

DyFuLM: An Advanced Multimodal Framework for Sentiment Analysis

Ruohan Zhou¹, Jiachen Yuan², Churui Yang³, Wenzheng Huang², Guoyan Zhang¹,
Shiyao Wei¹, Jiazhen Hu², Ning Xin², Md Maruf Hasan^{*2}

¹Department of Applied Mathematics, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, Jiangsu, China

²School of AI and Advanced Computing, XJTLU Entrepreneur College (Taicang),
Xi'an Jiaotong-Liverpool University, Suzhou, 215123, Jiangsu, China

³Department of Intelligent Science, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, Jiangsu, China

Ruohan.Zhou22@student.xjtlu.edu.cn, Jiachen.Yuan23@student.xjtlu.edu.cn, Churui.Yang24@student.xjtlu.edu.cn,
Wenzheng.Huang23@student.xjtlu.edu.cn, Guoyan.Zhang23@student.xjtlu.edu.cn, Shiyao.Wei22@student.xjtlu.edu.cn,
Jiazhen.Hu24@student.xjtlu.edu.cn, Ning.Xin21@student.xjtlu.edu.cn, MdMaruf.Hasan@xjtlu.edu.cn

Abstract—Understanding sentiment in complex textual expressions remains a fundamental challenge in affective computing. To address this, we propose a Dynamic Fusion Learning Model (DyFuLM), a multimodal framework designed to capture both hierarchical semantic representations and fine-grained emotional nuances. DyFuLM introduces two key modules: a Hierarchical Dynamic Fusion module that adaptively integrates multi-level features, and a Gated Feature Aggregation module that regulates cross-layer information flow to achieve balanced representation learning. Comprehensive experiments on multi-task sentiment datasets demonstrate that DyFuLM achieves 82.64% coarse-grained and 68.48% fine-grained accuracy, yielding the lowest regression errors (MAE = 0.0674, MSE = 0.0082) and the highest coefficient of determination ($R^2 = 0.6903$). Furthermore, the ablation study validates the effectiveness of each module in DyFuLM. When all modules are removed, the accuracy drops by 0.91% for coarse-grained and 0.68% for fine-grained tasks. Keeping only the gated fusion module causes decreases of 0.75% and 0.55%, while removing the dynamic loss mechanism results in drops of 0.78% and 0.26% for coarse-grained and fine-grained sentiment classification, respectively. These results demonstrate that each module contributes significantly to feature interaction and task balance. Overall, the experimental findings further validate that DyFuLM enhances sentiment representation and overall performance through effective hierarchical feature fusion.

Index Terms—Multimodal Framework, Sentiment Analysis, Dual-Encoder Model, Multi-task Learning, Feature Fusion

1 INTRODUCTION

With the rapid growth of the digital economy, hotel reviews on social media and online platforms have become a key factor influencing tourists' decisions [1]. Traditional methods, such as manual questionnaires, rating-based evaluations, and statistical summaries, provide limited interpretability and fail to capture the implicit emotions and contextual subtleties embedded in travelers' feedback. Compared to traditional methods, the ratings offer insight into perceptions of services and environments. Analyzing the rating of such reviews supports a better understanding of user behavior, enables service optimization, and improves tourism demand forecasting [2]. Sentiment analysis has been widely adopted in tourism re-

search, helping to allocate resources and personalized recommendations [3].

Conventional sentiment analysis methods typically classify text as positive, negative, or neutral based on sentiment lexicons, handcrafted statistical features, or traditional classification algorithms [4]. However, these approaches often overlook subtle or mixed emotions within the same review, for example, when a user praises the service of a hotel but criticizes its facilities [5, 6, 7]. The emergence of deep learning has transformed sentiment analysis, with large language models (LLMs) such as BERT exhibiting exceptional capabilities in contextual understanding and semantic representation [8]. Compared to traditional methods, models such as BERT and RoBERTa achieve higher accuracy and robustness in sentiment classification [9, 10]. Nevertheless, two significant challenges persist: (1) model performance varies across text types and domains due to differences in vocabulary and linguistic expression, which reduces accuracy [11]; (2) most existing models still rely on coarse sentiment polarity classification (positive or negative), lacking the ability to quantify emotional intensity [12].

To address the above challenges, we propose a Dynamic Fusion Learning Model (DyFuLM). The model integrates disentangled attention to achieve fine-grained semantic representation and improve contextual dependency to enhance robustness. By a multimodal framework, DyFuLM effectively improves emotion recognition in complex and mixed review scenarios. Evaluated on a large-scale dataset containing 515,738 reviews from 1,493 European luxury hotels collected between 2015 and 2017, DyFuLM demonstrates strong performance in sentiment classification tasks. To further enhance domain adaptability, we conduct fine-tuning within the target domain using authentic hotel reviews, which improves both accuracy and robustness of the model. Moreover, DyFuLM employs a multimodal framework that integrates coarse-grained sentiment classification, fine-grained emotion categorization, and emotion intensity regression, enabling the model to capture subtle and multifaceted emotional expressions with greater precision.

The main research contributions of this study are:

1. We propose a multimodal framework that integrates global semantic information with local emotional cues to achieve comprehensive contextual modeling and fine-grained feature representation. The framework jointly performs coarse-grained classification, fine-grained emotion categorization, and emotion intensity regression within a unified architecture. This joint optimization strategy effectively reduces task bias and enhances the model’s generalization ability across multiple emotional dimensions;
2. To further refine feature interaction, we introduce a multimodal framework that spans both encoders. This module utilizes gating and adaptive weighting to facilitate cross-layer and cross-model semantic alignment, thereby enhancing representational accuracy;

2 RELATED WORK

Sentiment analysis has progressed through three main stages. Early methods relied on traditional machine learning algorithms such as Naïve Bayes, Support Vector Machines [13], and Decision Trees [14], which extracted handcrafted features using TF-IDF or N-gram representations but failed to capture contextual dependencies [14]. With the advent of neural networks, CNNs were introduced for local feature extraction and LSTMs for sequential modeling [15, 16]. However, these architectures required large labeled datasets and suffered from gradient vanishing over long sequences. The emergence of Transformer-based models, particularly BERT and its variants (RoBERTa [17], DeBERTa [18]), marked a major paradigm shift by leveraging self attention and large-scale pretraining to achieve state-of-the-art (SOTA) performance across NLP tasks. Despite these advances, single encoder architectures still face inherent limitations in large scale sentiment analysis. Their unified parameter space restricts representational diversity, leading to domain overfitting. Relying solely on final layer features overlooks valuable linguistic information from intermediate layers, while shared representations in multi-task learning often cause semantic interference across tasks. To overcome these limitations, recent studies have explored multi-encoder frameworks. Poly-Encoder and TwinBERT employ dual-tower architectures for efficient semantic matching but perform fusion only at the final layer, limiting cross layer interaction [19, 20]. MoE-BERT introduces expert routing for dynamic feature selection, yet its discrete gating mechanism lacks fine-grained semantic alignment [21]. Similarly, concatenation based and attention-weighted fusion strategies remain static and struggle to adaptively integrate cross model information. To address these challenges, we propose DyFuLM, a multimodal framework learning model with three key innovations: (1) a hierarchical dual-encoder architecture that enhances contextual representation through global–local semantic complementarity, (2) a BiLSTM-guided dynamic fusion mechanism enabling adaptive cross-layer feature extraction, and (3) a gated aggregation module that facilitates fine-grained cross model interaction under hierarchical multi-task

guidance. Together, these innovations enable comprehensive and fine-grained sentiment modeling in large scale scenarios.

3 METHODOLOGY

This section presents the overall methodology of our study, comprising four core components: Data Preprocessing, Workflow Design, Model Architecture, and Architectural Comparisons. Each component outlines the key procedures and design principles that collectively underpin the DyFuLM framework.

3.1 Data Preprocessing

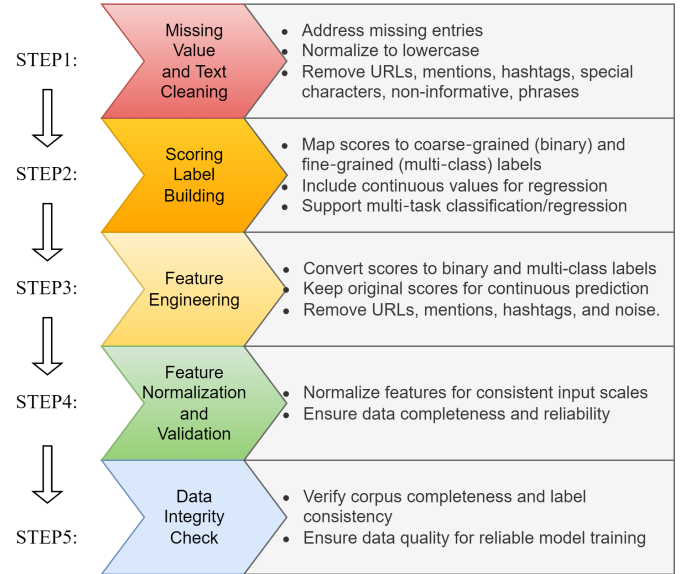


Fig. 1: Flowchart of Data Preprocessing

To ensure a clean and structured dataset for downstream tasks, the preprocessing procedure involved the Fig.1 and the following steps:

- 1) **Missing Value and Text Cleaning:** Missing entries were appropriately addressed. Texts were normalized by lowercasing and cleaned by removing URLs, mentions, hashtags, special characters, extra whitespace, and non-informative phrases.
- 2) **Scoring Label Building:** Reviewer scores were mapped to sentiment labels using both coarse-grained and fine-grained schemes to support flexible classification tasks.
- 3) **Feature Engineering:** Key features such as review length and temporal trends were extracted. Sentiment and length distributions were visualized, and geographic patterns were analyzed.
- 4) **Feature Normalization and Validation:** Textual and auxiliary features were transformed and normalized to ensure consistent scales among input variables. This procedure contributed to stable model training and better generalization performance.
- 5) **Data Integrity Check:** A final verification was conducted to confirm corpus completeness, label consistency, and data reliability, ensuring that the dataset

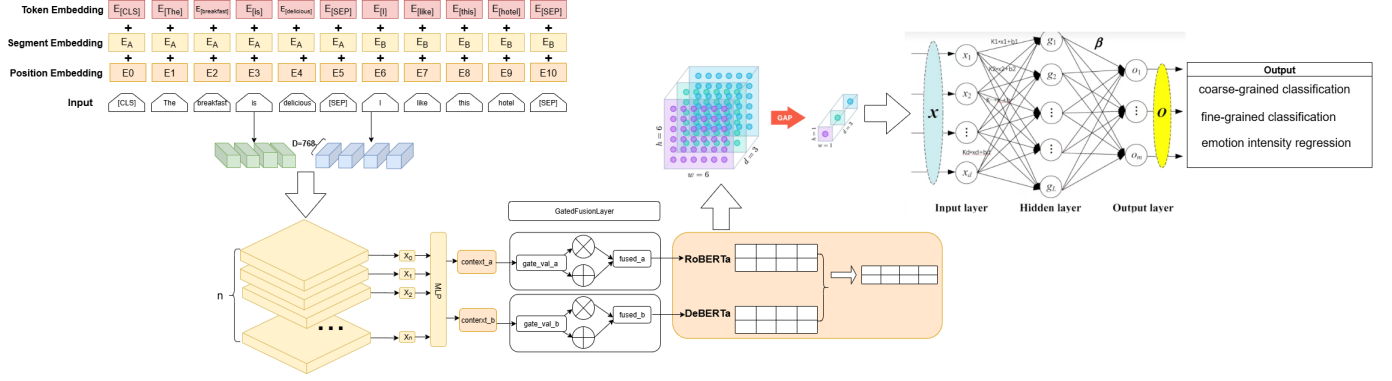


Fig. 2: DyFuLM Workflow

was suitable for subsequent model development and evaluation.

3.2 Workflow

Fig.2 delineates a multimodal emotion analysis framework. Initially, the preprocessed text is encoded through token, segment, and position embeddings to obtain the basic semantic representations. Subsequently, feature extraction is performed through a stack of layers, and the outputs are concatenated thereafter. A dynamic fusion feature processor, equipped with a gating mechanism, is then employed. This mechanism comprises two branches content_a and content_b, each integrating gate_val and load_val operations to dynamically regulate feature flow and fusion. Following this, features from multimodal are processed, followed by Global Average Pooling (GAP). Ultimately, a neural network consisting of input, hidden, and output layers yields results for coarse-grained classification, fine-grained classification, and emotion intensity regression.

3.3 Model Architecture

Traditional multi-task sentiment analysis models often rely on a single language encoder for semantic representation. While this structure provides effective global semantic understanding, it struggles to capture detailed emotional cues due to its limited representational depth. To overcome this limitation, we design a multimodal framework collaborative architecture that jointly models contextual semantics and fine-grained emotional features, improving both representation quality and feature integration.

Multimodal Framework Collaborative Hybrid Framework

DyFuLM introduces a cooperative multimodal framework design that enables dynamic interaction between different encoders. Specifically, RoBERTa contributes strong global contextual understanding and holistic sentiment representation, while DeBERTa provides finer-grained semantic discrimination and captures localized emotional cues. This collaborative integration allows DyFuLM to model both coarse and fine semantic structures more effectively, leading to improved downstream performance.

Hierarchical Dynamic Feature Fusion

Building upon the multimodal framework backbone, we further introduce a hierarchical dynamic feature fusion mechanism to leverage the diverse information captured at different layers of Transformers. Relying solely on the final layer may overlook informative intermediate representations. To address this, we design a layer-wise attention weighting strategy, allowing the model to dynamically select and aggregate features from multiple layers based on specific task needs. This enables a more flexible and comprehensive integration of multi-level semantics.

To capture inter-layer dependencies, we denote the hidden state of the l -th layer as $H^{(l)} \in \mathbb{R}^{T \times d}$, where T is the sequence length and d is the hidden dimension. A bidirectional LSTM is first applied to encode the representations from all layers, producing contextualized layer features: Eq.1[15]:

$$U^{(l)} = \text{BiLSTM}(H^{(1)}, H^{(2)}, \dots, H^{(L)}). \quad (1)$$

To adaptively integrate multi-layer information, a learnable attention weight $\alpha^{(l)}$ is assigned to each layer. After softmax normalization, the contribution of each layer is computed as: Eq.2[13]:

$$\alpha^{(l)} = \frac{\exp(w^T U^{(l)})}{\sum_{k=1}^L \exp(w^T U^{(k)})}. \quad (2)$$

Here, L denotes the total number of Transformer layers, and w is a learnable weight vector used to compute the attention score for each layer. The final fused representation is then obtained through a weighted sum: Eq.3

$$H_{\text{fused}} = \sum_{l=1}^L \alpha^{(l)} H^{(l)}. \quad (3)$$

This layer-wise attention mechanism allows the model to automatically focus on the most informative layers for each token, enhancing hierarchical feature representation.

Gating-based Cross-model Fusion

To enhance coordination between models, we introduce a gating mechanism to selectively integrate information between the representations of multimodal framework.

This mechanism acts as a controllable switch that adaptively determines, for each token, whether its final representation

should primarily rely on its own contextual features or incorporate complementary cues from the other model. Such an adaptive strategy avoids redundant information caused by simple concatenation and promotes more effective feature fusion.

Formally, let the token representation from RoBERTa be denoted as h_A , and the DeBERTa representation incorporating contextual information as c_B . The gating function is defined as: Eq.4

$$g = \sigma(W[h_A; c_B] + b), \quad (4)$$

where $\sigma(\cdot)$ denotes the sigmoid function, and W and b are learnable parameters. The operator $[\cdot; \cdot]$ indicates vector concatenation. The final fused representation is computed as: Eq.5

$$h_{\text{fused}} = g \odot h_A + (1 - g) \odot c_B. \quad (5)$$

This gating design enables the model to dynamically control the contribution of each source, achieving fine-grained and adaptive information fusion.

Multitasking Output Branches

To achieve multi-level task modeling, we design three independent yet hierarchically connected prediction heads. First, the coarse-grained head f_{coarse} takes the fused feature representation h as input and produces the coarse-level classification result: Eq.6

$$\hat{y}_c = f_{\text{coarse}}(h). \quad (6)$$

Then, the intensity head $f_{\text{intensity}}$ estimates the emotional strength from the same feature: Eq.7

$$\hat{y}_i = f_{\text{intensity}}(h). \quad (7)$$

Based on these outputs, a guidance function $g = \text{Guidance}(\hat{y}_c, \hat{y}_i)$ integrates both predictions to generate a guiding signal that recalibrates the original feature representation: Eq.8

$$h' = h \odot g, \quad (8)$$

where $\odot$ denotes multiplication of corresponding elements. Finally, the fine-grained head f_{fine} takes the recalibrated feature h' as input and produces the refined prediction: Eq.9

$$\hat{y}_f = f_{\text{fine}}(h'). \quad (9)$$

This hierarchical prediction mechanism from top to bottom allows the coarse classification and intensity modeling to mutually reinforce each other, enhancing both the model's representational capacity and predictive accuracy.

Here, h denotes the fused feature vector, $\hat{y}_c$ the coarse-level prediction, $\hat{y}_i$ the intensity prediction, g the guidance signal, and $\hat{y}_f$ the final fine-grained prediction.

3.4 Comparison of Model architectures

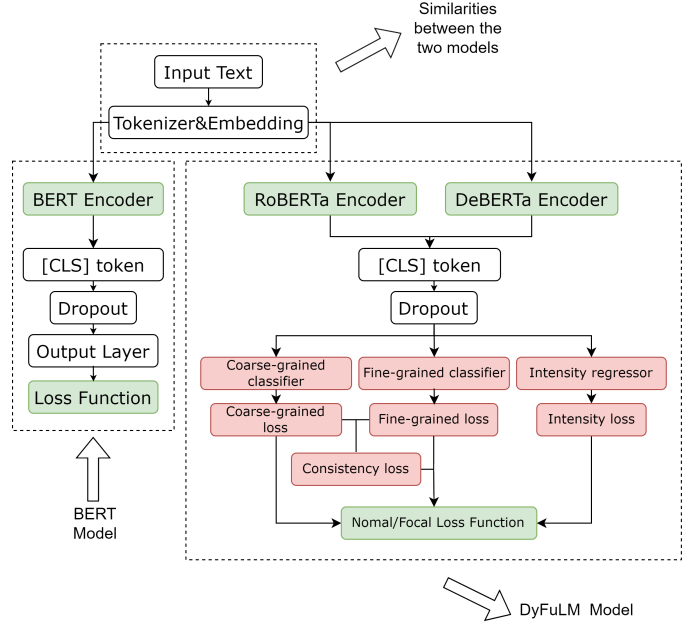


Fig. 3: Comparison between BERT and DyFuLM architectures (green denotes shared components, red denotes DyFuLM-specific modules)

DyFuLM adopts several key enhancements. Fig.3 illustrates the overall architecture, where DyFuLM integrates RoBERTa and DeBERTa as multimodal framework to enable parallel representation learning. To effectively combine their multi-layer features, a hierarchical dynamic fusion mechanism is employed. At the output stage, the model performs three tasks: coarse-grained classification, fine-grained classification, and emotion intensity regression, thereby supporting multi-dimensional sentiment modeling. This design preserves semantic understanding while substantially improving the model's granularity and generalization in emotion recognition.

4 EXPERIMENTS RESULTS AND ANALYSIS

4.1 Multivariate Time Series Data Display

This study utilizes a dataset of 515,738 user reviews from Booking.com, encompassing 1,493 hotels across Europe and collected between 2015 and 2017. Each record contains 17 attributes, including review text, rating, timestamp, user nationality, and hotel location. To minimize the influence of outliers and ensure consistency, the original review texts and ratings are directly used as standardized inputs for subsequent analysis.

Table I shows that user ratings for European hotels remained consistently high from 2015 to 2017, with a stable median of 8.8 and a 25th percentile of 7.5 across all three years, indicating a generally positive user perception. The average rating slightly fluctuated, with values of 8.32 in 2015, 8.42 in 2016, and 8.39 in 2017, yet remained at a high level overall. The maximum rating consistently reached 10, confirming a

TABLE I: Statistical characteristics of hotel ratings

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

stable scoring mechanism. As shown in Table I, the number of reviews peaked in 2016, marking a notable increase from 2015, followed by a slight decline in 2017. This fluctuation may be attributed to changes in platform activity or data collection policies.

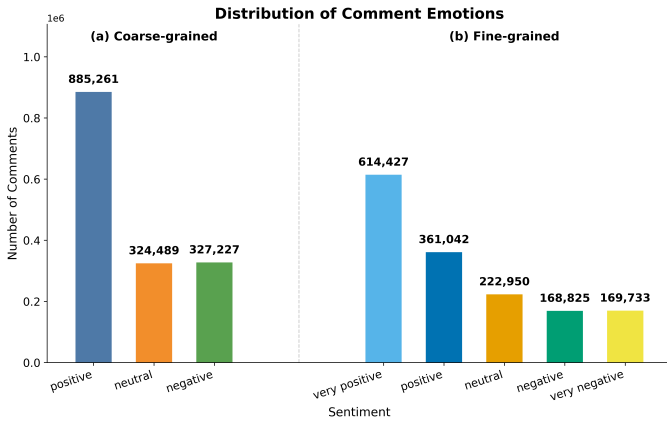


Fig. 4: Emotion Category Distribution of Comments

Fig.4 presents the fine-grained and coarse-grained sentiment distributions of the reviews. The fine-grained classification divides sentiments into five levels, allowing a more nuanced capture of users’ emotional variations. Among these, positive reviews (very positive and positive) exceed 975,469, forming the majority, while negative reviews total around 338,558. In contrast, the coarse-grained classification simplifies sentiment into three categories, offering a more intuitive overview. It shows 885,261 positive reviews, though at the cost of compressing sentiment detail. The fine-grained approach provides a clearer distinction in sentiment intensity, for example, differentiating between “very positive” and “positive” or between “slightly negative” and “strongly negative,” which is essential for modeling subtle emotional trends in user feedback.

4.2 Performance Metrics Overview

Table II presents a performance comparison between DyFuLM and several representative baselines. We evaluate the models across multiple dimensions. Specifically, Coarse Accuracy and Fine Accuracy measure classification performance at different granularities, while Coarse F1 and Fine F1 assess the balance between precision and recall. Additionally, we report the Mean Absolute Error (MAE) and Mean Squared Error (MSE) for regression tasks, where lower values indicate better prediction accuracy. We also include the R^2 (coefficient

of determination), which reflects the proportion of variance explained by the model.

4.3 Performance Comparison and Improvement Analysis

To benchmark our model, we evaluate it against two widely adopted baseline families.

Single Models. We select five representative models as baselines to ensure a comprehensive comparison across different Transformer developments. BERT serves as the standard baseline in NLP tasks [8]. RoBERTa, a widely adopted model that achieved state-of-the-art(SOTA) performance upon its release in 2019, reflects improvements in large scale pretraining [17]. DeBERTa-v3 represents advanced architectural optimization for contextual understanding [18]. DistilBERT serves as an efficiency oriented baseline that balances accuracy and speed [22]. Finally, EmoBERTa, fine-tuned for emotion recognition, provides a domain specific comparison point [23].

Hybrid Architectures. To ensure fair and representative evaluation, we include six widely used hybrid baselines that integrate pretrained Transformers with lightweight sequence learners or feature fusion schemes. These baselines are selected because they reflect mainstream strategies for enhancing contextual representations in sentiment analysis. Specifically, Simple + Concat and Gated Fusion represent typical feature fusion mechanisms [24, 25], while BERT/RoBERTa + CNN, + BiLSTM, and + GRU correspond to common Transformer encoder integration paradigms [26, 27, 28, 29]. Together, these baselines provide a comprehensive benchmark for evaluating the effectiveness of our approach.

TABLE II: Performance comparison with baselines (Values are presented to four decimal places, with optimal results for each metric in bold.)

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
<i>Single-Model</i>							
BERT	0.8128	0.6730	0.8041	0.6656	0.0718	0.0097	0.6366
RoBERT	0.8133	0.6734	0.8089	0.6692	0.0701	0.0098	0.6360
DeBERT	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
EmoBERT	0.8118	0.6670	0.8063	0.6547	0.0764	0.0100	0.6282
<i>Dual-Architecture</i>							
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
RoBERT+CNN	0.8142	0.6731	0.8055	0.6624	0.0733	0.0096	0.6430
RoBERT+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

We conduct a comprehensive comparison between DyFuLM and several representative baselines, including single-encoder models (BERT, RoBERTa, DeBERTa, DistilBERT, EmoBERTa) and dual-encoder architectures (Simple+Concat, Gated-Fusion, RoBERTa+CNN, BERT+CNN, RoBERTa+BiLSTM, BERT+GRU). As shown in Table II, DyFuLM consistently achieves superior results across all classification metrics, including Coarse Accuracy, Fine Accuracy, Coarse F1, and Fine F1, indicating a better balance between precision and recall.

In addition to classification performance, DyFuLM achieves the lowest MAE (0.0674) and MSE (0.0082) values among

all models, reflecting higher prediction accuracy in regression tasks. DyFuLM also leads in R^2 (0.6903), further emphasizing its strong generalization ability. These results confirm that DyFuLM not only improves overall classification performance but also exhibits strong generalization capabilities in minimizing prediction errors.

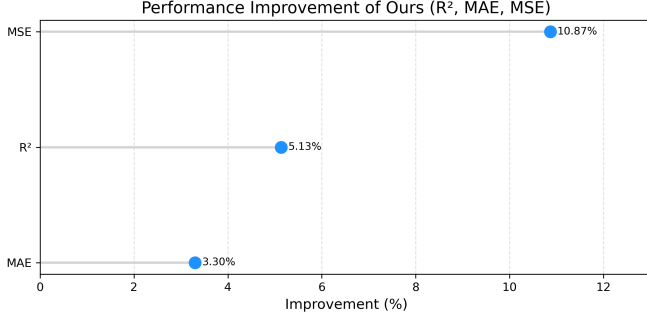


Fig. 5: Percentage of Performance Improvement

Fig.5 compares DyFuLM with the strongest baseline on the three regression metrics. First, MSE drops by 10.87%, indicating a substantial reduction in overall error. Second, the R^2 score rises by 5.13%, showing that the model explains a larger share of the variance in the target values. Finally, MAE decreases by 3.30%, confirming tighter point-wise predictions. Together, the simultaneous increase in R^2 and the decreases in both MAE and MSE demonstrate that the proposed model delivers more accurate and more reliable estimates across the full data range.

4.4 Ablation Study

To assess the effectiveness of each component in our hybrid multi-task sentiment analysis framework, we perform ablation studies by individually removing each module while keeping all other settings constant (learning rate = $1e-5$, batch size = 16, epochs = 6). All experiments are performed under identical hardware conditions, using an Intel Core i9-13900K CPU, an NVIDIA RTX 4090D GPU, and 64 GB RAM. This controlled setup ensures that performance variations are solely attributed to the excluded module, providing a clear and fair evaluation of its contribution.

TABLE III: Ablation study of model components (Values are presented to four decimal places, with optimal results for each metric in bold)

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
w/o GF+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

As shown in Table III, we conduct an ablation study to quantify the effect of each module. The first variant removes all modules, eliminating gated fusion, hierarchical guidance, and dynamic loss weighting. The second variant keeps only the gated fusion module to examine its independent contribution. The third variant removes the dynamic loss weighting while

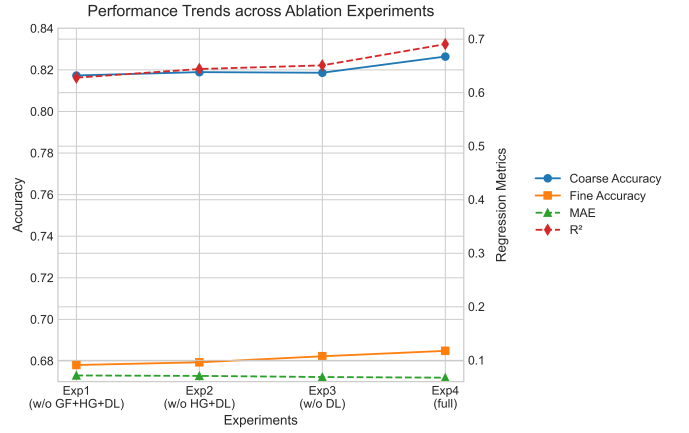


Fig. 6: Performance Trends across Ablation Experiments

preserving other components. The full model, DyFuLM, integrates all modules. This stepwise comparison clearly demonstrates how each design choice contributes to the overall performance.

The ablation results underscore the effectiveness of each proposed component. The complete DyFuLM model achieves the best overall results, with coarse-grained and fine-grained accuracies of 82.64% and 68.48%, respectively, along with the lowest MAE and MSE and the highest R^2 . When all modules are removed, the performance drops by 0.91% and 0.68%, highlighting the importance of module cooperation. Furthermore, retaining only the gated fusion module or removing the dynamic loss weighting mechanism leads to additional declines of 0.75% / 0.55% and 0.78% / 0.26%, respectively, indicating that both modules play key roles in feature interaction and task balance. Overall, the ablation results confirm the effectiveness of each component and demonstrate that their synergy significantly enhances DyFuLM’s semantic representation and sentiment analysis stability.

The ablation results in Fig.6 confirm the effectiveness of each proposed component. Among them, the gated fusion module yields the most substantial performance gain, followed by the hierarchical guidance and dynamic layer extraction modules. Together, these components form a unified framework that enhances both hierarchical understanding and fine-grained sentiment recognition, while also improving training stability.

5 CONCLUSION

In this study, we introduced DyFuLM, a multimodal framework for multi-dimensional sentiment analysis. By capturing both hierarchical semantics and fine-grained emotional cues, DyFuLM achieves this. The proposed hierarchical dynamic fusion and gated feature aggregation modules enable adaptive alignment across layers and models, improving contextual representation and sentiment understanding.

Comprehensive experiments demonstrated that DyFuLM consistently outperforms baseline models in tasks. It achieves higher coarse and fine-grained accuracy while maintaining

lower MAE and MSE, indicating a strong balance between expressiveness and precision. Ablation analyses confirmed that the gated fusion and hierarchical guidance modules are key to enhancing robustness and accuracy.

Despite these strengths, DyFuLM has several limitations. It currently focuses on textual input and has not incorporated multimodal signals such as images or audio. Moreover, its multimodal structure increases computational cost, which may limit scalability in real-time applications. In addition, evaluation has so far been limited to a specific domain, and broader validation is required to test its general applicability.

In future work, we plan to extend DyFuLM in three main directions. We will first integrate visual and acoustic modalities to develop multimodal sentiment analysis, enabling the model to capture richer emotional cues beyond text. Next, we aim to enhance generalization and robustness through cross-lingual and cross-domain transfer learning, addressing variations across datasets and contexts. Finally, we will improve computational efficiency by exploring lightweight architectures and model compression for practical deployment. These future efforts will further strengthen DyFuLM's adaptability and comprehensive understanding of human emotions in real-world scenarios.

ACKNOWLEDGMENT

This work was conducted at Xi'an Jiaotong-Liverpool University. The authors would like to thank the university for providing laboratory facilities and technical support. The study also benefited from the use of the Booking.com dataset.

REFERENCES

- [1] M. Gharzouli, A. K. Hamama, and Z. Khatibi, "Topic-based sentiment analysis of hotel reviews," *Current Issues in Tourism*, vol. 25, no. 9, pp. 1368–1375, 2022.
- [2] H. Li, H. Gao, and H. Song, "Sentiment mining of online reviews for the perception of hotel attributes," *Annals of Tourism Research*, vol. 102, p. 103647, 2023, doi: 10.1016/j.annals.2023.103647.
- [3] U. Gretzel, H. Werthner, C. Koo, and C. Lamsfus, "Conceptual foundations for understanding smart tourism ecosystems," *Computers in Human Behavior*, vol. 50, pp. 558–563, 2015, doi: 10.1016/j.chb.2015.03.043.
- [4] 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, doi: 10.1016/j.asej.2014.04.011.
- [5] X. Li, C. Wu, and F. Mai, "Parallel aspect-oriented sentiment analysis for sales forecasting with big data," *Production and Operations Management*, vol. 27, no. 10, pp. 1778–1794, 2018.
- [6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.
- [7] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.
- [8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, 2019.
- [9] M. E. Peters, S. Ruder, and N. A. Smith, "To tune or not to tune? Adapting pretrained representations to diverse tasks," in *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pp. 7–14, 2019.
- [10] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, "BERT applications in natural language processing: a review," *Artificial Intelligence Review*, vol. 58, no. 6, p. 166, 2025.
- [11] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, "Adapt or Get Left Behind: Domain Adaptation through BERT Language Models," arXiv preprint arXiv:1908.11860, 2019.
- [12] 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.
- [13] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86, Philadelphia, USA, 2002.
- [14] K. L. Tan, C. P. Lee, and K. M. Lim, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research," *Applied Sciences*, vol. 11, no. 16, Article 7827, 2021.
- [15] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, 2014.
- [16] K. S. Tai, R. Socher, and C. D. Manning, "Improved Semantic Representations from Tree-Structured Long Short-Term Memory Networks," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1556–1566, Beijing, China, 2015.
- [17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
- [18] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," arXiv preprint arXiv:2006.03654, 2021.
- [19] 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," arXiv preprint arXiv:1905.01969, 2020.
- [20] J. Lu, H. Zamani, V. Vinay, M. Joshi, S. Min, V. Zhong,

- and M. Grbovic, “TwinBERT: Distilling Knowledge to Twin-Structured BERT Models for Efficient Retrieval,” *arXiv preprint arXiv:2002.06275*, 2020.
- [21] J. Zuo, X. Zhao, T. Liu, and Y. Lin, “MoE-BERT: A Mixture of Experts Model Based on BERT for Multiple Downstream Tasks,” *arXiv preprint arXiv:2110.01799*, 2021.
 - [22] 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.
 - [23] T. Kim and P. Vossen, “EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa,” *arXiv preprint arXiv:2108.12009*, 2021.
 - [24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal Deep Learning,” *Proceedings of the 28th International Conference on Machine Learning (ICML)*, Bellevue, Washington, USA, pp. 689–696, 2011.
 - [25] K. Wei, J. Dai, D. Hong, and Y. Ye, “MGFNet: An MLP-dominated gated fusion network for semantic segmentation of high-resolution multi-modal remote sensing images,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 135, Article 104241, 2024.
 - [26] Y. Liu, S. Wang, and S. Yu, “A Bullet Screen Sentiment Analysis Method That Integrates the Sentiment Lexicon with RoBERTa-CNN,” *Electronics*, vol. 12, no. 2, Article 464, 2023.
 - [27] K. Kaur and P. Kaur, “BERT-CNN: Improving BERT for Requirements Classification using CNN,” *Procedia Computer Science*, vol. 218, pp. 2604–2611, 2023.
 - [28] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, “RoBERTa-BiLSTM: A Context-Aware Hybrid Model for Sentiment Analysis,” *Journal of LaTeX Class Files*, vol. 14, no. 8, pp. 1–10, 2021.
 - [29] L. Horne, M. Matti, P. Pourjafar, and Z. Wang, “GRU-BERT: A GRU-Based Method to Fuse BERT Hidden Layers for Twitter Sentiment Analysis,” *arXiv preprint arXiv:2010.03092*, 2020.