking certain words or characters within the text and then using BERT to predict and generate replacement words for these masked positions, thereby creating additional variant texts. This approach increases the diversity of the training sample without altering the semantics of the original text, enhancing the model's robustness to different expressions and contexts while improving the generalizability of the sentiment analysis model.

Let the original text sequence be $X = \{x_1, x_2, \ldots, x_n\}$. During variant sample construction, some tokens are first replaced with the special token [MASK] according to a predefined masking probability (15%), yielding the masked sequence $X\prime$ with partially replaced tokens. The BERT model leverages its large-scale corpora-learned context modeling capabilities to predict replacements for masked positions in $X\prime$ candidate sequences $Y = \{y_1, y_2, \ldots, y_n\}$. Each y_i strives to maintain semantic consistency with its original position. Multiple distinct yet semantically plausible variant texts are obtained through multiple prediction and decoding strategies (Top-p sample truncation). Together with the original text, these variant samples form the augmented training corpus. This significantly enhances data diversity and coverage, enabling the model to encounter more varied semantic expressions during training. Consequently, the model's generalization capability is strengthened, making it more adaptable to diverse textual inputs in practical applications. This improves the reliability and usability of the sentiment analysis tasks.

These variant samples, combined with the original text, form the augmented training corpus. This significantly enhances data diversity and coverage, exposing the model to a wider range of semantic expressions during training. The generation of diverse texts through MLM implicitly expands this category, particularly for minority sentiment samples. This mitigates bias caused by imbalanced category distributions, thereby improving the model's generalization capability on minority classes and overall robustness.

3.2. The Feature Extraction and Integration Module

Extracting deep semantic features is crucial for enhancing the model's performance in text sentiment analysis tasks. Traditional BOW models or static word vector methods (e.g., Word2Vec, GloVe) struggle to simultaneously capture contextual dependencies and global semantic information, resulting in limited expressive power. To overcome this limitation, the Chinese-RoBERTa model provided by Hugging Face (model name: hfl/Chinese-RoBERTa-wwm-ext) is introduced as a feature extractor. Based on the RoBERTa architecture, this model is pre-trained on large-scale Chinese corpora, including Wikipedia, news articles, and Q&A communities. It incorporates a whole-word masking strategy to capture more granular, context-dependent semantic features at the character and word levels.

In the implementation process, the input text is first subjected to word segmentation and encoding, and then, context-aware semantic vector representations are generated using the hfl/Chinese-RoBERTa-wwm-ext model. This process not only captures granular linguistic features but also effectively models overall semantic structures, providing robust feature support for subsequent sentiment classification tasks. Therefore, the Chinese-RoBERTa model is employed to obtain character-level and word-level semantic representations for feature extraction in text sentiment analysis.

3.2.1. Principle of Character—Level Text Representation

Considering text as a sequence of many characters, if the input text is T, it can be expressed as a sequence of characters $T = \{c_1, c_2, \ldots, c_m\}$, where c_i stands for the ith character in the text. The Chinese RoBERTa model uses an internal character embedding layer to map each character c_i into a low-dimensional vector space. The corresponding character vector representation v_{c_i} is then obtained, which is expressed as follows: $v_{c_i} = CharacterEmbedding(c_i)$. The model can capture fine-grained sentimental features, such as the subtle sentimental information contained in some individual words with sentimental tendencies, by breaking the text into smaller character units and thereby obtaining vector representations.

3.2.2. Principle of Word—Level Text Representation

At the word level, the Chinese-RoBERTa model is first applied to perform the segmentation operation on the input text (assuming that some suitable Chinese segmentation method is used to segment the text into word sequences), and the segmented text is set as $W = \{w_1, w_2, \ldots, w_n\}$, where w_j denotes the jth word. Similarly, the word embedding layer inside the model will map each word w_j into a word vector v_{w_j}, which can be expressed as: $v_{w_j} = WordEmbedding(w_j)$.

3.2.3. Integrating Character-Level and Word-Level Features

In order for the model to understand the text from the details to the whole multi-level, and accurately capture the key sentimental information, it is necessary to comprehensively consider the character-level and word-level features, and concatenate the character-level vector representation sequence and the word-level vector representation sequence to obtain the final semantic representation of the text V.

Assuming that the character vector and word vector are adjusted to the same dimension after some uniform dimensionality processing, the character-level vector is represented as $\{v_{c_1}, v_{c_2}, \ldots, v_{c_m}\}$ and the word-level vector is represented as $\{v_{w_1}, v_{w_2}, \ldots, v_{w_n}\}$, and the concatenated text semantic vector V can be expressed as:

$$V = \{v_{c_1}, v_{c_2}, \ldots, v_{c_m}, v_{w_1}, v_{w_2}, \ldots, v_{w_n}\} \quad (1)$$

Through the above-mentioned approach, the model will be able to understand the text from multiple levels, focusing on the subtle sentimental clues carried by individual characters, but also grasping the semantic and sentimental information expressed at the word- level, which is more in line with the requirements of sentiment analysis for the comprehensiveness of semantic understanding of the text.

3.2.4. Feature Fusion with Gated Attention Mechanism

After obtaining character-level and word-level features, a gated attention mechanism is used to carry out feature fusion. This mechanism can automatically adjust the attention to different levels of features according to the specific conditions of the text, so as to enhance the feature expression ability of the model and optimize the effect of text sentiment analysis.

Specifically, the gated value is generated with the help of fully connected layer and Sigmoid activation function. Let the weight matrix of the fully connected layer be $W_g \in \mathrm{R}^{d_{in} \times d_{out}}$ (where d_{in} is the dimension of the input features, which is the dimension of the concatenated character-level and word-level feature vectors, i.e., $d_{in} = d_c + d_w$, d_{out} is the output dimension, which is set to 1, indicating that the generation of a gated value), and the bias vector is $d_{out} \in \mathrm{R}^{d_{out}}$.

For the input character-level and word-level feature vectors, they are first concatenated together to obtain the composite feature vector $x = \{\mathrm{c}_1, c_2, \ldots, c_m, \mathrm{w}_1, w_2, \ldots, w_n\}$ (with dimension d_{in}). Then the gated value g is computed through the fully connected layer with Sigmoid activation function as follows:

$$z = W_g x + b_g \tag{2}$$

$$g = \sigma(z) \tag{3}$$

where σ denotes the Sigmoid function, defined as $\sigma(z) = \frac{1}{1+e^{-z}}$, which maps the z value output from the fully-connected layer to the interval $(0, 1)$, and the resulting gated value g is used to control the fusion weights of character-level and word-level features. The fused feature vector f can be calculated as follows: $f = \mathrm{g} \cdot \mathrm{C} + (1 - \mathrm{g}) \cdot \mathrm{W}$, which realizes the dynamic adjustment of the weights of character-level and word-level features in the fusion process according to the gated value.

3.3. Sequence Information Processing Module

3.3.1. Principle and Role of BiGRU Layer

In the sequence information processing module, the BiGRU layer is crucial, which focuses on capturing text sequence information, and strengthens the model's ability to understand the text with the help of bi-directional context fusion information, which can effectively capture the embedded sentimental information in both short text and long text scenarios.

After the input is processed by the gated attention mechanism, let the feature sequence be $X = \{x_1, x_2, \ldots, x_t\}$, where x_t denotes the input feature vector at the moment t, and T is the length of the sequence. BiGRU consists of a forward GRU and a backward GRU, and the forward GRU processes the input sequences sequentially to compute the hidden state sequence $\overrightarrow{h}_t$, whose update formula is as follows:

$$\overrightarrow{z}_t = \sigma\left(W_z x_t + U_z \overrightarrow{h}_{t-1} + b_z\right) \tag{4}$$

$$\overrightarrow{r}_t = \sigma\left(W_r x_t + U_r \overrightarrow{h}_{t-1} + b_r\right) \tag{5}$$

$$\overrightarrow{h}_t = tanh\left(W_h x_t + U_h\left(\overrightarrow{r}_t \bigodot \overrightarrow{h}_{t-1}\right) + b_h\right) \tag{6}$$

where W_z, U_z, W_r, U_r, W_h, U_h are the corresponding weight matrices, b_z, b_r, b_h are the bias vectors, σ is the Sigmoid function, and $\odot$ denotes the element corresponding to the multiplication operation.

Similarly, the backward GRU processes the input sequence in reverse from the end of the sequence to obtain the hidden state sequence h_t, which is computed by the similar formula as the forward, except that the processing order is reversed. Finally, the output h_t of BiGRU is obtained by concatenating the forward and backward hidden states i.e.,: $h_t = \left[\overrightarrow{h}_t, \overleftarrow{h}_t\right]$.

3.3.2. Principle and Role of Multi-Head Self-Attention Layer

MHSA layer is also essential in sequence information processing, which relies on multiple different attention heads. Focusing on the key parts of the text, it strengthens the ability to capture the important features of sentiment, helps the model to properly deal with the complex information in the long text, and accurately mines the sentimental information.

If the features input to the MHSA layer are represented as $Q(Query)$, $K(Key)$, $V(Value)$, these are usually transformed appropriately by the output of the previous BiGRU layer, with dimensions d_q, d_k, d_v, respectively. for the single-head attention mechanism, the formula for calculating the attention score is as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

In MHSA, the attention mechanism with h heads is used, and h = 4 heads attention layer is chosen in this experiment, each attention head i ($i = 1, 2, \ldots, h$) has its own independent parameter matrix W_{Q_i}, W_{k_i}, W_{v_i} which is used to linearly transform the input Q, K and V to get the Q_i, K_i and V_i corresponding to the different heads, respectively. Attention output Attention is computed separately for each head:

$$Attention(Q,\ K, V) = softmax\left(\frac{Q_i {K_i}^T}{\sqrt{d_{k_i}}}\right)V_i \tag{8}$$

The outputs of the individual heads are concatenated and passed through a linear transformation layer (with weight matrix W_o) to obtain the final output O of the MHSA:

$$O = W_o[Attention_1(Q,\ K, V); Attention_2(Q,\ K,\ V); \ldots; Attention_h(Q,\ K,\ V) \tag{9}$$

Although the MHSA layer does not directly address class imbalance, it enhances contextual modeling and captures key features. This approach assigns higher weights to rare sentiment trigger words in minority class samples, effectively mitigating the under-representation risk in minority classes. Consequently, it improves the model's semantic understanding capabilities and classification robustness when handling imbalanced data.

3.4. Classification Prediction Output Module

This module is responsible for further mapping the text semantic vectors processed by sequence modeling and gated attention mechanism to the sentiment category labeling space to output the final classification prediction results. The two core structures here mainly include the fully connected layer and the Softmax normalization layer.

Let $H \in R^d$ denotes the text vector produced by the multi-head attention layer, where d indicates the dimension of the final fused feature representation. Firstly, this vector

is input to the fully connected layer to complete the linear mapping and obtain the unnormalized classification score vector $Z \in R^C$, where C is the number of categories for sentiment classification. In this experiment, $C = 3$, which corresponds to positive, neutral, and negative affective tendencies, respectively. The linear mapping process of the fully connected layer can be expressed in Equation (10):

$$Z = W_o H + b_o \tag{10}$$

where $W_o \in R^{C \times d}$ is the weight matrix and $b_o \in R^C$ is the bias term.

Secondly, the classification score vector z is converted to an output $\hat{y} \in R^C$ in the form of a probability distribution through the Softmax layer, whose probability formula for class i is shown in Equation (11):

$$\hat{y}_i = \frac{e^{Z_i}}{\sum_{j=1}^{C} e^{Z_j}} i = (1, 2 \ldots, C) \tag{11}$$

That is, $\hat{y}$ is obtained which can be used as the model's prediction of the sentiment classification of the input text.

3.5. Loss Function and Optimization Strategy Module

The model chooses cross entropy as the loss function and uses AdamW optimizer [28] for parameter updating to ensure that the model gradually converges to the performance optimal solution during the training process.

An iterative training strategy with a fixed number of rounds is adopted in the model training phase, combined with the batch grouping mechanism to handle large-scale samples. Meanwhile, to prevent model overfitting, an early stopping strategy and a validation set monitoring mechanism are introduced to update the model weights only when the validation set performance continues to improve.

4. Experimental Results

4.1. Dataset Acquisition and Preprocessing

This study focuses on the application of AI courses in education and selects the Chinese University MOOC platform (MOOC), the Baidu PaddlePaddle AI Studio, and iFLYTEK AI College as data sources. These platforms demonstrate strong course volume, user participation, and content coverage. The corresponding review data provide a reliable reflection of the actual performance of the courses in terms of teaching quality, content design, and learning experience. The MOOC platform integrates high-quality courses from leading institutions and is widely acknowledged for its excellence in artificial intelligence education. The Baidu PaddlePaddle AI Studio has developed a comprehensive curriculum system that covers the entire AI learning process and has collected substantial evaluation data. With its solid technological foundation, iFLYTEK AI College delivers courses that combine strong academic rigor with practical applicability, and its evaluations offer valuable insights for assessing instructional quality and learning outcomes.

We used web crawler technology to crawl the comment data of 57 AI-related courses from the aforementioned platforms, which cover key areas such as machine learning and deep learning. After recording the key information in the comments, 88,572 comments were collected and distributed among the courses (see Table 1 for details of the distribution).

Table 1. Status of data on comments.

Data Sources	Number of Courses	Number of Reviews
MOOC	27	37,886
Baidu PaddlePaddle AI Studio	13	4892
iFLYTEK AI College	17	15,794

A term thesaurus was constructed by extracting English words from the reviews. All reviews containing thesaurus terms are retained. Concurrently, corresponding rules are developed to remove special characters and other contents to avoid accidental deletion of technical terms. In terms of improving data quality, blank and repetitive reviews are deleted; short texts containing the terms are retained according to rules; HTML tags are cleaned up; traditional and simplified text are standardized; and all-English and all-numeric reviews and reviews with less than 6 words are eliminated. Word segmentation is used to process the text to remove repetitive parts and reduce repetition-related interference. Then, meaningless high-frequency words are removed, and a threshold is set to retain high-value words, combined with the deactivation word list, to improve the stability of theme modeling. After the above processing, 54,404 structured data are finally retained. This enhances the readability of the text, reduces data noise, and provides strong support for subsequent analysis.

4.2. Data Labeling and Classification

The Taiwan NTUSD Sentiment Dictionary and Dalian University of Technology Sentiment Dictionary were selected to enhance the efficiency and analytical accuracy of sentiment analysis on Chinese course reviews. The Taiwan NTUSD dictionary boasts broad applicability and a rich vocabulary of sentiment terms. This dictionary not only distinguishes positive and negative sentiments but also assigns intensity scores to corresponding sentiments (negative sentiment scored -2, positive sentiment scored +1). This facilitates rapid tri-classification annotation, enabling the preliminary determination of a review's sentiment category. Meanwhile, the Dalian University of Technology Sentiment Dictionary provides more refined vocabulary and scoring systems, making it suitable for complex sentiment analysis scenarios. It accurately assesses each review's sentiment, thereby enhancing the overall analysis precision. By applying these dictionaries sequentially, first using the NTUSD dictionary for preliminary classification to determine the general sentiment orientation of reviews, then employing the Dalian University of Technology dictionary for refined scoring, the synergistic advantages of both are fully leveraged. This ensures high-precision analysis results, laying a solid foundation for future research. To validate annotation quality, a 5% random sample from the original dataset was manually reviewed by the research team. The calculated Cohen's kappa coefficient (0.89) demonstrated strong annotation consistency. The dataset exhibited a significant imbalance after annotation: 37,635 positive comments, 10,385 neutral comments, and 6384 negative comments.

The dataset is divided into training, validation, and test sets at a ratio of 70%/15%/15% to ensure model training and evaluation of performance. Each subset maintains the same sentiment label distribution as the entire dataset, thereby preventing the effect of partitioning bias on model performance. The training set is used for model learning, the validation set is used for parameter tuning and early stopping mechanisms, and the test set serves as an independent performance evaluation benchmark. The main comparison experiments and ablation studies were conducted using the full dataset to comprehensively assess the model's robustness under data imbalance and maximum sample size. Verification results based on manually reviewed samples demonstrate that the model achieves an accuracy of 0.89, precision of 0.87, recall of 0.91, and F1 score of 0.91 in

the sentiment classification tasks. This validates the reliability of both annotation quality and model performance.

4.3. Experimental Parameter Settings

To ensure the stability and effectiveness of the sentiment analysis model in the training process, hyperparameter settings are used for key parts of the model structure and training process. It mainly covers data preprocessing parameters, pre-trained model configuration, feature extraction module, attention mechanism structure, feature fusion and classification module, loss function, and optimizer settings. Each parameter is selected considering the model's expressive and generalization abilities, while some parameters are fine-tuned along with the experimental scenarios. Table 2 shows the parameters' configurations.

Table 2. Table of model parameters.

Parameter	Value
Maximum sequence length	256
Batch Size	16
BiGRU Hidden Layer Dimension	256
BiGRU Output Dimension	512
Number of Multi-head Attention layers	4
Gated Full Connectivity Layer Output Dimension	1
Output Dimensions in Full connectivity layer	3
Dropout rate	0.3
Learning rate	2×10^{-5}

4.4. Baseline Model Comparison Experimental Analysis

4.4.1. Experiments on Three-Classification Sentiment Analysis

Five representative and varied Chinese sentiment datasets were selected to comprehensively evaluate the performance and generalization capabilities of sentiment classification models across diverse scenarios: simplifyweibo_4_moods [29], online_shopping_10_cats [30], dmsc_v2 [31], the Chinese Taobao [32] review dataset (hereafter referred to as cn-taobao), and a dataset constructed by the authors. These datasets span diverse text sources, including social media, e-commerce reviews, and movie critiques. They exhibit robust textual complexity and sentiment label diversity, providing strong support for evaluating model performance across cross-domain scenarios.

The simplifyweibo_4_moods dataset was compiled from Sina Weibo corpora by the Natural Language Processing Laboratory at Nanjing University and is publicly available on the Hugging Face platform. This dataset contains approximately 360,000 short Weibo texts with sentiment labels, originally categorized into four sentiments: "joy," "anger," "disgust," and "sadness." The labels were remapped to align with the three-classification task in this study: "joy" was assigned to positive sentiment, "anger" and 'disgust' were merged into negative sentiment, and "sadness" was assigned to neutral sentiment. The final three-class dataset comprises 200,000 positive comments, 50,000 neutral comments, and 100,000 negative comments based on this mapping. Second, the Fudan University School of Computer Science and Technology released the online_shopping_10_cats dataset, which is included in the Chinese NLP corpus collection, on GitHub. This dataset covers e-commerce reviews from 10 product categories, totaling approximately 60,000 entries. This study selected and retained reviews with moderate star ratings to prevent extreme reviews from biasing model training, reconstructing three sentiment categories based on review content. The final sample distribution comprises 34,620 positive reviews, 17,824 neutral reviews, and 7556 negative reviews. Third, the dmsc_v2 dataset (Douban Movie Short Comments v2), released by the Institute of Computing Technology, Chinese Academy of

Sciences, contains more than 2 million short movie reviews from Douban, covering 28 films and more than 700,000 users. Considering computational resources and experimental efficiency, 100,000 entries were randomly sampled from the dataset and categorized based on rating ranges: 1–2 stars as negative, 3 stars as neutral, and 4–5 stars as positive. The final dataset comprises 63,972 positive reviews, 25,341 neutral reviews, and 10,687 negative reviews after classification. Fourth, the cn-taobao dataset originates from the Alibaba Cloud's Tianchi platform and is one of the most widely used Chinese review datasets in e-commerce scenarios. The original sentiment labels are explicitly annotated and adopted in this study. The sample distribution comprises 10,657 positive reviews, 2561 neutral reviews, and 3436 negative reviews. Additionally, a custom Chinese review dataset was constructed based on online user feedback to better align with this research scenario, supplementing the diversity of sentiment expressions in specific domains.

This study selects a total of six models for comparative analysis, including TCN-BiLSTM, IDM-GAT, RoCSAOL, Transformer-TextCNN, RoBERTa-BiGRU-Double Attention, and the Proposed Model. The five baseline models used for comparison are briefly described as follows:

(1) TCN-BiLSTM [33]: The hybrid word embedding technique is used to realize text vectorization, and after capturing local features by time-series convolutional network, the sequence information is extracted using improved bidirectional LSTM, weighting features by combining with the attention mechanism, and finally completing the classification by Softmax.

(2) IDM-GAT [34]: This model enhances sentiment information expression by analyzing each aspect word and its context in a sentence using relative positional coding. The BGRU layer encodes the vector representation of aspect words and context, the interactive attention mechanism captures the word-level dependencies between aspect words and sentiment contexts, and the word nodes and their dependencies at the sentence level are weighted by the multi-head dependency graph attention mechanism. The multi-head dependency graph attention mechanism weights the dependencies. Finally, the two levels of attention vectors are aggregated to improve the model's understanding of sentiment analysis.

(3) RoCSAOL [35]: This model combines the pre-trained RoBERTa model, optimizes the ordered neuron LSTM through convolution and pooling operations, and uses a self-attention mechanism to mine sentiment semantics in the text.

(4) Transformer-TextCNN [36]: Combining the absolute and sinusoidal position encoding mechanisms of BERT to enhance the perception of location information, the global semantic modeling capability of Transformer and the advantage of TextCNN in capturing local key features to collaboratively extract global and local features of Chinese short text.

(5) RoBERTa-BiGRU-Double Att [37]: RoBERTa is utilized to obtain context-aware word vectors, and BiGRU captures the sequence features at the character level, introducing the double attention mechanism to compute the weight distribution of characters on the overall semantics and within the sentence.

As shown in Table 3, the Proposed Model demonstrates significant performance advantages on multiple public and self-built datasets, especially in the core evaluation metrics of F1score, classification ability under the AUC curve, and precision–recall balance.

In the F1 score metric comparison, the model obtains the highest score of 0.9345 on the sim_weibo dataset, which is ahead of TCN-BiLSTM (0.9254) and IDM-GAT (0.9007), and improves about 24.42% compared to RoCSAOL (0.6903). Meanwhile, the model still maintains a high level in other datasets, such as online_10, dmsc_v2, cn-taobao, and self-built datasets, showing strong generalization ability. At the recall level, the average score of this model is 0.8969, which is significantly higher than TCN-BiLSTM (0.8888) and IDM-GAT (0.7955) and improves by 22.33% compared to RoCSAOL (0.6736), reflecting the robustness and completeness in covering sentiment categories.

Table 3. Comparison of this model with other models (three classifications).

Model	Dataset	Precision	Recall	F1 Score	AUC
TCN-BiLSTM	sim_weibo	0.9266	0.9249	0.9254	0.9714
	online_10	0.8671	0.8737	0.8750	0.9565
	dmsc_v2	0.9054	0.9107	0.9082	0.9437
	cn-taobao	0.8269	0.8341	0.8293	0.9046
	self-built	0.9129	0.9006	0.9066	0.9513
IDM-GAT	sim_weibo	0.9068	0.8977	0.9007	0.9585
	online_10	0.7391	0.6836	0.7009	0.8748
	dmsc_v2	0.8024	0.7936	0.8017	0.8762
	cn-taobao	0.6954	0.7029	0.6991	0.8148
	self-built	0.8976	0.8997	0.8961	0.9332
RoCSAOL	sim_weibo	0.7273	0.7412	0.6903	0.8260
	online_10	0.6317	0.6198	0.6046	0.7984
	dmsc_v2	0.7098	0.7124	0.7103	0.8247
	cn-taobao	0.5761	0.5809	0.5767	0.7032
	self-built	0.7140	0.7135	0.6730	0.8297
Transformer-TextCNN	sim_weibo	0.9257	0.9271	0.9301	0.9724
	online_10	0.8764	0.8852	0.8819	0.9438
	dmsc_v2	0.9093	0.9104	0.9081	0.9369
	cn-taobao	0.8432	0.8327	0.8398	0.9154
	self-built	0.9185	0.8975	0.9128	0.9653
RoBERTa-BiGRU-Double Att	sim_weibo	0.8876	0.8792	0.8803	0.9347
	online_10	0.8029	0.8257	0.8091	0.9139
	dmsc_v2	0.8439	0.8491	0.8502	0.9026
	cn-taobao	0.7892	0.7743	0.7675	0.8321
	self-built	0.8264	0.8157	0.8259	0.9027
Proposed Model	sim_weibo	0.9370	0.9341	0.9345	0.9737
	online_10	0.8963	0.8550	0.8568	0.9428
	dmsc_v2	0.9249	0.9091	0.9124	0.9427
	cn-taobao	0.8517	0.8496	0.8503	0.9219
	self-built	0.9317	0.9366	0.9240	0.9643

This study employs the F1 score as the core metric and plots a bar chart for comparison with current mainstream models. As shown in Figure 2, in most cases, the bar height corresponding to this model occupies a leading position across five distinct domain text datasets, visually demonstrating its outstanding classification performance.

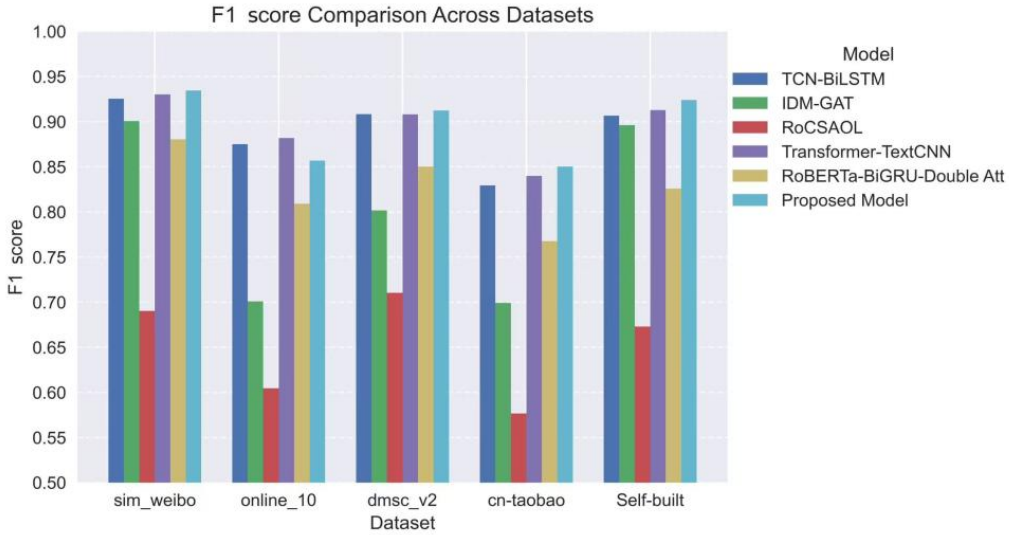


Figure 2. Bar chart of comparing the F1 score on various datasets.

Observing the trend in bar height variations, this model exhibits the narrowest performance fluctuation range (0.8503–0.9345) across the five datasets, with a standard deviation of only 0.034-significantly lower than other models. This model demonstrated the smallest decrease in bar height on the online_10 and cn-taobao datasets, indicating strong ODG capabilities. In contrast, the IDM-GAT model exhibits a pronounced performance collapse on the online_10 dataset (marked by a steep drop in bar height). Its F1 score plummeted from 0.9007 on sim_weibo to 0.7009, representing a 19.98% decline. This highlights the model's sensitivity to shifts in data distribution.

Our model achieved a maximum F1 score of 0.9345 on the sim_weibo dataset, which is 0.44% higher than the second-best Transformer-TextCNN model (0.9301). Our model achieved an F1 score of 0.9240 on the self-built dataset, surpassing the Transformer-TextCNN model's 0.9128 by 1.12%, demonstrating the model's strong generalization capability on real-world datasets. Notably, on the dmsc_v2 dataset, our model outperformed all the comparison models with an F1 score of 0.9124. This represents a significant improvement over both TCN-BiLSTM (0.9082) and Transformer-TextCNN (0.9081), indicating the unique advantage of the model in handling complex textual features.

4.4.2. Experiments on Binary Sentiment Analysis

This study aims to validate the effectiveness of the Proposed Model in binary sentiment classification. To this end, a comparative analysis was conducted using both a self-constructed dataset and publicly available corpora. Three public datasets were selected for comparison: weibo_senti_100k [30] (containing 119,988 Weibo comments), waimai_10k [30] (with 11,987 food delivery reviews), and ChnSentiCorp_htl_all [30] (comprising 7357 hotel reviews). All datasets consist of short Chinese texts, aligning with the self-built data in terms of textual characteristics to ensure validity in cross-scenario evaluations. Since both our model and the self-built dataset were originally designed for three-class classification, a sentiment lexicon-based annotation method was applied to convert the data into binary classes for the purpose of this study. The weibo_senti_100k dataset, curated by the Natural Language Processing Laboratory at Nanjing University and originally part of the Chinese-NLP-Corpus collection, is publicly available on Hugging Face. It contains 69,988 positive and 50,000 negative comments, and is widely used in sentiment analysis within social media contexts. The waimai_10k dataset was assembled by Tsinghua University researchers and open-source contributors, also included in Chinese-NLP-Corpus and released on Hugging Face. It includes 4000 positive and 7987 negative reviews, reflecting user-generated content characteristics in the food delivery domain, making it suitable for open-domain sentiment analysis. ChnSentiCorp_htl_all, another dataset from Chinese-NLP-Corpus available on Hugging Face, contains 5573 positive and 1784 negative comments. It focuses on consumer reviews of hotel services, facilities, and environments, with relatively formal text styles appropriate for structured review parsing. The self-constructed dataset consists of 45,063 positive and 9341 negative samples. Experiments across these four datasets allow for a comprehensive evaluation of the model's robustness and generalization capability across diverse short-text domains.

The following six comparison models were used in the experiments:

(1) BERT-CNN-attention [38]: The pre-training model based on the BERT obtains contextually relevant word vectors, and a convolutional neural network is used to extract key local features, which are fused through the attention mechanism for classification prediction.

(2) Bi-LSTM-CNN [8]: Bi-directional LSTMs and CNNs are used in parallel to extract long-range dependent features and deep abstract features, respectively, and improve the classification performance through feature fusion.

(3) TextCNN-BiLSTM-Att [7] combines convolutional network and bidirectional LSTM to capture local features and global sequence information, respectively, realizing multi-level feature dynamic fusion through self-attention mechanism.

(4) BERT-attention-Fasttext [39] utilizes BERT for word vectorization to generate dynamic word vectors containing contextual semantic information, and captures the relationship between global and local through the attention mechanism.

(5) ERNIE-BiLSTM [40]: a pre-trained language model ERNIE is used to generate dynamic word vectors, and semantic features are extracted through a bi-directional LSTM network to complete sentiment classification.

(6) ERNIE-CNN [41]: A fine-tuned ERNIE is utilized to extract features from the preprocessed comment dataset to generate word vectors containing comprehensive semantic information. The deep CNN component captures localized features from text, thus capturing semantic nuances at multiple levels.

As shown in the experimental results in Table 4, this model demonstrates significant advantages in the sentiment classification task on the wei-bo-100k dataset. In terms of classification accuracy, the model obtained a high precision of 0.9298, which achieved an absolute improvement of 13.51% and 24.83% compared to BERT-CNN-attention (0.7947) and BiLSTM-CNN (0.6815), respectively, and fully demonstrates its superiority in positive sentiment discrimination.

Table 4. Comparison of the model in this paper with other models (dichotomous).

Model	Dataset	Precision	Recall	F1 Score	AUC
BERT-CNN-attention	wei-bo-100k	0.7947	0.7986	0.8011	0.9028
	waimai_10k	0.7287	0.7194	0.7209	0.8324
	ChnSentiCorp_	0.6919	0.6787	0.6659	0.7422
	self-built	0.8149	0.8207	0.8176	0.9124
Bi-LSTM-CNN	wei-bo-100k	0.6815	0.6761	0.6808	0.7871
	waimai_10k	0.6376	0.6319	0.6401	0.7195
	ChnSentiCorp_	0.6053	0.6133	0.6110	0.6904
	self-built	0.6629	0.6497	0.6632	0.7127
TextCNN-BiLSTM-Att	wei-bo-100k	0.8231	0.8312	0.8328	0.9282
	waimai_10k	0.7854	0.7797	0.7842	0.8567
	ChnSentiCorp_	0.7344	0.7720	0.7660	0.8701
	self-built	0.8007	0.7986	0.8018	0.9062
BERT-attention-Fasttext	wei-bo-100k	0.7243	0.7189	0.7207	0.8359
	waimai_10k	0.6679	0.6703	0.6718	0.7724
	ChnSentiCorp_	0.6482	0.6531	0.6489	0.7593
	self-built	0.6971	0.7024	0.6958	0.8124
ERNIE-BiLSTM	wei-bo-100k	0.8641	0.8597	0.8624	0.9152
	waimai_10k	0.6943	0.7025	0.6932	0.7872
	ChnSentiCorp_	0.6339	0.6471	0.6409	0.7382
	self-built	0.7027	0.6984	0.7019	0.7792
ERNIE-CNN	wei-bo-100k	0.7864	0.7791	0.7847	0.8356
	waimai_10k	0.7325	0.7401	0.7380	0.7139
	ChnSentiCorp_	0.7069	0.7127	0.7082	0.7947
	self-built	0.7597	0.7647	0.7609	0.7239
Proposed Model	wei-bo-100k	0.9298	0.9295	0.9267	0.9742
	waimai_10k	0.8682	0.8707	0.8695	0.9271
	ChnSentiCorp_	0.7509	0.7685	0.7546	0.8612
	self-built	0.9158	0.9104	0.9087	0.9698

In terms of recall metrics, this model achieves a recall of 0.9295, compared with BERT-CNN-attention (0.7986) and TextCNN-BiLSTM-Att (0.8312), which is an improvement

of 13.09% and 9.83%, respectively, showing its stronger coverage ability and detection completeness for negative sentiment categories.

In terms of the F1 score of the comprehensive performance index, this model reaches 0.9267, which is 20.60% and 6.43% higher than the traditional models BERT-attention-Fasttext (0.7207) and ERNIE-BiLSTM (0.8624), respectively, indicating that it achieves a more desirable balance between classification accuracy and robustness. In addition, the present model obtains a high score of 0.9742 in the AUC, which is much higher than that of BERT-CNN-attention (0.9028) and BiLSTM-CNN (0.7871), with classification performance boundaries improved by 7.14% and 18.71%, respectively. This further demonstrates the advantages of the present model in terms of category separability and generalization ability.

4.5. Experimental Analysis and Discussion

4.5.1. Module Validation

To evaluate the functional contribution of the model components, ablation experiments were conducted on the underlying architecture (Chinese-RoBERTa + Dual-Stream Transformer + Gated Attention + Generative Enhancement). Given the unbalanced nature of the data, P/R/F1 score/AUC was selected as the evaluation index, and the experimental results are detailed in Table 5.

Table 5. Results of ablation experiments.

Experiment	Precision	Recall	F1 Score	AUC
Proposed model	0.9317	0.9366	0.9240	0.9643
Fine-tuned RoBERTa	0.9039	0.9011	0.8964	0.9499
Freeze RoBERTa	0.8287	0.7633	0.7906	0.9215
Remove data enhancements	0.8995	0.8967	0.8950	0.9400
Remove BiGRU	0.8902	0.8932	0.9010	0.9367
Remove MHSA	0.9087	0.8905	0.9079	0.9477
Remove gated mechanism	0.9057	0.9089	0.9067	0.9458
Remove gated attention mechanism	0.8894	0.8911	0.8853	0.9498
Remove Transformer	0.8301	0.8292	0.8389	0.9246
Remove character feature extraction	0.7918	0.7932	0.7885	0.8941

The experimental results show that the full model achieves the best performance in both F1 score and AUC metrics (F1 score = 0.9240, AUC = 0.9643), significantly outperforming all ablation variants and the strong baseline fine-tuned RoBERTa (F1 score = 0.8964, AUC = 0.9499). Notably, the fine-tuning experiments for fine-tuned RoBERTa employed tightly controlled hyperparameter settings (MAX_LEN = 256, BATCH_SIZE = 16, LEARNING_RATE = 2×10^{-5}, NUM_CLASSES = 3); thus, this result can be considered an upper bound on RoBERTa's performance on this task. In comparison, the full model still achieves an F1 score improvement of approximately 2.8%, demonstrating that the proposed downstream architecture is not redundant but rather introduces independently valuable structural innovations on top of the semantic representation of RoBERTa.

The following conclusions can be drawn from the experimental results:

(1) In terms of context modeling capabilities, although RoBERTa has captured global dependencies through the transformer layer, its output mainly reflects the "general correlation" learned in the pre-trained phase and lacks explicit modeling of local sequential patterns in downstream sentiment classification tasks. The introduction of BiGRU provides a structured inductive bias for the model, which can reintegrate RoBERTa's token representation sequentially and progressively, thereby improving the model's robustness in short texts and noisy samples. The additionally stacked MHSA layer is not a simple repetition of RoBERTa's attention mechanism but a task-specific attention learner that can refocus

on key trigger and polarity words in downstream sentiment classification. The results of the ablation experiment further verified the necessity of the two: the F1 score dropped to 0.9010 and 0.9079, respectively, after removing BiGRU or MHSA, indicating that the two provide complementary gains beyond the RoBERTa representation.

(2) Regarding feature fusion, the gated attention mechanism proposed in this paper dynamically weights character- and word-level features rather than simply concatenating or summing them. This mechanism allows the model to flexibly adjust the importance of the two types of features based on the semantic characteristics of different samples. The ablation results show that removing the gated mechanism reduces the F1 score to 0.9067, a performance degradation of approximately 1.7%, validating the effectiveness of dynamic weighting in avoiding redundancy and feature interference.

(3) The effectiveness of the data augmentation module has also been fully demonstrated. After removing it, the performance of the model across all metrics decreased by an average of approximately 3.3%, with Recall and F1 score performance dropping particularly sharply. This demonstrates that MLM-based augmentation can significantly alleviate class imbalance, increase corpus diversity, and enhance generalization capabilities.

(4) In an end-to-end training strategy, freezing RoBERTa's parameters resulted in a significant performance drop (F1 score = 0.7906, AUC = 0.9215), demonstrating that relying solely on downstream architecture without updating pre-trained weights fails to fully realize the model's potential. This result highlights the need for full parameter fine-tuning to fully leverage pre-trained semantic knowledge and adapt to downstream tasks.

We conducted 10 independent runs of each experimental setting with different random seeds to verify the proposed model's stability and effectiveness and its key components. Precision, recall, F1 score, and AUC mean and standard deviation were calculated (Table 6).

Table 6. Results of ablation experiments (mean ± std).

Experiment	Precision	Recall	F1 Score	AUC
Proposed Model	0.932 ± 0.006	0.937 ± 0.005	0.924 ± 0.007	0.964 ± 0.004
Fine-tuned RoBERTa	0.904 ± 0.007	0.901 ± 0.008	0.896 ± 0.009	0.950 ± 0.005
Freeze RoBERTa	0.829 ± 0.012	0.763 ± 0.014	0.791 ± 0.013	0.922 ± 0.006
Remove data enhancements	0.900 ± 0.008	0.897 ± 0.009	0.895 ± 0.010	0.940 ± 0.006
Remove BiGRU	0.890 ± 0.010	0.893 ± 0.011	0.901 ± 0.009	0.937 ± 0.005
Remove MHSA	0.909 ± 0.009	0.891 ± 0.010	0.908 ± 0.008	0.948 ± 0.004
Remove Gated mechanism	0.906 ± 0.008	0.909 ± 0.009	0.907 ± 0.009	0.946 ± 0.005
Remove Gated Attention mechanism	0.889 ± 0.011	0.891 ± 0.010	0.885 ± 0.009	0.950 ± 0.005
Remove Transformer	0.830 ± 0.012	0.829 ± 0.013	0.839 ± 0.011	0.925 ± 0.007
Remove Character Feature Extraction	0.792 ± 0.015	0.793 ± 0.014	0.789 ± 0.013	0.894 ± 0.010

The following conclusions can be drawn from the experimental results:

(1) The proposed full model outperforms its variants in all metrics and exhibits lower standard deviations, demonstrating not only superior average performance but also greater robustness under different start conditions.

(2) Paired *t*-tests were used to statistically analyze performance differences among key modules. The results showed that after removing the gated attention mechanism, the F1 score of the model dropped significantly from 0.924 ± 0.007 to 0.885 ± 0.009 (t = 50.80, $p = 2.23 \times 10^{-12}$, $p < 0.05$), indicating that this module significantly improved model performance. When the data augmentation module was removed, the F1 score dropped from 0.924 ± 0.007 to 0.895 ± 0.010 (t = 30.67, $p = 2.04 \times 10^{-10}$, $p < 0.05$), also showing a significant difference. This demonstrates that the gated attention mechanism effectively integrates character-level and word-level features, improving feature interaction and in-

formation utilization, while the data augmentation strategy plays a key role in expanding corpus diversity and alleviating class imbalance.

(3) Removing the BiGRU or MHSA results in F1 score dropping to 0.901 ± 0.009 and 0.908 ± 0.008, respectively, validating their complementary contributions beyond the RoBERTa representation. Removing the gated mechanism also results in F1 score dropping to 0.907 ± 0.009, confirming the superiority of dynamic weighted fusion over simple concatenation. Furthermore, removing the transformer encoder or retaining only single-character feature extraction significantly degrades model performance (F1 score were 0.839 ± 0.011 and 0.789 ± 0.013, respectively), demonstrating that deep context modeling and multi-granular feature representation are equally essential for sentiment classification.

(4) Freezing the RoBERTa parameters results in an F1 score of only 0.791 ± 0.013, which is significantly lower than that of full fine-tuning. This result highlights the need for full parameter training to unlock the potential of pre-trained models and adapt them to downstream tasks.

The experimental results not only verify the independent contribution and complementary role of each module in performance improvement but also fully demonstrate the innovation and effectiveness of the proposed gated attention mechanism in multi-granularity semantic fusion.

4.5.2. Model Performance Visualization

This study systematically analyzes the training process and results using multidimensional visualization methods to evaluate the proposed model's performance. During the model performance assessment phase, systematic analysis of validation metric trends across different training epochs reveals that the model achieves optimal performance at the 4th epoch (as shown in Figure 3). The training curve clearly demonstrates that the proposed model exhibits outstanding performance across multiple key evaluation metrics. The model converged rapidly within the first four training cycles: the training loss decreased significantly from an initial value of 0.310 to 0.110, while the validation loss simultaneously dropped from 0.245 to 0.120. This phenomenon fully demonstrates the model's efficient learning capability. Notably, the F1 score steadily improved from 0.8947 to 0.9240 during the first four epochs and remained stable in subsequent training cycles (finally stabilizing at 0.9102). This indicates that the model achieves an optimal balance between precision and recall, which are two mutually constraining metrics.

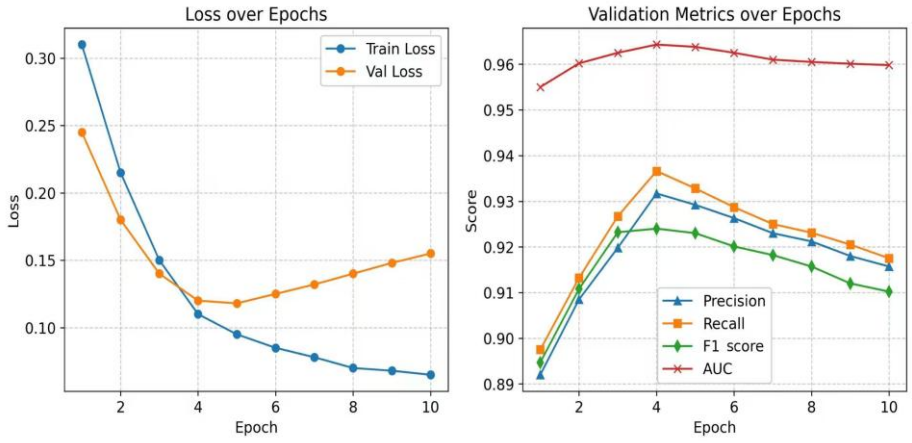


Figure 3. Line chart of validation metric trends.

For the three-classification task, we comprehensively evaluated the performance of six text classification models (TCN-BiLSTM, IDM-GAT, RoCSAOL, Transformer-TextCNN, RoBERTa-BiGRU-Double Att, and our proposed model) on five standard datasets

(sim_weibo, online_10, dmsc_v2, cn-taobao, and self-built) through systematic experimental design and visual analysis, as illustrated in the Figure 4. Figure 4 depicts the metric evolution during training, which primarily serves to analyze the model's convergence characteristics and learning dynamics.

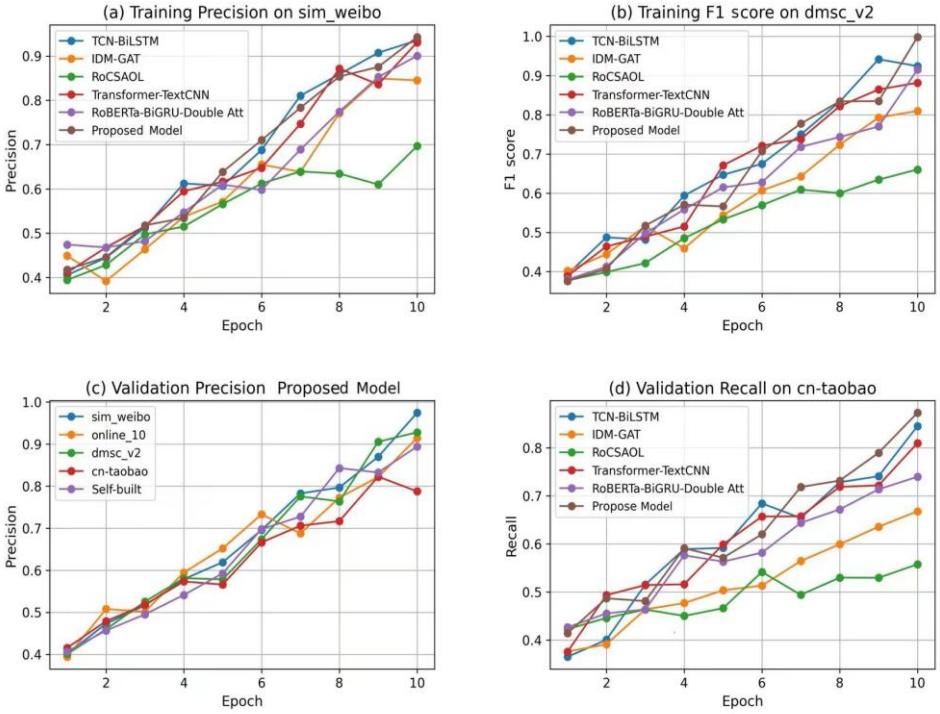


Figure 4. Iteration charts of on different datasets with different metrics.

Figure 4a demonstrates that our model exhibits a significant precision advantage during training on the sim_weibo dataset, ultimately achieving a peak value of 0.95. This represents improvements of 1.0% and 3.0% over the next-best TCN-BiLSTM (0.94) and Transformer-TextCNN (0.92) models, respectively. The convergence curve reveals that the proposed model rapidly converges to a stable state within the first three training epochs, whereas in comparison models, 5–6 epochs are typically required to reach stability. This rapid convergence demonstrates the model's efficient feature learning capability. The RoCSAOL model performs relatively poorly, achieving only 0.70 precision, highlighting the limitations of traditional methods when processing short social media texts.

Figure 4b demonstrates that our model achieves the highest overall performance with an F1 score of 0.94 on the dmsc_v2 dataset, outperforming baseline models by an average of 3–5 percentage points. TCN-BiLSTM and Transformer-TextCNN followed closely with 0.93 and 0.91, respectively, while IDM-GAT and RoBERTa-BiGRU-Double Att demonstrated moderate performance (0.82–0.83). RoCSAOL lagged significantly behind at 0.68. Observing the convergence process, our model not only achieved superior final performance but also maintained a consistent lead throughout training without noticeable performance fluctuations.

Figure 4c shows that the proposed model maintains outstanding performance across the five domain datasets. Specifically, it achieves a peak precision of 0.92 on the sim_weibo dataset, 0.90 on our self-built dataset, 0.91 on dmsc_v2, and 0.88 on online_10. Even on the cn-taobao e-commerce dataset, it attains a respectable level of 0.84. The standard deviation of precision across the five datasets was only 0.028, which was significantly lower than the 0.05–0.08 fluctuation range observed in the comparison models.

Figure 4d demonstrates that our model significantly outperforms other approaches in recall comparisons on the cn-taobao dataset with a recall of 0.84. This represents a 2.0% improvement over TCN-BiLSTM and Transformer-TextCNN (both at 0.82), a 16.0% improvement over IDM-GAT (0.68), and a substantial 29.0% improvement over RoCSAOL (0.55). The high recall rate of this model indicates its ability to effectively capture the features of challenging samples, thereby reducing the number of missed detections.

Figure 5 presents a radar chart comparing multiple metrics to comprehensively evaluate the performance of 7 distinct models across four datasets (wei-bo-100k, waimai_10k, ChnSentiCorp_, and our self-built dataset) in binary sentiment analysis experiments.

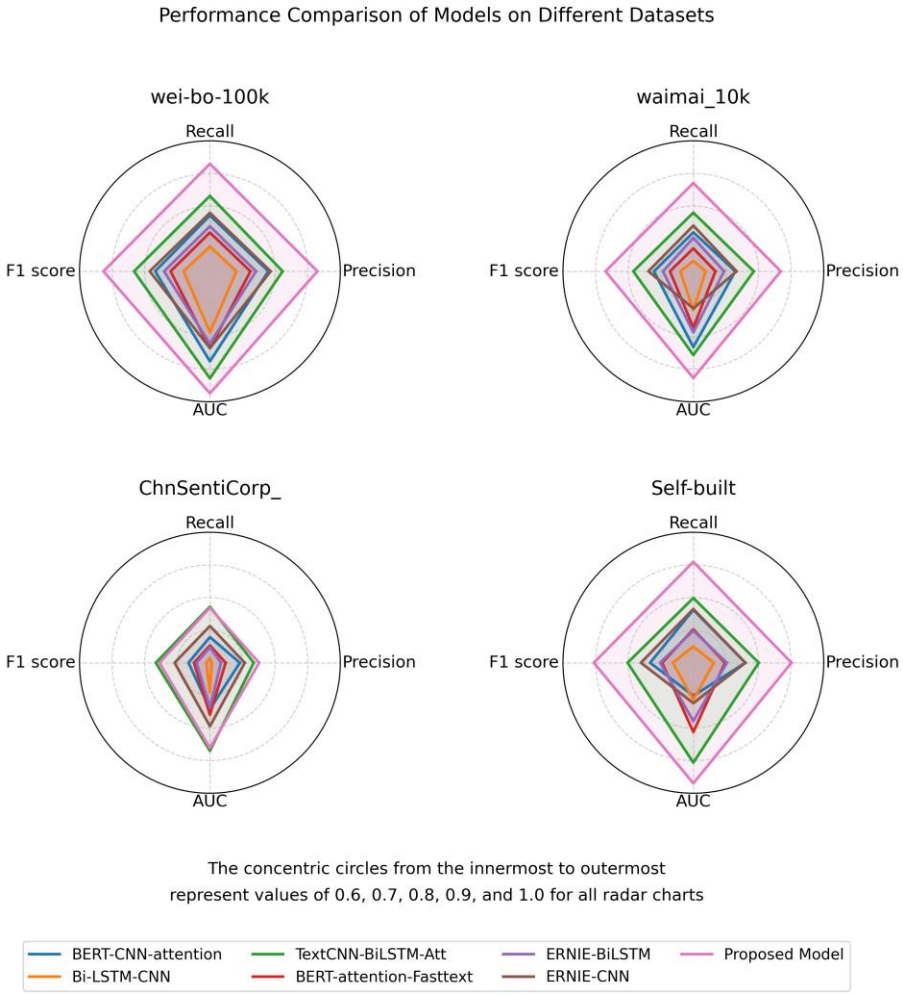


Figure 5. Radar Chart for multi-indicator comparison.

As shown in Figure 5, this model demonstrates comprehensive and significant superior performance, with its core competitiveness manifested in three aspects: First, in terms of metric balance, the model achieves a perfect equilibrium between Precision (0.9298) and Recall (0.9295) on the weibo-100k dataset (with a difference of only 0.0003), far surpassing the average 0.05 gap observed in other models. This balance enables its F1 score (0.9267) to approach the optimal theoretical value. Second, regarding cross-dataset stability, the AUC standard deviation across 4 datasets is only 0.046, demonstrating stronger robustness compared to TextCNN-BiLSTM-Att (0.058) and ERNIE-CNN (0.084). Third, its F1 score (0.7546) on the ChnSentiCorp_ dataset maintains a relative advantage over the second-place model (0.7660) in domain adaptability. Moreover, its AUC value (0.9698) on the self-built specialized domain dataset approaches the theoretical maximum, demonstrating outstanding domain knowledge capture capabilities.

4.5.3. Performance Analysis Under an Uneven Distribution of Categories

In sentiment classification tasks, imbalanced category distribution is a key factor that affects model performance. An excessive proportion of majority class samples often causes models to favor the majority class during training, thereby reducing their ability to recognize minority class samples. For example, in a three-class dataset, positive reviews in simplifyweibo_4_moods account for approximately 57%, neutral reviews account for only 14%, and negative reviews account for approximately 29%. The online_shopping_10_cats dataset features negative reviews at only 12.6% and positive reviews at approximately 57.7%. Moreover, datasets such as dmsc_v2 and cn-taobao exhibit significantly fewer minority class samples. In binary classification tasks, datasets such as waimai_10k, ChnSentiCorp_htl_all, and our in-house dataset also exhibit pronounced imbalances in positive-to-negative sample ratios. This category distribution imbalance poses challenges for model training and generalization capabilities, particularly in terms of reducing minority class sample identification accuracy.

The proposed model employs a multi-module collaborative design to enhance the semantic modeling capabilities of minority samples while maintaining overall performance stability. Experimental results demonstrate that our model achieves F1 scores of 0.9345, 0.8568, 0.9124, 0.8503, and 0.9240 for the simpl_weibo, online_10, dmsc_v2, cn-taobao, and self-built datasets, respectively, in the three-classification task, outperforming all comparison models (e.g., TCN-BiLSTM, IDM-GAT, RoCSAOL, Transformer-TextCNN, and RoBERTa-BiGRU-Double Att). For binary classification tasks, on the weibo_senti_100k, waimai_10k, ChnSentiCorp_htl_all, and self-built datasets, the F1 Score reached 0.9267, 0.8695, 0.7546, and 0.9087, respectively, significantly outperforming traditional deep models and BERT-based models. This demonstrates that multi-granularity feature modeling effectively mitigates bias caused by class imbalance. BiGRU combined with MHSA enhances contextual modeling capabilities, improving semantic understanding of low-frequency class samples.

Ablation experiments revealed each module's role in class imbalance mitigation. Removing the data augmentation module caused the F1 score to drop from 0.9240 to 0.8950, with a significant decline in recall, indicating that data augmentation is crucial for expanding minority class corpora and enhancing recognition capabilities for rare samples. Removing either the BiGRU or MHSA module reduced the F1 score to 0.9010 and 0.9079, respectively, demonstrating that sequence modeling and global context capture are effective means of mitigating class imbalance.

In the three-classification task, analyzing the confusion matrix with positive, negative, and neutral classes arranged in the x- and y-coordinate order allows for an intuitive observation of each module's contribution to minority class recognition (as shown in Figure 6). After removing the BERT augmentation, the confusion matrix reveals that the diagonal values for the neutral class (minority class) are 1789, while those for the positive and negative classes are 7236 and 1171, respectively. Notably, a high number of positive instances (230) are misclassified as neutral, indicating limited minority class detection capability. After introducing BERT augmentation, the positive, negative, and neutral diagonal entries increased to 7460, 1174, and 1868, respectively, with a significant reduction in misclassifications on off-diagonal entries. This demonstrates that BERT augmentation effectively enhances minority class recognition by capturing global semantics and key trigger words, making it a crucial module for mitigating the imbalanced data problem. Removing BiGRU caused neutral diagonal misclassifications to drop to 1763, positive to 7288, and negative to 1179, with overall misclassifications increasing. This demonstrates that sequence modeling is equally crucial for capturing the features of minority sentiment. The neutral class diagonal was 1680, positive 7443, and negative 1116 after removing the

MHSA, with a significant increase in neutral misclassifications. This demonstrates that global attention plays a prominent role in reinforcing the contextual information of minority classes. Comprehensive comparisons reveal that BERT data augmentation, followed by MHSA and BiGRU, is the most effective method for mitigating category imbalance.

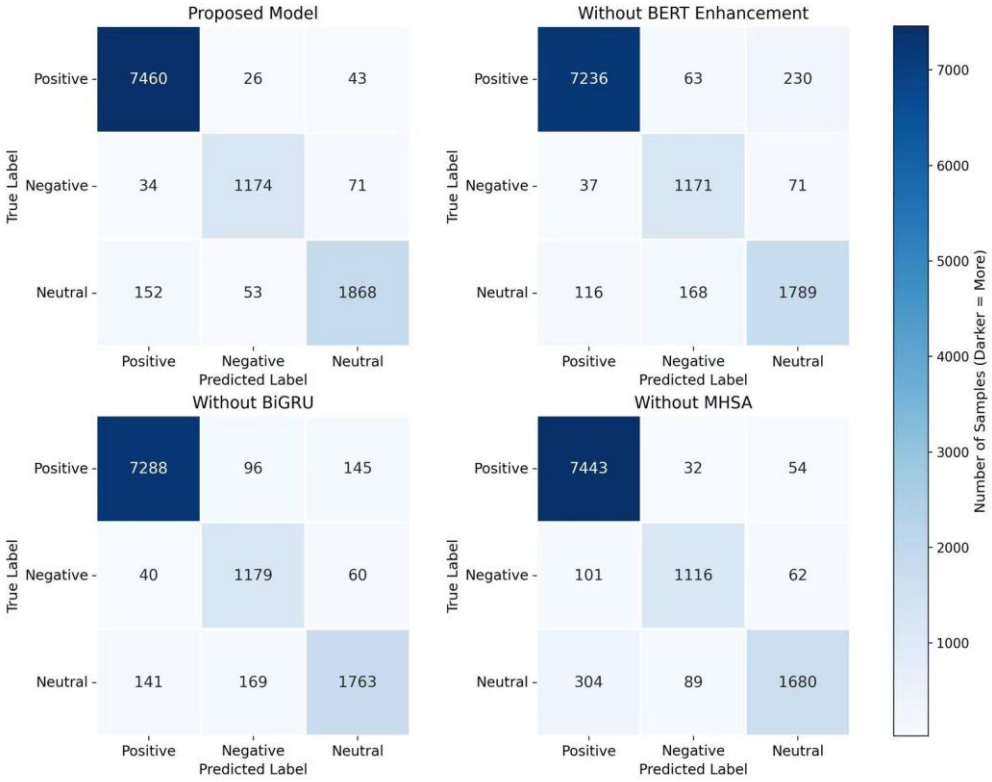


Figure 6. Three-classification confusion matrix comparison chart.

In binary classification tasks, the confusion matrix further demonstrates the model's advantage in identifying minority classes (as shown in Figure 7). Within this model, the positive and negative classes are relatively balanced. Only 152 instances of the minority-negative class were misclassified as positive, accounting for approximately 11.8% of the total samples. This indicates that the model effectively mitigates class imbalance. In the Without BERT Enhancement model, the number of positive instances misclassified as negative increases to 95, whereas the number of negative instances misclassified as positive slightly decreases to 108, resulting in a marginal drop in the overall F1 score. This demonstrates that BERT data augmentation effectively enhances minority class coverage through semantic enrichment. The number of negative class misclassifications increases to 160 without BiGRU, highlighting the importance of sequence modeling in capturing the contextual dependencies of minority classes. Without MHSA, misclassifications of negative classes rise to 180, whereas positive class misclassifications slightly increase to 7450, indicating that global attention has the greatest impact on capturing key features of minority classes. Thus, BERT data augmentation most directly improves minority class coverage. BiGRU enhances local sequence modeling to reduce minority class misclassifications, while MHSA strengthens global context modeling, making it a key module for addressing class imbalance.

In summary, this model significantly improves minority class recognition in both three-classification and binary classification tasks through multi-module collaborative design while maintaining stable majority class accuracy. It naturally mitigates the impact of class imbalance. In our experiments, the self-built dataset performs better on the three-class task (Accuracy = 0.9317, Precision = 0.9366, Recall = 0.9240, AUC = 0.9643) than the binary

task (Accuracy = 0.9158, Precision = 0.9104, Recall = 0.9087, AUC = 0.9698), mainly due to the more balanced class distribution and explicit modeling of the neutral category, which enhances sentiment boundary clarity and semantic richness.

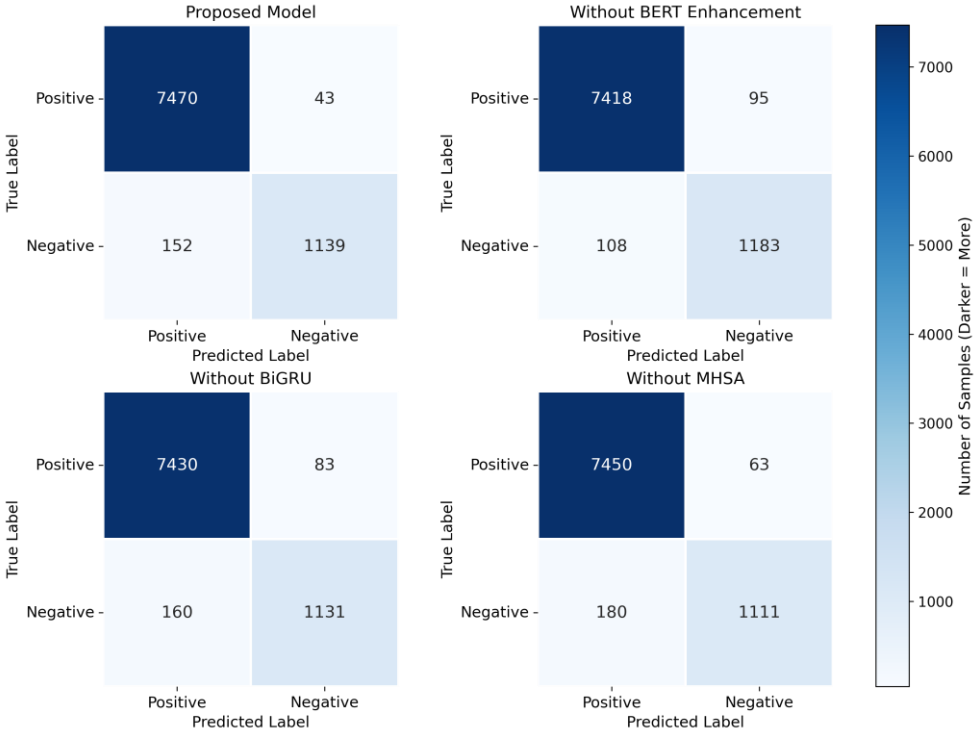


Figure 7. Comparison of Binary Classification Confusion Matrices.

5. Conclusions

5.1. Research Summary

Currently, text sentiment analysis research faces three core challenges: First, the lack of a dynamic feature extraction mechanism results in insufficient semantic characterization. Adapting traditional methods to dynamic changes in text granularity is difficult, and key features can easily be lost or incorrectly extracted. Second, the model architecture is inadequate in terms of text-type adaptability. Existing schemes lack unified processing capability for cross-domain text, which limits the generalization of the analysis and the scene migration capability to some extent. Third, a more significant bottleneck in domain adaptive technology. Single-scenario-optimized models exhibit more obvious performance degradation when applied across tasks.

Therefore, this study proposes an innovative model for sentiment analysis. Specifically, first, a hybrid BERT-Chinese-RoBERTa pre-training strategy is adopted to achieve joint data and feature optimization using a generative enhancement method for masked language models to effectively capture multi-level semantic associations. Second, the gated dual-stream transformer module is designed to dynamically adjust the weights of character-level and word-level features using the channel selective attention mechanism to enhance the model's adaptability to the representation of heterogeneous texts. Finally, the BiGRU-Multiple Attention Joint Learning Network is constructed to improve the recognition accuracy of fine-grained sentiment elements through temporal modeling and global dependency to capture co-optimization.

This model achieves an F1 score of 92.4% on mainstream sentiment datasets, demonstrating an average improvement of 8.7% over baseline models and outstanding performance in small-to-medium-scale tasks. Through multi-module designs, including BERT

data augmentation, fusion of character-level and word-level features, and collaborative modeling of BiGRU and MHSA, the model effectively mitigates the challenge of identifying minority class samples caused by category distribution imbalance. This approach enhances the model's robustness and generalization capabilities by significantly improving minority class recognition while maintaining majority class accuracy.

5.2. Limitations and Future Prospects

Although the proposed sentiment analysis model achieves significant performance improvements on multiple benchmark datasets, it still has certain limitations. First, the model's computational complexity is high when processing large-scale data, and its inference efficiency cannot fully meet the requirements of real-time applications. Second, the model heavily relies on training data when transferring across domains. The performance may fluctuate if the target domain has different language styles or sentimental expressions. Finally, this research focuses primarily on text modalities and has not yet incorporated multimodal sentiment information, such as images and speech, into the modeling framework, resulting in limited application scenarios.

Future research can address several optimization areas to address these limitations. First, optimizing computational efficiency and model lightweighting can enhance large-scale data processing and real-time application capabilities. Second, exploring cross-domain adaptation techniques and transfer learning methods can enhance the generalizability of the model in heterogeneous text environments. Finally, the integration of multimodal information into the model framework can expand the number of application scenarios.

Against the backdrop of the rapid development of large language models (LLMs), the proposed gated dual-stream architecture and generative augmentation strategy demonstrates promising performance and robustness in small-scale and medium-scale sentiment analysis tasks, demonstrating its potential to complement LLMs. On one hand, this model can be embedded within the LLM framework as a lightweight, specialized module, enhancing fine-grained sentiment recognition and handling of imbalanced data. On the other hand, when combined with the powerful semantic understanding and cross-task transfer capabilities of LLMs, the model offers the potential for further expansion in multimodal sentiment analysis and cross-domain adaptation. In summary, the research findings of this paper not only possess independent application value in current small- and medium-scale tasks but also provide a viable approach for future exploration of its integration with LLMs. It is expected to play a significant role in sentiment analysis's intelligentization and multi-scenario expansion.

Author Contributions: Conceptualization, S.Y. and J.X.; methodology, S.Y. and J.X.; software, J.X.; validation, Z.L. and Y.S.; formal analysis, J.X.; investigation, S.Y.; resources, Z.L.; data curation, J.X.; writing—original draft preparation, S.Y. and J.X.; writing—review and editing, S.Y., Z.L. and Y.S.; visualization, J.X.; supervision, S.Y.; project administration, S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by 2017 Higher Education Research Project grant number 2017G04. The APC was funded by Dalian University of Foreign Languages.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: During the preparation of this manuscript, all authors used DeepSeek R1 for the purposes of polishing. All authors have reviewed and edited the output and take full responsibility for the content of this publication. We would like to thank all the anonymous reviewers for their constructive comments and suggestions that greatly improved the quality of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. An, J.; Zainon, W.M.N.W. Integrating Color Cues to Improve Multimodal Sentiment Analysis in Social Media. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106874. [CrossRef]
2. Jiang, H.; Zhao, C.; Chen, H.; Wang, C. Construction method of domain sentiment lexicon based on improved TF-IDF and BERT. *Comput. Sci.* **2024**, *51*, 162–170. [CrossRef]
3. Wang, H.; Wang, Y.; Marius, G.P. Research on text sentiment analysis based on sentiment dictionary and deep learning. *Comput. Digit. Eng.* **2024**, *52*, 451–455. [CrossRef]
4. Li, Z.; Li, R.; Jin, G.H. Sentiment Analysis of Danmaku Videos Based on Naïve Bayes and Sentiment Dictionary. *IEEE Access* **2020**, *8*, 75073–75084. [CrossRef]
5. Tarımer, İ.; Çoban, A.; Kocaman, A.E. Sentiment Analysis on IMDB Movie Comments and Twitter Data by Machine Learning and Vector Space Techniques. *arXiv* **2019**, arXiv:1903.11983. [CrossRef]
6. Dervenis, C.; Kanakis, G.; Fitsilis, P. Sentiment Analysis of Student Feedback: A Comparative Study Employing Lexicon and Machine Learning Techniques. *Stud. Educ. Eval.* **2024**, *83*, 101406. [CrossRef]
7. Long, Y.; Li, Q. Sentiment analysis model of Chinese commentary text based on self-attention and TextCNN-BiLSTM. *J. Shihezi Univ. Nat. Sci.* **2025**, *43*, 111–121. [CrossRef]
8. Huang, J.; Xu, X. A short text sentiment analysis model integrating output features of Bi-LSTM and CNN. *Comput. Digit. Eng.* **2024**, *52*, 3326–3330. [CrossRef]
9. Sun, S.; Wei, G.; Wu, S. Sentiment analysis method integrating hypergraph enhance and dual contrastive learning. *Comput. Eng. Appl.* **2025**, 1–12. [CrossRef]
10. Zhang, B.; Zhou, W. Transformer-Encoder-GRU (T-E-GRU) for Chinese Sentiment Analysis on Chinese Comment Text. *Neural Process Lett.* **2023**, *55*, 1847–1867. [CrossRef]
11. Hu, J.; Yu, Q. Sentiment analysis of Chinese text based on BERT-BGRU-Att model. *J. Tianjin Univ. Technol.* **2024**, *40*, 85–90. [CrossRef]
12. Tembhurne, J.V.; Lakhotia, K.; Agrawal, A. Twitter Sentiment Analysis using Ensemble of Multi-Channel Model Based on Machine Learning and Deep Learning Techniques. *Knowl. Inf. Syst.* **2025**, *67*, 1045–1071. [CrossRef]
13. Li, H.; Ma, Y.; Ma, Z.; Zhu, H. Weibo Text Sentiment Analysis Based on Bert and Deep Learning. *Appl. Sci.* **2021**, *11*, 10774. [CrossRef]
14. Sun, J.; Wei, C. Research on gold futures price forecasting based on text sentiment analysis and LightGBM-LSTM model. *J. Nanjing Univ. Inf. Sci. Technol.* **2024**, 1–16. [CrossRef]
15. Xu, X.; Tian, K. A novel financial text sentiment analysis-based approach for stock index prediction. *J. Quant. Technol.* **2021**, *38*, 124–145. [CrossRef]
16. Duan, W.; Xue, T. FinBERT-RCNN-ATTACK: Emotional analysis model of financial text. *Comput. Tchnology Dev.* **2024**, *34*, 157–162. [CrossRef]
17. Huang, S.; Raga, R.C. A deep learning model for student sentiment analysis on course reviews. *IEEE Access* **2024**, *12*, 136747–136758. [CrossRef]
18. Liang, Z.; Chen, J. An Al-BERT-Bi-GRU-LDA algorithm for negative sentiment analysis on Bilibili comments. *PeerJ Comput. Sci.* **2024**, *10*, e2029. [CrossRef]
19. Liyih, A.; Anagaw, S.; Yibeyin, M. Sentiment analysis of the Hamas-Israel war on YouTube comments using deep learning. *Sci. Rep.* **2024**, *14*, 13647. [CrossRef] [PubMed]
20. Guo, Y.; Yao, X. Research on user reviews of online health community based on topic mining and sentiment analysis. *J. Mod. Inf.* **2025**, *45*, 135–145.
21. Wen, T.; Xu, G. Research on sentiment-topic recognition on Service quality of mixed anxiety and depression disorder. *Inf. Res.* **2024**, 10–19. [CrossRef]
22. Zhang, W.; Zhang, J.; Guo, Y. Sentiment analysis of online medical reviews based on BERT and semantics collaboration through dual-channel. *J. Med. Inform.* **2024**, *45*, 30–35. [CrossRef]
23. Mishev, K.; Gjorgjevikj, A.; Vodenska, I.; Chitkushev, L.T.; Trajanov, D. Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access* **2020**, *8*, 131662–131682. [CrossRef]
24. Yan, C.; Liu, J.; Liu, W.; Liu, X. Sentiment Analysis and Topic Mining Using a Novel Deep Attention-Based Parallel Dual-Channel Model for Online Course Reviews. *Cogn. Comput.* **2023**, *15*, 304–322. [CrossRef]
25. Wang, X. Predicting Dropout and Sentiment Analysis of Course Evaluation in MOOCs Based on Deep Learning. Master's Thesis, Henan University, Zhengzhou, China, 2024.
26. Gao, H. Sentiment analysis model of online course review based on MacBERT. *Electron. Des. Eng.* **2023**, *31*, 39–43. [CrossRef]
27. Rizinski, M.; Peshov, H.; Mishev, K.; Jovanovik, M.; Trajanov, D. Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex). *IEEE Access* **2024**, *12*, 7170–7198. [CrossRef]
28. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *ICLR* **2019**. [CrossRef]

29. Wang, T.; Yang, W. Review of Text Sentiment Analysis Methods. *Comput. Eng. Appl.* **2021**, *57*, 11–24. [CrossRef]
30. Liang, Y.M.; Fan, J.; Chai, W.Z. Multi-scale feature fusion sentiment classification based on bidirectional cross attention. *J. Comput. Appl.* **2025**, *45*, 2773–2782.
31. Xu, G.X.; Liu, L.Y.; Wang, J.C.; Chen, Z. Text sentiment analysis based on Bert and Hypergraph with dual attention network. *Appl. Res. Comput.* **2024**, *41*, 786–793.
32. Tianchi. 2020. Available online: https://tianchi.aliyun.com/dataset/131940 (accessed on 7 June 2022).
33. Ren, C.; Qiu, Q.; Lie, S. Text sentiment analysis fusing TCN and BiLSTM. *Comput. Eng. Des.* **2024**, *45*, 2090–2096. [CrossRef]
34. Ai, S.; Chen, H.; Cui, X.; An, R. Aspect-based sentiment analysis model for short text with interactive dependent multi-head graph attention. *J. Harbin Univ. Sci. Technol.* **2024**, *29*, 10–17. [CrossRef]
35. Ye, H. Research on Sentiment Analysis of Short Text User Comments Based on Deep Learning. Master's Thesis, Nanchang University, Nanchang, China, 2024.
36. Li, H.; Wang, Y.; Wang, X. Chinese short text sentiment classification: Research on Transformer-TextCNN model with location-aware enhancement. *Comput. Eng. Appl.* **2025**, *61*, 216–226. [CrossRef]
37. Wu, J.; Jia, Y.; Wang, Z.; Zhang, J.; Yan, W.; Gao, A.; Che, S. Sentiment analysis of Weibo epidemic public opinion text based on deep learning. *Comput. Technol. Dev.* **2024**, *34*, 175–183. [CrossRef]
38. Chen, T.; An, J. Sentiment classification of microblog short text based on feature fusion. *Front. Data Comput.* **2020**, *2*, 21–29.
39. Liu, F.; Yu, W. Text sentiment analysis model based on BERT and Attention. *Softw. Guide* **2023**, *22*, 66–71. [CrossRef]
40. Hsieh, Y.H.; Zeng, X. Sentiment Analysis: An ERNIE-BiLSTM Approach to Bullet Screen Comments. *Sensors* **2022**, *22*, 5223. [CrossRef] [PubMed]
41. Wang, M.; You, H.; Ma, H.; Sun, X.; Wang, Z. Sentiment Analysis of Online New Energy Vehicle Reviews. *Appl. Sci.* **2023**, *13*, 8176. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.