DyFuLM: A Dynamic Collaborative Dual-Encoder Network for Fine-Grained Sentiment Analysis

Jiachen Yuan[1] [0009-0000-8763-0440], Ruohan Zhou[1,†][0009-0006-7189-8040], Wenzheng Huang[2,†][0009-0004-0969-7331], Churui Yang[3,‡][0009-0009-2770-8758], Guoyan Zhang[1,‡][0009-0009-1793-0443], Shiyao Wei[1,‡][0009-0006-1981-775X], Jiazhen Hu[2][0009-0004-4309-6644], Ning Xin[1][0009-0003-7718-6552]*, Md Maruf Hasan[2][0000-0001-9809-9376]**

[1]Department of Applied Mathematics, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China
[2]School of AI and Advanced Computing, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China
[3]Department of Intelligent Science, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China
† Ruohan Zhou and Wenzheng Huang contributed equally as co-second authors.
‡ Churui Yang, Guoyan Zhang, and Shiyao Wei contributed equally as co-third authors.
Jiachen.Yuan23@student.xjtlu.edu.cn

Abstract. Fine-grained sentiment analysis in review text requires simultaneous modeling of global semantics, subtle affective cues, and cross-task dependencies between sentiment categories and sentiment intensity. Existing single pre-trained language models (PLMs) and simple late-fusion strategies often collapse layer-wise, encoder-wise, and task-wise heterogeneity into one shared representation, which limits multi-dimensional sentiment modeling. To address this issue, this paper proposes DyFuLM, a collaborative dual-encoder network for fine-grained sentiment analysis. DyFuLM combines RoBERTa and DeBERTa, uses Hierarchical Dynamic Fusion to aggregate informative intermediate-layer signals within each encoder, and applies Gated Feature Aggregation to coordinate cross-encoder representations. Hierarchical Guidance and Dynamic Loss Weighting further support coarse-grained classification, fine-grained classification, and sentiment intensity regression in one unified framework. Experiments on a hotel review benchmark with 515,738 records show that DyFuLM achieves 82.64% coarse-grained accuracy, 68.48% fine-grained accuracy, MAE = 0.0674, MSE = 0.0082, and R^2 = 0.6903, outperforming strong single-model and hybrid baselines. Ablation results show that the gains arise from coordinated modeling at the layer, encoder, and task levels. The framework prioritizes predictive quality in multi-dimensional sentiment understanding, while lightweight deployment remains outside the current scope.

* Corresponding author.
** Corresponding author.

Keywords: Pre-trained Language Models, Dual-Encoder Fusion, Fine-Grained Sentiment Analysis, Multi-Task Learning, Emotion Intensity Regression

1 Introduction

With the rapid growth of the digital economy, hotel reviews on online platforms have become an important source of information shaping tourists' decisions [1]. Compared with traditional questionnaires and manually summarized evaluations, review texts provide richer evidence about user perceptions of service quality, facilities, environment, and overall experience. Analyzing such reviews can support service optimization, demand forecasting, and user-oriented recommendation in the tourism industry [2].

Conventional sentiment analysis methods usually classify text as positive, negative, or neutral by relying on sentiment lexicons, handcrafted features, or traditional machine-learning algorithms [3]. Although such methods are computationally efficient, they often fail to capture contextual dependence and subtle emotional variation. This limitation is especially evident in hotel reviews, where users may simultaneously praise one aspect of a hotel while criticizing another [4]. Such mixed and nuanced expressions make coarse polarity classification insufficient for practical decision support.

Pre-trained language models such as BERT and its variants have substantially advanced contextual representation learning for sentiment analysis [5, 10, 11]. Nevertheless, two challenges remain in fine-grained review understanding. First, single-encoder architectures may not generalize consistently across heterogeneous review styles because different PLMs encode different semantic biases and contextual patterns [6]. Second, many existing studies still focus primarily on coarse polarity, while real review mining often requires finer emotion categorization and the estimation of sentiment intensity [7].

A further difficulty lies in the interaction among three forms of heterogeneity. Different Transformer layers capture lexical, syntactic, and semantic cues at different abstraction levels; different encoders emphasize different contextual biases; and the downstream tasks exhibit different difficulty levels and convergence behaviors. When these factors are collapsed into a single last-layer representation, fine-grained affective cues are easily diluted and multi-task optimization becomes unstable. A more suitable strategy is to coordinate layer selection, encoder collaboration, and task interaction within one unified pipeline.

DyFuLM addresses this problem through an ordered collaborative architecture. Hierarchical Dynamic Fusion first aggregates informative intermediate-layer signals within each encoder, Gated Feature Aggregation then coordinates the complementary representations produced by RoBERTa and DeBERTa, and Hierarchical Guidance together with Dynamic Loss Weighting regulates information flow and optimization across the three output tasks. This design targets predictive quality for multi-dimensional sentiment analysis on complex review data rather than lightweight deployment.

The framework is evaluated on a large-scale hotel review dataset containing 515,738 records from 1,493 European hotels collected between 2015 and 2017. The present experiments focus on within-domain validation for hotel review analysis. Experimental

results show that DyFuLM achieves consistently strong performance across coarse-grained classification, fine-grained classification, and sentiment intensity regression.

The main contributions of this work are summarized as follows:

1. A collaborative dual-encoder framework is proposed to model coarse-grained classification, fine-grained classification, and sentiment intensity regression within a unified architecture.
2. A Hierarchical Dynamic Fusion module and a Gated Feature Aggregation module are designed to coordinate layer-wise and encoder-wise heterogeneity, enabling richer semantic interaction between RoBERTa and DeBERTa.
3. A multi-task training strategy combining Hierarchical Guidance and Dynamic Loss Weighting is introduced to improve task interaction and optimization balance in multi-dimensional sentiment analysis.

2 Related Work

Early sentiment analysis mainly relied on lexicon-based methods and traditional machine-learning classifiers, which were limited in contextual modeling and semantic compositionality [3]. Neural architectures such as CNNs and LSTMs later improved local feature extraction and sequence modeling, but long-range dependence and transferable semantics still remained difficult to capture consistently [8, 9].

Transformer-based PLMs such as BERT, RoBERTa, and DeBERTa have since become dominant because large-scale pretraining and self-attention substantially improve contextual representation learning [5, 10, 11]. However, fine-grained review understanding still poses two challenges for single-encoder PLMs: the last hidden layer alone may not fully preserve informative intermediate-layer semantics, and one encoder often provides limited representational diversity for texts containing mixed or weakly expressed emotions [6, 7].

Recent multi-task affective models have shown that related supervision signals can improve emotion and sentiment prediction. All-in-One jointly models emotion, sentiment, and intensity prediction, while Multi-Task Learning Transformers compares transformer backbones for emotion classification and intensity prediction in social media [16, 17]. Multi-encoder and hybrid fusion paradigms provide another line of development. Poly-Encoder explores dual-tower structures for richer interaction, and gated fusion is widely used to combine heterogeneous representations [12, 15]. DistilBERT and EmoBERTa further illustrate complementary directions toward compact and emotion-aware PLM design [13, 14].

Fine-grained sentiment extraction has also begun to explore dual encoders and contrastive learning. External-knowledge enhanced dual encoder and contrastive learning for aspect sentiment triplet extraction and Sentiment Triplet Extraction With Multi-View Contrastive Learning show that structured fine-grained sentiment tasks benefit from richer encoder interaction and stronger supervision [18]. Beyond classification, recent affective language systems have also incorporated higher-level reasoning in domain-specific settings such as mental health counseling. Compared with prior work,

DyFuLM focuses on review-level multi-dimensional sentiment analysis and organizes layer-wise fusion, encoder-wise collaboration, and task-wise guidance into one collaborative sentiment pipeline.

3 Methodology

3.1 Dataset Description and Preprocessing

This study uses the Hotel Reviews Dataset collected from Booking.com, which contains 515,738 user review records from 1,493 European hotels between 2015 and 2017. Each review includes textual comments and structured attributes such as rating score, review date, reviewer nationality, and hotel address. The preprocessing pipeline contains three core steps: text cleaning removes missing values, duplicated records, and obvious formatting artifacts; sentence consolidation merges the positive and negative comment fields into a unified textual sequence when necessary; and label construction maps numerical reviewer scores to both coarse-grained and fine-grained sentiment labels so that the three downstream tasks can be trained under a unified data pipeline. Other normalization and consistency checks are applied as standard preparation steps to ensure stable optimization and reliable analysis.

All dataset statistics reported in the experiments refer to the 515,738 raw review records. Coarse-grained and fine-grained label frequencies are derived from task-specific mappings of the same review corpus and should therefore be interpreted within their own label spaces rather than compared directly with the raw record count.

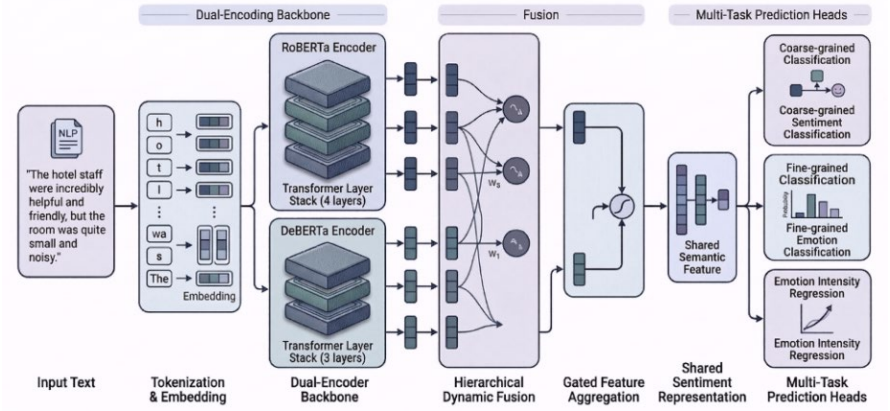


Fig. 1. Overall architecture of DyFuLM. The framework consists of input encoding, dual-encoder representation learning, hierarchical fusion, gated aggregation, and multi-task prediction. Hierarchical Guidance acts on prediction refinement, whereas Dynamic Loss Weighting is applied in the training objective and is not shown in the inference graph.

3.2 Workflow and Model Architecture

As shown in Fig. 1, DyFuLM follows a four-stage pipeline. In the input stage, the review text is tokenized and converted into embeddings. In the encoding stage, the same input is processed in parallel by a dual-encoder backbone composed of RoBERTa and DeBERTa. In the fusion stage, Hierarchical Dynamic Fusion extracts informative intra-encoder layer semantics and Gated Feature Aggregation combines the two encoder branches into a shared sentiment representation. In the prediction stage, this representation is used for coarse-grained classification, fine-grained classification, and sentiment intensity regression.

DyFuLM addresses three coupled sources of heterogeneity in fine-grained review modeling: layer-wise heterogeneity inside each PLM, representational heterogeneity across encoders, and optimization heterogeneity across tasks. Traditional multi-task sentiment analysis models often depend on a single language encoder. Although such architectures can provide strong global semantic understanding, they may be less effective at capturing subtle emotional cues or heterogeneous semantic patterns in complex reviews. RoBERTa provides strong contextual semantic modeling, while DeBERTa contributes finer contextual discrimination through disentangled attention. Their combination yields a richer feature basis for downstream sentiment tasks.

3.2.1 Hierarchical Dynamic Fusion

Hierarchical Dynamic Fusion aggregates informative intermediate-layer representations within each encoder branch rather than relying only on the final hidden layer. The following equations describe the operation on one encoder branch; the same procedure is applied independently to the RoBERTa branch and the DeBERTa branch before cross-encoder aggregation.

Let the hidden state of the l-th layer from one encoder be denoted by

$$H^{(l)} = [h_1^{(l)}, \dots, h_T^{(l)}] \in \mathbb{R}^{T\times d}, \quad l = 1, \dots, L,$$

where T is the sequence length, d is the hidden dimension, and L is the number of Transformer layers. For each token position t, the layer-wise representations

$$x_t = [h_t^{(1)}; h_t^{(2)}; \cdots; h_t^{(L)}] \in \mathbb{R}^{L\times d}$$

are fed into a bidirectional LSTM to model inter-layer dependence:

$$U_t = \mathrm{BiLSTM}(x_t), \qquad U_t \in \mathbb{R}^{L\times d_u}, \tag{1}$$

where d_u is the hidden size of the BiLSTM output.

A learnable scoring vector $w \in \mathbb{R}^{d_u}$ is then used to compute the importance of each layer:

$$e_t^{(l)} = w^{\top} u_t^{(l)}, \tag{2}$$

where $u_t^{(l)}$ is the l-th row of U_t. The normalized attention weight is defined as

$$\alpha_t^{(l)} = \frac{\exp\left(e_t^{(l)}\right)}{\sum_{k=1}^{L} \exp\left(e_t^{(k)}\right)}, \tag{3}$$

with $\alpha_t^{(l)} \in \mathbb{R}$ and $\sum_{l=1}^{L} \alpha_t^{(l)} = 1$. The fused token representation for that encoder branch is obtained by

$$h_t^{\text{layer}} = \sum_{l=1}^{L} \alpha_t^{(l)} h_t^{(l)}, \tag{4}$$

where $h_t^{\text{layer}} \in \mathbb{R}^d$. The same HDF procedure is applied independently to the RoBERTa and DeBERTa branches before cross-encoder aggregation.

3.2.2 Gated Feature Aggregation

After intra-encoder fusion, DyFuLM performs cross-encoder aggregation to coordinate the representations produced by RoBERTa and DeBERTa. Let $h_t^A \in \mathbb{R}^d$ denote the HDF-enhanced token representation from the RoBERTa branch and $h_t^B \in \mathbb{R}^d$ denote the corresponding HDF-enhanced token representation from the DeBERTa branch. The gate vector is computed as

$$z_t = \sigma\left(W_g[h_t^A; h_t^B] + b_g\right), \tag{5}$$

where $W_g \in \mathbb{R}^{d \times 2d}$, $b_g \in \mathbb{R}^d$, and $z_t \in \mathbb{R}^d$. The final token-level fused representation is

$$h_t = z_t \odot h_t^A + (1 - z_t) \odot h_t^B, \tag{6}$$

where $\odot$ denotes element-wise multiplication. To obtain a sentence-level feature for downstream tasks, mean pooling is applied over the fused token representations:

$$s = \frac{1}{T} \sum_{t=1}^{T} h_t, \qquad s \in \mathbb{R}^d. \tag{7}$$

The resulting vector s is regarded as the shared sentiment representation.

3.2.3 Hierarchical Guidance

Based on the shared sentiment representation s, the model performs three related tasks. The coarse-grained classifier predicts the high-level sentiment category:

$$\hat{y}_c = f_{\text{coarse}}(s), \tag{8}$$

where $\hat{y}_c$ denotes the probability distribution over coarse-grained sentiment classes. The intensity regressor predicts sentiment strength:

$$\hat{y}_i = f_{\text{intensity}}(s). \tag{9}$$

To avoid symbol conflict with the gate vector, the task-guidance vector is denoted by r. It is generated from the coarse-grained and intensity predictions through

$$r = \phi([\hat{y}_c; \hat{y}_i]), \tag{10}$$

where $r \in \mathbb{R}^d$ serves as a feature-wise recalibration signal. The guided representation is then defined as

$$s' = s \odot r, \tag{11}$$

and the fine-grained classifier predicts the detailed sentiment category by

$$\hat{y}_f = f_{\text{fine}}(s'). \tag{12}$$

This hierarchical guidance mechanism allows high-level sentiment polarity and intensity information to refine fine-grained category recognition.

3.3 Task Objectives and Computational Scope

Let $\mathcal{L}_c$, $\mathcal{L}_f$, and $\mathcal{L}_i$ denote the losses for coarse-grained classification, fine-grained classification, and intensity regression, respectively. The two classification objectives use cross-entropy loss, whereas the regression objective uses mean squared error:

$$\mathcal{L}_c = -\sum_{k=1}^{K_c} y_{c,k} \log \hat{y}_{c,k}, \tag{13}$$

$$\mathcal{L}_f = -\sum_{k=1}^{K_f} y_{f,k} \log \hat{y}_{f,k}, \tag{14}$$

$$\mathcal{L}_i = \| y_i - \hat{y}_i \|_2^2. \tag{15}$$

The overall training objective is

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_f \mathcal{L}_f + \lambda_i \mathcal{L}_i, \tag{16}$$

where $\lambda_c + \lambda_f + \lambda_i = 1$. During training, the loss weights are updated adaptively according to task loss magnitude, loss variation trend, and relative convergence behavior. Tasks with larger losses, slower loss reduction, or weaker convergence receive larger weights so that under-optimized objectives can obtain more training emphasis. In the ablation setting without DL, the task weights remain fixed at lambda_c = lambda_f = lambda_i = 1/3, so the comparison isolates adaptive weighting against equal static weighting.

DyFuLM employs two full-size encoders together with additional fusion and guidance modules. This design increases training and inference cost relative to single-encoder baselines, but it preserves richer intermediate and cross-encoder information for the three downstream tasks. Accordingly, the present study concentrates on predictive performance and task consistency. Parameter compression, latency-aware redesign, and lightweight deployment are left for future work.

4 Experimental Results and Analysis

4.1 Experimental Setup

Unless otherwise stated, all models are trained and evaluated using the same data split and preprocessing pipeline. The dataset is divided into training, validation, and test sets using a ratio of 8:1:1. Input reviews are truncated or padded to a maximum sequence length of 256. The model is optimized using AdamW with a learning rate of 1e-5, batch size of 16, and 6 training epochs. Model selection is based on the best validation Macro-F1. All experiments are conducted on a platform equipped with an Intel Core i9-13900K CPU, an NVIDIA RTX 4090D GPU, and 64 GB RAM.

4.2 Dataset Statistics

This study uses a dataset of 515,738 hotel review records collected from Booking.com, covering 1,493 hotels in Europe from 2015 to 2017. Each record contains 17 attributes, including review text, rating, timestamp, reviewer nationality, and hotel location. The original review texts and rating scores were used as the primary inputs for subsequent analysis after preprocessing and label mapping.

Table 1 shows that hotel ratings remained consistently high between 2015 and 2017. The median score stayed at 8.8 and the 25th percentile remained at 7.5 across all three years, indicating generally positive user perception. The average rating fluctuated only slightly, from 8.3198 in 2015 to 8.4247 in 2016 and 8.3905 in 2017.

Table 1. Statistical characteristics of hotel ratings.

Indicator	Maximum Value	25% Quartile	Mean	Median
2015 Score	10.0000	7.5000	8.3198	8.8000
2016 Score	10.0000	7.5000	8.4247	8.8000
2017 Score	10.0000	7.5000	8.3905	8.8000
Overall Score	10.0000	7.5000	8.3951	8.8000

The maximum score remained 10 throughout the observation period, suggesting a stable rating scale.

4.3 Performance Metrics Overview

Table 2 compares DyFuLM with representative baselines across classification and regression tasks. Coarse Accuracy and Fine Accuracy evaluate classification performance at two levels of granularity. Coarse F1 and Fine F1 denote Macro-F1, which is used throughout this paper unless otherwise stated. For regression, Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2) are reported. Lower MAE and MSE indicate better regression performance, whereas higher values are preferred for the remaining metrics. Macro-F1 is particularly important in

this setting because it complements accuracy by reducing the dominance of majority positive labels in the mapped sentiment spaces.

4.4 Performance Comparison and Improvement Analysis

To benchmark the proposed method, the comparison includes representative PLM baselines and hybrid architectures. BERT serves as a standard baseline in natural language processing [5]. RoBERTa and DeBERTa are included for their strong contextual modeling capability [10, 11]. DistilBERT provides a compact comparison point [13], and EmoBERTa offers a task-relevant baseline for emotion-oriented representation learning [14]. The hybrid baselines combine PLMs with direct concatenation, gated fusion, or lightweight sequence learners, which together provide a comprehensive benchmark for evaluating the effectiveness of DyFuLM [15].

As shown in Table 2, DyFuLM achieves the best results across all reported metrics. Hierarchical Dynamic Fusion strengthens token-level semantics by using informative intermediate-layer signals rather than last-layer pooling alone. Gated Feature Aggregation then coordinates the complementary information encoded by RoBERTa and DeBERTa, reducing the representational bias of a single PLM. Finally, Hierarchical Guidance and Dynamic Loss Weighting promote more coherent interaction between classification and regression objectives.

These effects appear in both discrete and continuous prediction tasks. DyFuLM improves coarse-grained and fine-grained classification, indicating stronger semantic discrimination and better balance between majority and minority labels under Macro-F1. It also attains the lowest MAE and MSE together with the highest R^2, suggesting that the shared sentiment representation is more effective for continuous sentiment intensity estimation. Relative to the strongest baseline, the regression gains reach 3.30% on MAE, 10.87% on MSE, and 5.13% on R^2. Taken together, these results show that the

Table 2. Performance comparison with baselines. Best and second-best results for each metric are highlighted in red and blue, respectively. Lower values are better for MAE and MSE.

Model	C. Acc	F. Acc	C. F1	F. F1	MAE	MSE	R^2
DyFuLM	0.8264	0.6848	0.8215	0.6814	0.0674	0.0082	0.6903
Gain vs. 2nd best Single-Model	1.24%↑	1.09%↑	1.24%↑	1.43%↑	3.30%↓	10.87%↓	5.13%↑
BERT	0.8128	0.6730	0.8041	0.6656	0.0718	0.0097	0.6366
RoBERTa	0.8133	0.6734	0.8089	0.6692	0.0701	0.0098	0.6360
DeBERTa	0.8163	0.6774	0.8109	0.6704	0.0700	0.0094	0.6485
DistilBERT	0.8116	0.6708	0.8031	0.6633	0.0717	0.0098	0.6327
EmoBERTa	0.8118	0.6670	0.8063	0.6547	0.0764	0.0100	0.6282
Hybrid Model							
Simple+Concat	0.8149	0.6732	0.8114	0.6690	0.0714	0.0097	0.6367
Gated Fusion	0.8152	0.6744	0.8097	0.6718	0.0717	0.0098	0.6355
RoBERTa+CNN	0.8142	0.6731	0.8055	0.6624	0.0733	0.0096	0.6430
RoB-ERTa+BiLSTM	0.8130	0.6713	0.8062	0.6596	0.0707	0.0099	0.6319
BERT+CNN	0.8104	0.6666	0.8019	0.6592	0.0724	0.0099	0.6306
BERT+GRU	0.8153	0.6764	0.8097	0.6700	0.0697	0.0092	0.6566

additional computational cost introduced by the dual-encoder backbone translates into stronger predictive quality on this benchmark rather than into efficiency-oriented gains.

4.5 Ablation Study

To assess the contribution of each component, ablation experiments are conducted under identical settings, including the same learning rate, batch size, number of epochs, data split, and hardware platform. GF denotes the Gated Feature Aggregation module, HG denotes the Hierarchical Guidance strategy, and DL denotes the Dynamic Loss Weighting mechanism. In the variant without DL, task weights are kept fixed at lambda_c = lambda_f = lambda_i = 1/3 during training.

Table 3 reports the effect of removing different modules. The first variant removes GF, HG, and DL simultaneously. The second removes HG and DL while keeping GF. The third removes only DL. The complete DyFuLM model includes all components.

The ablation results reveal a cumulative contribution pattern rather than an isolated single-module effect. Comparing w/o GF+HG+DL with w/o HG+DL isolates the effect of GF: after GF is added back, coarse accuracy increases from 0.8173 to 0.8189, fine accuracy increases from 0.6780 to 0.6793, and the regression metrics also improve. Comparing w/o HG+DL with w/o DL isolates the effect of HG: after HG is added back, fine accuracy increases from 0.6793 to 0.6822 and fine F1 increases from 0.6769 to 0.6790, while MAE and MSE are further reduced. Finally, comparing w/o DL with the full DyFuLM model isolates the effect of DL: the w/o DL setting already retains multi-task learning, but it uses equal static weights of 1/3 for the three objectives, whereas the full model updates them adaptively. Once dynamic loss weighting is enabled, all three tasks improve jointly. The monotonic improvement across Table 3 shows that GF improves cross-encoder interaction, HG refines task-aware representation learning, and DL stabilizes multi-task optimization.

Table 3. Ablation study of model components. Best and second-best results for each metric are highlighted in red and blue, respectively. Lower values are better for MAE and MSE.

Experiments	Coarse Acc	Coarse F1	Fine Acc	Fine F1	MAE	MSE	R^2
DyFuLM	0.8264	0.8215	0.6848	0.6814	0.0674	0.0082	0.6903
vs. 2nd best	0.92%↑	0.81%↑	0.38%↑	0.35%↑	1.75%↓	12.77%↓	6.09%↑
w/oGF+HG+DL	0.8173	0.8129	0.6780	0.6729	0.0715	0.0100	0.6277
w/o HG+DL	0.8189	0.8147	0.6793	0.6769	0.0707	0.0095	0.6439
w/o DL	0.8186	0.8149	0.6822	0.6790	0.0686	0.0094	0.6507

5 Discussion

The experimental results indicate that fine-grained review understanding benefits from coordinated modeling at three levels. Layer-wise fusion preserves informative intermediate semantics that are often omitted by last-layer pooling. Encoder-wise aggregation allows RoBERTa and DeBERTa to compensate for each other's representational bias.

Task-wise guidance then connects coarse polarity, fine-grained categorization, and intensity regression through a shared sentiment representation.

Several limitations remain. First, the dual-encoder design introduces additional computational cost relative to single-encoder baselines, so the present experiments evaluate accuracy-oriented behavior rather than deployment efficiency. Second, the data come from hotel reviews collected between 2015 and 2017, so the current evidence should be interpreted as within-domain validation rather than broad cross-domain generalization. Third, the mapped label spaces are imbalanced, especially for minority sentiment categories. Reporting Macro-F1 reduces the dominance of majority classes, but minority-category robustness still deserves more targeted evaluation. These limitations are closely related to the scope clarified in the revised study: the present paper emphasizes predictive quality, uses Macro-F1 to account for class imbalance, and avoids overstating cross-domain generalization beyond the hotel-review benchmark.

Future work will explore compression and distillation so that the collaborative representation learned by DyFuLM can be transferred to a lighter student model. Another direction is to extend the framework toward richer affective language applications that couple sentiment understanding with higher-level reasoning.

6 Conclusion

This paper proposes DyFuLM, a dynamic collaborative dual-encoder framework for multi-dimensional sentiment analysis. By combining RoBERTa and DeBERTa, DyFuLM captures both global contextual semantics and fine-grained affective cues. Hierarchical Dynamic Fusion improves the use of intermediate-layer information, Gated Feature Aggregation coordinates heterogeneous encoder representations, and Hierarchical Guidance together with Dynamic Loss Weighting strengthens multi-task interaction across coarse-grained classification, fine-grained classification, and sentiment intensity regression. Experiments on a large-scale hotel review benchmark show that DyFuLM consistently outperforms strong single-model and hybrid baselines across both classification and regression tasks, while the ablation study confirms that each proposed component contributes positively to the final performance.

The current evidence should therefore be interpreted as within-domain validation on hotel reviews and as a performance-oriented study rather than an efficiency-oriented deployment framework. Future work will extend DyFuLM toward compression-aware deployment and broader affective language applications.

7 Acknowledgment

This research is supported by Xi'an Jiaotong-Liverpool University, China, through a Research Development Fund Grant (RDF-21-02-049). Md Maruf Hasan acknowledges the financial support for the project titled "Development of a Trustworthy Framework for the Application of Artificial Intelligence Technologies in National Security and Social Good".

References

1. M. Gharzouli, A. K. Hamama, and Z. Khattabi, "Topic-based sentiment analysis of hotel reviews," *Current Issues in Tourism*, vol. 25, no. 9, pp. 1368–1375, 2022.
2. Y. Wen, Y. Liang, and X. Zhu, "Sentiment analysis of hotel online reviews using the BERT model and ERNIE model—Data from China," *PLOS ONE*, vol. 18, no. 3, p. e0275382, 2023.
3. W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.
4. R. Y. K. Lau, W. Zhang, and W. Xu, "Parallel aspect-oriented sentiment analysis for sales forecasting with big data," *Production and Operations Management*, vol. 27, no. 10, pp. 1775–1794, 2018.
5. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.
6. A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, "Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification," in *Proceedings of LREC*, pp. 4933–4941, 2020.
7. M. S. Akhtar, T. Garg, and A. Ekbal, "Multi-task learning for aspect term extraction and aspect sentiment classification," *Neurocomputing*, vol. 398, pp. 247–256, 2020.
8. Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proceedings of EMNLP*, pp. 1746–1751, 2014.
9. K. S. Tai, R. Socher, and C. D. Manning, "Improved Semantic Representations from Tree-Structured Long Short-Term Memory Networks," in *Proceedings of ACL*, pp. 1556–1566, 2015.
10. Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
11. P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," *arXiv preprint arXiv:2006.03654*, 2020.
12. S. Humeau, K. Shuster, M.-A. Lachaux, and J. Weston, "Poly-Encoders: Transformer Architectures and Pre-Training Strategies for Fast and Accurate Multi-Sentence Scoring," in *Proceedings of ICLR*, 2020.
13. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT: A distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
14. T. Kim and P. Vossen, "EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa," *arXiv preprint arXiv:2108.12009*, 2021.
15. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," in *Proceedings of ICML*, pp. 689–696, 2011.
16. M. S. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, and S. Kurohashi, "All-in-One: Emotion, Sentiment and Intensity Prediction Using a Multi-Task Ensemble Framework," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 285–297, 2022.
17. Q. Labeed and X. Liang, "Multi-Task Learning Transformers: Comparative Analysis for Emotion Classification and Intensity Prediction in Social Media," in *Proceedings of ICPRS*, pp. 1–7, 2024.
18. Y. Huang, X. Zhou, R. Yin, and P. Shi, "External-knowledge enhanced dual encoder and contrastive learning for aspect sentiment triplet extraction," *PLOS ONE*, vol. 21, no. 3, p. e0340792, 2026.