Dynamic Tuning and Multi-Task Learning Based Model for Multimodal Sentiment Analysis

Anonymous ACL submission

Abstract

Multimodal sentiment analysis aims to uncover human affective states by integrating data from multiple sensory sources. However, previous studies have focused on optimizing the model architecture, neglecting the impact of objective function settings on model performance. Given this, this study introduces a new framework - DMMSA, which integrates uni and multimodal sentiment analysis tasks, utilizes the intrinsic correlation of sentimental signals, and enhances the model's understanding of complex sentiments. In addition, it reduces task complexity by incorporating coarse-grained sentiment analysis. Meanwhile, the framework embeds a contrastive learning mechanism within the modality, enhancing the ability to distinguish between similar and dissimilar features. We conducted experiments on CH-SIMS, MOSI, and MOEI. The results showed that DMMSA outperformed the baseline method in classification and regression tasks when the model structure was unchanged, and only the optimization objectives were replaced.

1 Introduction

Multimodal Sentiment Analysis (MSA) revolves around integrating and synergistic parsing of diverse heterogeneous data modalities (Yang et al., 2024; Du et al., 2024; Liu et al., 2024), encompassing many information forms such as text, visual, auditory, and even biometric markers. (Sun et al., 2023) With the evolution of social media ecosystems and the proliferation of multimedia content, information presentation has evolved from pure text to richly illustrated content, culminating in today's prevalent video-based information (Yang et al., 2023; Huang et al., 2024).

Traditional unimodal sentiment analysis is confined mainly to the textual domain (Zeng et al., 2024). In contrast, MSA encompasses a comprehensive interpretation of multiple perceptual channels, including visual cues (e.g., facial expression, scene color, and body movement) and audio characteristics (e.g., pitch amplitude, frequency distribution, and speech tempo) (Hu et al., 2022; Ging et al., 2020). MSA has garnered significant attention in recent research. On the one hand, human sentimental expression inherently possesses cross-modal properties, with text, speech, and even haptic cues intricately interwoven to form sentiments, rendering a single modality insufficient to fully unveil the complexity of sentiments. (Lu et al., 2024; Shi et al., 2024) On the other hand, MSA technologies, through their deep fusion of multiple signal sources, significantly enhance the accuracy of sentiment recognition and understanding, fulfilling the high-precision sentimental intelligence demands in domains such as intelligent customer service, VR/AR experience optimization, and precise mental health assessment (Truong and Lauw, 2019; Feng et al., 2024).

Multimodal fusion has emerged as a core technique for understanding video contexts, demonstrating its value across numerous downstream tasks (Liang et al., 2022; Mai et al., 2022; Sun et al., 2020). Prior research has proposed a series of fusion techniques for MSA. For instance, Yu et al (Yu et al., 2020). employ a self-supervised joint learning strategy called self-mm, it integrating discriminative information learned from individual unimodal tasks with shared similarity information from the multimodal task during the late fusion stage, thereby enhancing model performance.

While multimodal fusion techniques are crucial for models (Fu et al., 2024; Jiang et al., 2024), setting optimization objectives is equally indispensable in model construction (Yang et al., 2023). Suitable optimization objectives effectively guide the model towards continuous performance optimization throughout training (Yang et al., 2023). Moreover, as shown in Figure 1, optimization objective setting and model structure optimization focus on different modules, complementing each other.

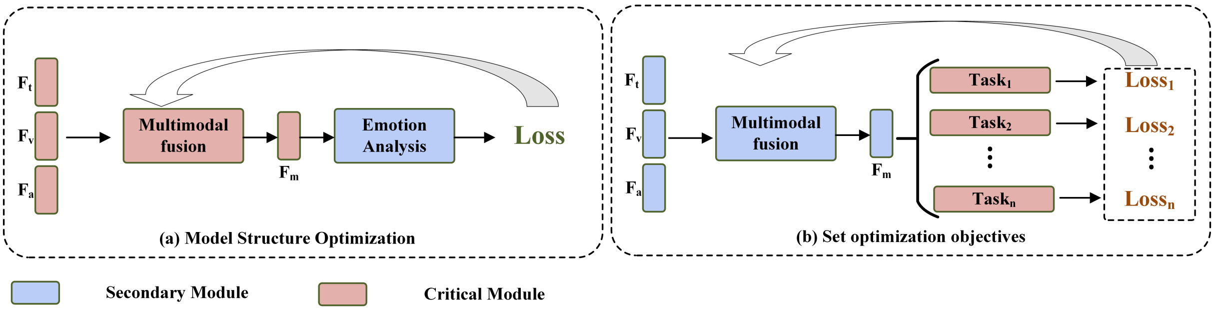

Figure 1: (a) Illustrates critical and secondary modules in the process of optimizing model architecture. Such approaches focus on enhancing the unimodal feature extraction and multimodal feature fusion modules. (b) Presents the modules of primary concern in our work. We concentrate on improving model performance by setting appropriate optimization objectives while maintaining the unchanged structure of other modules.

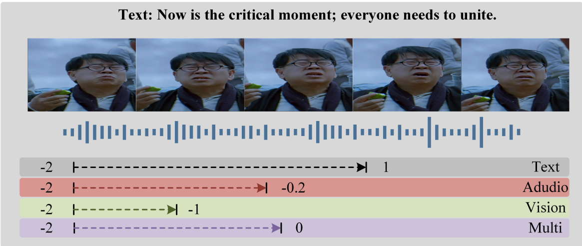

Figure 2: An example from the CH-SIMS dataset.

In MSA tasks, unimodal sentiments directly impact overall sentiments (Aslam et al., 2023; Truong and Lauw, 2019; Liu et al., 2019). As shown in Figure 2, each modality carries unique sentimental tendencies. Therefore, the model should be able to consider both uni and multimodal sentiments comprehensively. Another challenge faced by MSA tasks lies in their broad range of sentiment ratings. For example, the MOSI dataset requires models to accurately map samples to the sentiment intensity scale of [-3,+3], increasing the prediction difficulty.

Given these challenges, we propose DMMSA, a **D**ynamic Tuning and **M**ulti-Task Learning **MSA** model. DMMSA ensures the model can capture unimodal signals in detail and integrate multimodal information through collaborative optimization of unimodal and multimodal tasks. The model is equipped with a text-oriented contrastive learning module to promote feature decoupling and enhance the depth and accuracy of sentimental understanding. Furthermore, incorporating coarse-grained sentiment classification tasks to converge the prediction range has improved the accuracy of sentiment intensity determination. We implemented Global Dynamic Weight Generation(GDWG) to avoid negative transfer effects and achieve joint adjustment of model parameters, thereby maximizing overall performance.

The main contributions of this paper can be summarized as follows:

1. We propose the Multi Nt-Xent loss to guide the model in decomposing unimodal features and establishing text-centered contrastive relations.

2. By employing coarse-grained sentiment analysis tasks, we effectively converge the prediction range, reducing the complexity of modeling sentimental intensity.

3. To address the issue of unequal convergence rates among different tasks during multitask training, we propose the GDWG strategy, effectively mitigating negative transfer effects arising from such mismatches.

Our model is evaluated on three benchmark datasets: CH-SIMS[6], MOSI[9], and MOSEI[10]. The results showed that DMMSA outperformed the baseline method in classification and regression tasks when the model structure was unchanged, and only the optimization objectives were replaced. Additionally, we conduct comprehensive ablation studies, substantiating the efficacy of each component within our proposed architecture.

2 Related Work

2.1 Multimodal Sentiment Analysis

As a core topic in affective computing research, MSA has primarily been focused on representation learning and multimodal fusion strategies by past scholars. In representation learning, Wang et al (Wang et al., 2019). introduced the Recurrent Attended Variation Embedding Network (RAVEN), tailored for fine-grained structural modeling of non-verbal subword sequences, dynamically adjusting word-level representations in response to non-verbal cues. Regarding multimodal fusion techniques, Zaden et al (Zadeh et al., 2017). designed the Tensor Fusion Network to deeply model intra-modal and inter-modal relationships in online video analysis, addressing the transient variability of spoken language, sign language, and audio signals. Subsequently, they advanced the

Memory Fusion Network (Zadeh et al., 2018), employing attention mechanisms for interactive information integration across different views. Sun et al (Sun et al., 2023)., attentive to heterogeneity issues, proposed an attention-based cross-modal fusion scheme that facilitates modal interactions through attention mechanisms, promoting the effective alignment of distinct modal features. However, these efforts overlooked the potential impact of optimization objective design on model performance.

Yu et al (Yu et al., 2020). recognized the significant influence of unimodal sentimental expressions on overall affective states, leading to the construction of the CH-SIMS dataset, encompassing both uni- and multimodal sentiment intensity measures. Their study employed the L1 loss function for multimodal and unimodal sentiment analysis as a joint optimization objective. Experimental results revealed that incorporating unimodal sentiment analysis tasks enhanced the model's accuracy in predicting holistic sentimental dispositions. Yang et al. (Yang et al., 2023). decomposed unimodal representations into similarity and dissimilarity components, utilizing a text-centric contrastive learning approach. However, when implementing multi-task learning, they failed to adequately account for potential negative transfer effects resulting from pronounced disparities in task convergence rates and loss scales.

In contrast, the DMMSA model introduces a GDWG mechanism, enabling the model to adaptively adjust task weights based on the relative rates of loss decrease during training, effectively mitigating the detrimental impact of negative transfer on model performance. Moreover, DMMSA incorporates coarse-grained sentiment analysis tasks to constrain the prediction scope.

2.2 Contrastive Learning

Contrastive learning systematically constructs and discriminates between feature differences in positive and negative sample pairs to reveal intrinsic structural relationships within data (Hu et al., 2022; Yang et al., 2023; Khosla et al., 2021; Lei et al., 2021). This strategy has proven particularly effective in multimodal feature fusion research (Li et al., 2020). Specifically, Radford et al (Radford et al., 2021). employed multimodal contrastive learning techniques to align image-text pairs, effectively alleviating inherent data heterogeneity between visual and textual modalities and fostering widespread application in diverse multimodal downstream tasks such as visual question answering and caption generation. Similarly, Akbari et al (Akbari et al., 2021). trained a vision-audio-text translation model using the same contrastive learning approach, successfully achieving deep alignment among these three modalities.

In performing MSA tasks, Yang et al. (Yang et al., 2023). devised two contrastive learning mechanisms, intra-modal contrast, and inter-modal contrast, to guide the model toward generating features that embody homogeneity across modalities and capture heterogeneity between them. This strategy ensures that the model attends equally to commonalities and differences in modal interactions during modeling. Nonetheless, while this method yielded promising results, it did not address the limitation of traditional NT-Xent loss functions, which are tailored for single positive pair settings and ill-suited for scenarios involving multiple positive pairs; NT-Xent loss is

$$L_{NTX} = -\sum_{(a,p)\in P} \log \frac{\exp(sim(a,p)/\tau_m)}{\sum_{(a,k)\in N\cup P} \exp(sim(a,k)/\tau_m)} \tag{1}$$

where, τ_m is the temperature coefficient controlling the similarity distribution. (a, p) and (a, k) denote positive and negative sample pairs, respectively. N represents the set of negative pairs, while P signifies the set of positive pairs. Assuming the model has already converged, the formula can be further simplified as follows:

$$L_{NTX} = -\sum_{(a,p)\in P} \log \frac{1}{n} \tag{2}$$

where, the symbol n denotes the number of positive sample pairs. Observing the above formula, it becomes evident that when dealing with a single positive pair scenario, i.e., $n = 1$, the Contrastive Loss (CL) value precisely equals zero. However, the CL manifestly fails to converge to zero in situations involving more than one positive pair, i.e., $n > 1$. In light of this limitation, this paper, while leveraging contrastive learning strategies to aid the model in extracting both similar and dissimilar features, proposes an improvement to the NT-Xent loss function tailored to accommodate multiple positive instances, namely the Multi NT-Xent loss:

$$L_{MNTX} = -\log \frac{\sum_{(a,p)\in P} \exp(sim(a,p)/\tau_m)}{\sum_{(a,k)\in N\cup P} \exp(sim(a,k)/\tau_m)} \tag{3}$$

Under the condition of model convergence, the loss function can be further simplified as:

$$L_{MNTX} = -\log \frac{n}{n} \tag{4}$$

Consequently, the model is effectively guided in its contrastive learning tasks, whether faced with a single positive pair or multiple ones.

3 Methodology

3.1 Problem Formulation

MSA aims to decipher sample sentimental states by harnessing multiple signals, encompassing text (I_t), visual (I_v), and audio (I_a) modalities. Task types within this domain are typically categorized into two broad classes: classification and regression. Focusing on the latter, the proposed DMMSA model takes I_t, I_v, and I_a as inputs, yielding an output sentimental intensity value $y*$, constrained within the actual interval $[-R, R]$, where R defines the upper and lower bounds of the sentiment score.

3.2 Model Architecture

The overall architecture of the DMMSA model is depicted in Figure 3, following a processing flow outlined as follows: Given an input sample, unimodal data sources (I_t, I_v, I_a) are first subjected to feature extraction through dedicated unimodal encoders. Subsequently, a feature decomposition layer disassembles the encoded unimodal features, extracting similarity and dissimilarity components. Ultimately, these decomposed features are fed into a multimodal MLP module, which generates the final sentiment analysis output.

In pursuit of optimized model training, DMMSA is engineered to concurrently execute four tasks: ① Fine-grained Multimodal Sentiment Regression (MSR), aimed at precise quantification of sentimental intensity; ② Coarse-grained Multimodal Sentiment Classification (MSC), serving to restrict the prediction space; ③ Unimodal Sentiment Analysis, reinforcing the learning of unimodal representations; and ④ Contrastive Learning, enhancing the model's discriminative ability between similar and dissimilar features. Strategically, the integration of unimodal and multimodal sentiment tasks encourages the model to account for both multimodal and unimodal sentiments, while the coarse-grained classification task imposes a bounded prediction scope, enhancing localizationaccuracy. Contrastive learning further refines feature discriminability.

When constructing the loss function to multitask joint training, we introduce a GDWG mechanism cognizant of the potential disparity in gradient update rates among different tasks. This method is designed to balance the gradient descent rates across all tasks, effectively mitigating negative transfer effects arising from gradient misalignment and ensuring the stability and efficiency of the overall learning process. The resultant aggregate loss is

$$L_{MSA} = L_{MSR} + L_{MSC} + \lambda_{Uni} L_{Uni} + \lambda_{CL} L_{CL} \tag{5}$$

where, λ denotes the weights assigned to each task by the GDWG method. L_{MSR} represents the loss for Multimodal Sentiment Regression, L_{MSC} stands for the loss associated with Multimodal Sentiment Classification, L_{Uni} signifies the loss for Unimodal Sentiment Analysis, and L_{CL} denotes the Contrastive Learning loss. MSA is the core task of our model. To ensure the stability and coherence of its learning process, we have fixed the weights associated with L_{MSR} and L_{MSC}, which are closely tied to the performance of the MSA task. Meanwhile, we adjust the weights of the L_{uni} and L_{CL} tasks to enhance the MSA task's learning efficacy while minimizing any potential perturbations they may introduce to the learning trajectory of the MSA task.

L_{MSR}: The Multimodal Sentiment Regression loss aims to guide the model in integrating signals from different modalities to estimate the sentimental intensity of samples accurately. Herein, we feed the fused decomposed similarity and dissimilarity features into a Multimodal MLP for sentiment intensity prediction, associating its output with the given multimodal sentimental intensity labels via a Smooth L1 loss function to derive this loss. The formulaic expression is as follows:

$$y^* = MLP([T_s; T_d; A_s; A_d; V_s; V_d]) \tag{6}$$

$$L_{mul} = \left\{ \begin{array}{l} 0.5 * (\frac{y^* - y}{\varphi})^2, if(\frac{y^* - y}{\varphi}) < 1 \\ (y^* - y) - 0.5\varphi, otherwise \end{array} \right\} \tag{7}$$

where y^* represents the predicted result, y represents the multimodal sentiment label, and φ controls the smoothness.

L_{MSC}: The Multimodal Sentiment Classification loss aims to guide the model in coarse-grained categorization of sentimental states, thereby constraining the prediction space and facilitating precise targeting in sentiment analysis tasks. Here, we first map the sentimental intensity labels of samples to predefined sentimental polarity categories (e.g., positive, negative, neutral) according to pre-established rules, forming an sentimental polarity label set. Subsequently, the decomposed multimodal features are effectively concatenated and passed as input to a sentiment classifier, which

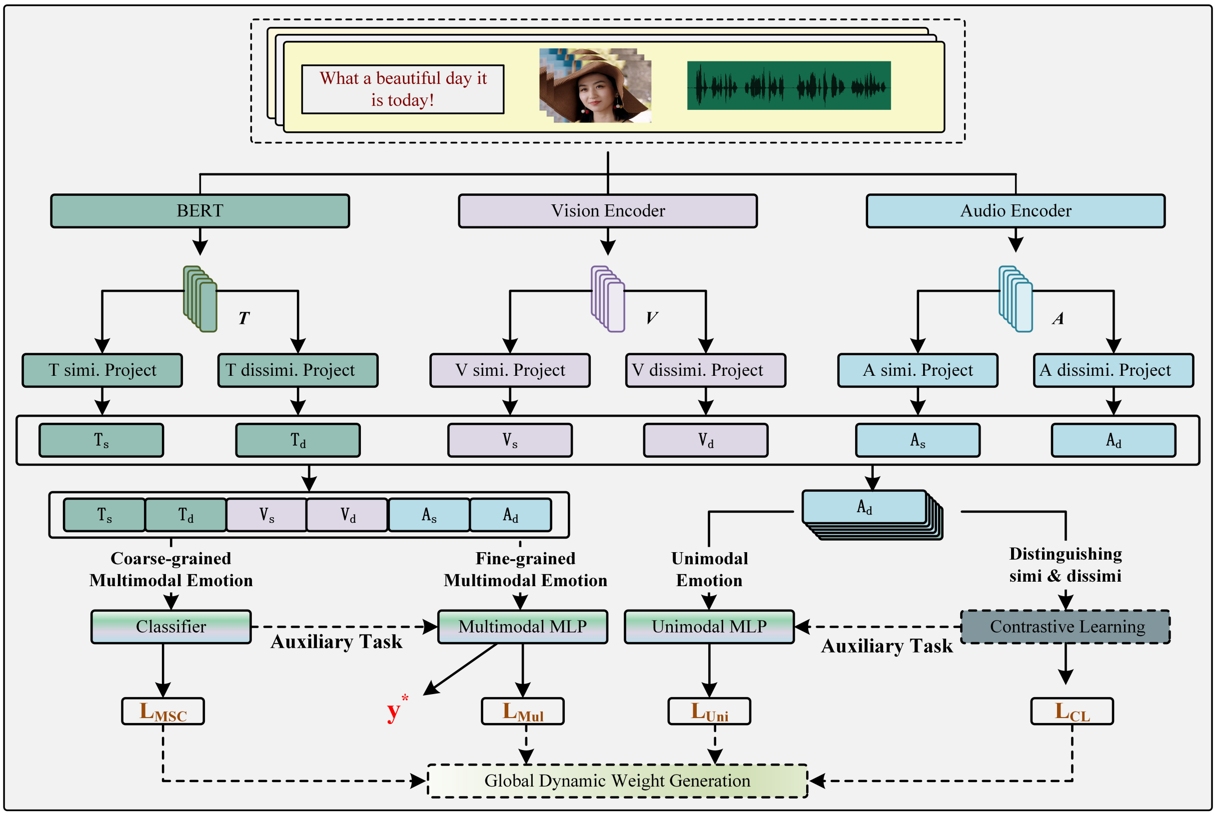

Figure 3: The overall framework of the DMMSA model.

yields the probability distribution for each sample across various sentimental polarities. Finally, the classifier's predicted probability distribution is compared with the actual assigned sentimental polarity labels, with the cross-entropy loss function employed to quantify the loss between the two. The specific formulaic expression is as follows:

$$y_{MSC} = Classifier([T_s; T_d; A_s; A_d; V_s; V_d]) \tag{8}$$

$$L_{MSC} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(y_{MSC}^{i,c}) \tag{9}$$

where, N denotes the number of samples, and C represents the number of categories.

L_{Uni}: The Unimodal Sentiment Analysis loss aims to guide the model in delving into the sentimental information embedded within each modality. Here, to ensure consistent treatment of modal features, we feed the similarity features (T_s, V_s, A_s) and dissimilarity features (T_d, V_d, A_d) of each modality separately into a weight-sharing multilayer perceptron (MLP) layer. The MLP layer outputs six sentiment predictions $u^* = MLP([T_s, T_d, A_s, A_d, V_s, V_d])$, with similarity features used to infer the multimodal sentiment label y and dissimilarity features employed to predict the corresponding unimodal sentiment labels $y^{t/v/a}$. In the absence of unimodal labels, the dissimilarity feature prediction task adjusts to predict the multimodal label y instead, maintaining the coherence of model training. Finally, a Smooth L1 loss function is employed for each prediction to measure the loss between the prediction and the respective ground truth label $u = [y, y, y, y^t, y^v, y^a]$. The specific formulaic expression is as follows:

$$L_{mul} = \left\{ \begin{array}{l} 0.5 * (\frac{u^*-u}{\varphi})^2, if(\frac{u^*-u}{\varphi}) < 1 \\ (u^* - u) - 0.5\varphi, otherwise \end{array} \right\} \tag{10}$$

L_{CL}: The Contrastive Loss aims to guide the model in effectively performing feature decomposition, allowing it to discern similarities and dissimilaritys among features sensitively. Here, considering that text modality data often assumes a dominant role in MSA tasks, with other modalities providing auxiliary information to enhance prediction accuracy, we opt to use text data as the reference anchor for constructing positive and negative sample pairs [3]. The specific configuration is as follows:

$$N = \{(T_s, T_d), (T_s, V_d), (T_s, A_d)\} \tag{11}$$

$$P = \{(T_s, V_s), (T_s, A_s)\} \tag{12}$$

Subsequently, we employ our proposed Multi NT-Xent Loss to guide the model in maximizing similarity between positive sample pairs while minimizing similarity between negative sample pairs. The calculation formula for Multi NT-Xent Loss is given by Equation (3).

Dataset	Train	Valid	Test	Total
CH-SIMS	1368	456	457	2281
MOSI	1284	229	686	2199
MOSEI	16326	1871	4659	22856

Table 1: Dataset-specific partitioning details.

3.3 Global Dynamic Weight Generation

In multitasking learning scenarios, distinct tasks often exhibit asynchronous convergence patterns, leading specific tasks to stabilize prematurely or tardily [7][8]. This inconsistency in convergence rates can engender negative transfer, where the learning process of one task adversarially impacts the performance of other tasks, thereby compromising overall model effectiveness. To address this challenge, we introduce a GDWG mechanism. This mechanism aims to adaptively adjust the relative weights of individual tasks during training, specifically by assessing the descent rate of each task's loss function at every training stage and, based on these assessments, generating weight values for each task. The specific mathematical expression is presented below:

$$w_k(t-1) = \frac{L_k(t-1)}{L_k(1)} \tag{13}$$

$$\lambda_k(t-1) = \frac{\exp(w_k(t-1)/\tau)}{\sum_j^J \exp(w_j(t-1)/\tau)} \tag{14}$$

where, $w_k(t)$ denotes the relative decay rate of task k at the t-th training stage, $\lambda_k(t)$ represents the weight value assigned to task k at stage t, and $L_k(t)$ signifies the loss incurred by task k at stage t. J signifies the total number of tasks subject to adjustment, while τ is a temperature coefficient that governs the magnitude of weight updates, with smaller values indicating greater weight update amplitude. All tasks under consideration are initially assigned equal weights during the model's initialization phase. Subsequently, their actual loss values at the first training stage, $L_k(1)$, serve as respective baseline loss references.

4 Experiments

4.1 Datasets

To evaluate the performance of the DMMSA model, we selected three representative MSA datasets: CH-SIMS (Yu et al., 2020), MOSI (Zadeh et al., 2016), and MOSEI (Zadeh et al., 2018). CH-SIMS, a resource for MSA in Chinese, comprises 2,281 video samples, with sentiment labels expressed as scores within the continuous interval [-1, +1]. MOSI, an English dataset, includes 2,199 video clips and employs a [-3, +3] sentimental intensity rating system.

Model	Acc-3(↑)	Acc-5(↑)	MAE(↓)	Corr(↑)
LF-DNN	66.91	41.62	0.420	0.612
MFN(A)	65.73	39.47	0.435	0.582
LMF	64.68	40.53	0.441	0.576
TFN	65.12	39.30	0.432	0.591
Mult(A)	64.77	37.94	0.453	0.561
Self-MM	64.73	43.15	0.414	0.598
ConFEDE	68.36	43.72	0.3924	0.6351
DMMSA	**69.63**	**46.92**	**0.3778**	**0.66**

Table 2: Results of the Comparative Experiments on the CH-SIMS Dataset.

MOSEI, an extended English MSA collection derived from MOSI, significantly expands the scale to 22,856 video segments, maintaining the [-3, +3] sentiment scoring range. The specific details of the dataset division are presented in Table 1.

4.2 Baseline Models and Evaluation Metrics

We compared our method with LF-DNN (Yu et al., 2020), MFN (Zadeh et al., 2018), LMF (Liu Z, 2018), TFN (Zadeh et al., 2017), MulT(A) (Tsai et al., 2019), self-MM (Yu et al., 2021), MISA(A) (Hazarika et al., 2020), MAG-BERT (Rahman et al., 2020), Self-MM (Yu et al., 2021), and ConFEDE (Yang et al., 2023).

We report the model's performance on classification and regression tasks following prior work. For classification, we compute theaccuracy of 3-class prediction (Acc-3) and 5-class prediction (Acc-5) on CH-SIMS, as well as theaccuracy of 2-class prediction (Acc-2) and 7-class prediction (Acc-7) on MOSI and MOSEI. Here, Acc-2 and F1-score for MOSI and MOSEI are reported in two forms: "negative/non-negative" and "negative/positive" (excluding 0). We present Mean Absolute Error (MAE) and Pearson correlation (Corr) regarding regression. All metrics except MAE are better when higher.

4.3 Controlled Experiment

Tables 2 and 3 summarize the performance comparison of various methods. The listed experimental results are based on the average of five runs with different random seeds, with the performance data for all baseline models except ConFEDE sourced from published literature.

On the CH-SIMS dataset, DMMSA demonstrates superior overall performance in classification and regression tasks compared to all baseline models. Relative to the baseline model ConFEDE, we achieve increases of 1.27% in Acc-3 and 3.20% in Acc-5. This phenomenon is primarily attributed to the coarse-grained sentiment analysis task integrated into DMMSA, enhancing the model's classification task performance. Moreover, DMMSA ex-

Model	MOSI					MOSEI				
	Acc-2	F1	Acc-7	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr
LF-DNN (Yu et al., 2020)	77.52/78.63	77.46/78.63	34.52	0.955	0.658	80.60/82.74	80.85/82.52	50.83	0.58	0.709
MFN(A) (Zadeh et al., 2018)	77.4/-	77.3/-	34.1	0.965	0.632	78.94/82.86	79.55/82.85	51.53	0.573	0.718
LMF (Baltrušaitis et al., 2018)	-/82.5	-/82.4	33.2	0.917	0.695	80.54/83.48	80.94/83.36	51.59	0.576	0.717
TFN (Zadeh et al., 2017)	-/80.8	-/80.7	34.9	0.901	0.698	78.50/81.89	78.96/81.74	51.60	0.573	0.714
MulT(A) (Tsai et al., 2019)	-/83.0	-/82.8	40.0	0.871	0.698	81.15/84.63	81.56/84.52	52.84	0.559	0.733
MISA(A) (Hazarika et al., 2020)	81.8/83.4	81.7/83.6	42.3	0.783	0.776	83.6/85.5	83.8/85.3	52.2	0.555	0.756
MAG-BERT (Rahman et al., 2020)	82.13/83.54	81.12/83.58	41.43	0.790	0.766	79.86/86.86	80.47/83.88	50.41	0.583	0.741
ConFEDE (Yang et al., 2023)*	83.85/85.55	83.83/**85.76**	43.82	0.725	0.789	80.7/84.38	81.2/84.32	51.96	0.555	0.753
DMMSA*	**83.97/85.70**	**83.92**/85.70	**45.39**	**0.710**	**0.793**	**82.63/86.27**	**83.04/86.21**	**53.91**	**0.527**	**0.777**

Table 3: Comparison experiment results on MOSI and MOSEI. In Acc-2 and F1, the left side of "/" represents "negative/non-negative", and the right side represents "negative/positive".

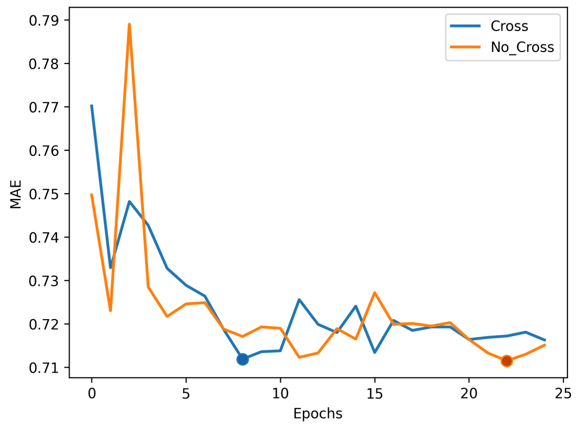

Figure 4: The evolution of MAE over epochs for different optimization strategies on the MOSI validation dataset. The "Cross" represents DMMSA incorporating the coarse-grained sentiment analysis task, whereas the "Non_Cross" corresponds to DMMSA with the coarse-grained sentiment analysis task removed. Circles mark the lowest MAE values.

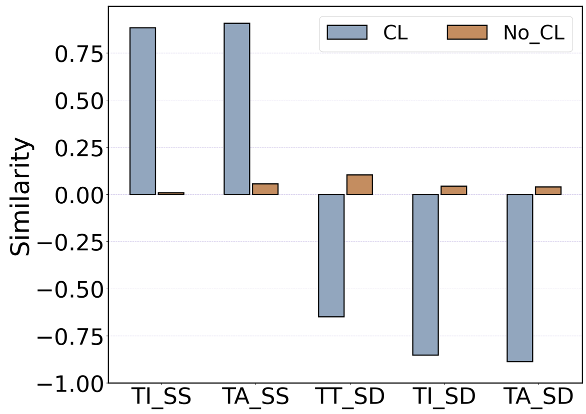

Figure 5: The similarity between similar and dissimilarity features. "CL" denotes the similarity of a model incorporating the contrastive learning task. "No_CL" represents the similarity of a model removing the contrastive learning task. "T", "I", and "A" respectively denote text, image, and audio modalities, with "S" and "D" signifying similarity features and dissimilarity features, respectively.

hibits notable advancements in MAE and Corr metrics. This result is because it incorporates uni and multimodal sentiment analysis to capture their interdependencies effectively, and it embeds a coarse-grained sentiment analysis task that contributes to constraining the sentimental prediction scope and simplifying the sentiment analysis task. As illustrated in Figure 4, DMMSA exhibits more minor MAE fluctuations and faster convergence during training compared to the model without the inclusion of coarse-grained sentiment analysis, further substantiating the positive role of this task in model optimization.

To further validate the efficacy of our proposed approach, we conducted experiments on the MOSI and MOSEI datasets lacking unimodal sentiment labels. Table 3 presents the results. DMMSA outperforms all baseline models on both datasets, demonstrating its exceptional performance even when unimodal sentiment labels are unavailable. This phenomenon is mainly due to the design of the contrastive learning task. As shown in Figure 5, even when unimodal feature labels are missing, the model can still be guided by the contrastive learning task to identify and separate uni and multimodal features effectively.

Of particular concern is that DMMSA's improvement in Acc-5, MAE, and Correlation metrics exceeds its improvement in Acc-2 and Acc-3. This phenomenon stems from the higher requirements for model performance and feature quality in complex tasks compared to simple tasks; simple tasks often only require lower-level features to achieve good performance, while the advantage of DMMSA lies in extracting higher-quality features, so its performance gain is more significant when task difficulty increases.

To confirm this hypothesis, we designed an incremental experiment; Table 4 presents the results. We can observe that the performance of DMMSA on Acc-2 has reached convergence when trained with 60% data, and its performance will not improve with the increase of training data. On the contrary, the performance of DMMSA on Acc-7 and regression tasks continuously improves with the increase of training data.

Data	Acc-2	F1	Acc-7	MAE	Corr
MOSEI	82.63/**86.27**	83.04/**86.21**	**53.91**	**0.527**	**0.777**
MOSEI*0.8	82.41/86.15	82.87/86.14	53.39	0.532	0.772
MOSEI*0.6	**83.77**/86.12	**84.01**/85.99	52.77	0.538	0.769
MOSEI*0.4	82.78/86.09	83.16/86.03	53.36	0.540	0.767
MOSEI*0.2	80.68/85.03	81.27/85.06	52.37	0.551	0.760
MOSEI*0.1	81.89/85.08	82.33/85.04	52.18	0.552	0.755

Table 4: Performance of DMMSA under varying amounts of training data.

4.4 Ablation Study and Analysis

We conducted an ablation study on the proposed method to investigate the individual contributions of each module to model performance. Table 5 shows the results.

We can observe that the model's performance decreases to varying degrees under the three ablation strategies. "w/o MSC" exhibits decreases across all performance metrics. The decline can primarily be attributed to the loss of practical constraints on the sentimental prediction range after removing the MSC task. As Figure 6 illustrates, when the MSC task is incorporated, the model performs a preliminary prediction of the sample's sentiments, which confines its regression prediction search space (as denoted by "MSC" in Figure 6). For instance, if the model preliminarily assigns the sample to the "Negative" sentiment region, it restricts subsequent predictions to occur only within this area, thereby preventing excessive divergence from the actual sentimental state.

Model	Acc-3	Acc-5	MAE	Corr
DMMSA	**69.63**	**46.92**	0.3778	0.66
w/o MSC	68.41	44.68	0.3807	0.656
w/o CL	69.41	46.74	0.3828	0.651
w/o GDWG	69.32	46.17	**0.3776**	**0.663**

Table 5: The ablation experiments on CH-SIMS. "w/o CL" signifies the exclusion of the contrastive learning(CL) task.

Under the "w/o CL" configuration, the model's MAE and Corr indicators significantly decreased. This result is mainly because the core objective of CL tasks is to assist the model in effectively distinguishing and extracting similar and dissimilar features from single modalities. Once the CL task is removed, the model loses the feature discrimination ability promoted by this mechanism, making it difficult to accurately distinguish and utilize these critical sentimental features. So, it weakens its performance in regression tasks.

In the setting "w/o GDWG," the model exhibits a mild upward trend in performance on regression tasks, whereas a marked decline is observed in its efficacy on classification tasks. The underlying cause of this phenomenon lies in the model's loss

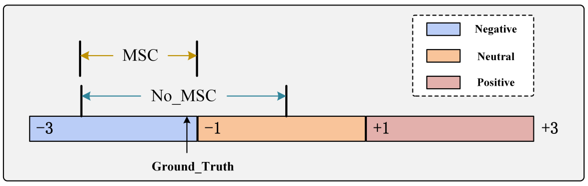

Figure 6: Visualizing sentiment intensity range, with "MSC" denoting sentiment intensity prediction scope when incorporating MSC tasks and "No_MSC" indicating the scope without it.

of the effective regulatory mechanism for gradient convergence rates and magnitude differences among various tasks during the training cycle. As a result, the model tends to over-optimize a single task at the expense of neglecting the learning requirements of other tasks, culminating in an evident imbalance in overall performance. The core function of the GDWG module resides in its ability to dynamically adjust the weight allocation for each task based on the real-time global descent rate of respective task loss functions. It prevents the model from overly concentrating on any particular task to the detriment of the learning progress of other tasks, thereby effectively mitigating learning skew arising from inter-task competition.

5 Conclusion

This study introduces DMMSA, an affective analysis framework that integrates multi-task learning strategies with dynamic tuning mechanisms to enhance the accuracy of modeled understanding of complex human sentiments by exploiting intrinsic correlations between uni-modal and multimodal sentimental signals. Specifically, DMMSA systematically extracts and decomposes sentiment representations from multimodal inputs into similarity and dissimilarity components, which are then deepened through coarse-grained sentiment classification tasks and contrastive learning mechanisms acting on the interplay of sentimental representations. To comprehensively validate the DMMSA, we evaluate it on three representative MSA datasets: CH-SIMS, MOSI, and MOSEI. Experimental results demonstrate that DMMSA surpasses various benchmark models across all overall performance metrics on all datasets. Moreover, through a series of ablation experiments, we further substantiate the indispensable contribution of each constituent module within DMMSA to the overall performance improvement, thereby affirming this design's methodological soundness and effectiveness.

6 Limitation

Although we have alleviated the negative transfer effects caused by differences in task convergence rates using the Global Dynamic Weight Generation (GDWG) strategy, this problem still exists and becomes a key factor restricting the performance improvement of DMMSA. Table 4 shows that as the training sample size increases, the performance of DMMSA on Acc-2 decreases, while on Acc-5, MAE, and Correlation indicators, it shows an upward trend. Therefore, the focus of subsequent research will be exploring the optimization path of GDWG to suppress negative transfer more effectively.

References

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221.

Ajwa Aslam, Allah Bux Sargano, and Zulfiqar Habib. 2023. Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks. *Applied Soft Computing*, page 110494.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Jun Du, Jianhang Jin, Jian Zhuang, and Cheng Zhang. 2024. Hierarchical graph contrastive learning of local and global presentation for multimodal sentiment analysis. *Scientific Reports*, 14(1):5335.

Wenjun Feng, Xin Wang, Donglin Cao, and Dazhen Lin. 2024. An autoencoder-based self-supervised learning for multimodal sentiment analysis. *Information Sciences*, page 120682.

Yao Fu, Biao Huang, Yujun Wen, and Pengzhou Zhang. 2024. Fdr-msa: Enhancing multimodal sentiment analysis through feature disentanglement and reconstruction. *Knowledge-Based Systems*, page 111965.

Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. 2020. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33:22605–22618.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

Jian Huang, Yuanyuan Pu, Dongming Zhou, Jinde Cao, Jinjing Gu, Zhengpeng Zhao, and Dan Xu. 2024. Dynamic hypergraph convolutional network for multimodal sentiment analysis. *Neurocomputing*, 565:126992.

Xun Jiang, Xing Xu, Huimin Lu, Lianghua He, and Heng Tao Shen. 2024. Joint objective and subjective fuzziness denoising for multimodal sentiment analysis. *IEEE Transactions on Fuzzy Systems*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+language omni-representation pretraining. *Preprint*, arXiv:2005.00200.

Yi Liang, Turdi Tohti, and Askar Hamdulla. 2022. Multimodal false information detection method based on text-cnn and se module. *Plos one*, 17(11):e0277463.

Shengchao Liu, Yingyu Liang, and Anthony Gitter. 2019. Loss-balanced task weighting to reduce negative transfer in multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*.

Zhizhong Liu, Bin Zhou, Dianhui Chu, Yuhang Sun, and Lingqiang Meng. 2024. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 101:101973.

Lakshminarasimhan V B et al Liu Z, Shen Y. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *arXiv preprint arXiv:1806.00064*.

Qiang Lu, Xia Sun, Zhizezhang Gao, Yunfei Long, Jun Feng, and Hao Zhang. 2024. Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis. *Information Processing & Management*, 61(1):103538.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Hang Shi, Yuanyuan Pu, Zhengpeng Zhao, Jian Huang, Dongming Zhou, Dan Xu, and Jinde Cao. 2024. Co-space representation interaction network for multimodal sentiment analysis. *Knowledge-Based Systems*, 283:111149.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999.

Quoc-Tuan Truong and Hady W Lauw. 2019. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 305–312.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Dingkang Yang, Mingcheng Li, Dongling Xiao, Yang Liu, Kun Yang, Zhaoyu Chen, Yuzheng Wang, Peng Zhai, Ke Li, and Lihua Zhang. 2024. Towards multimodal sentiment analysis debiasing via bias purification. *arXiv preprint arXiv:2403.05023*.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Ying Zeng, Wenjun Yan, Sijie Mai, and Haifeng Hu. 2024. Disentanglement translation network for multimodal sentiment analysis. *Information Fusion*, 102:102031.

A Baseline

LF-DNN: Concatenating unimodal features and analyzing sentiments. (Yu et al., 2020)

MFN: Firstly, employ LSTM for View-specific interaction. Then, utilize the attention mechanism for Cross-view interaction, and finally summarize through time with a Multi-view Gated Memory. (Zadeh et al., 2018)

LMF: By parallelly decomposing tensors and weights, utilize modality-specific low-rank factors to perform multimodal fusion. (Liu Z, 2018)

TFN: The authors propose a novel model called Tensor Fusion Network (TFN), which can learn end-to-end dynamics within and across modalities. It adopts a new multimodal fusion method (tensor fusion) to model the dynamics across modalities. (Zadeh et al., 2017)

MulT: The core of MulT lies in its cross-modal attention mechanism, which offers a potential cross-modal adaptation by directly attending to low-level features in other modalities to fuse multimodal information. (Tsai et al., 2019)

MISA: The model learns Modality-Invariant and Modality-Specific representation spaces for each

	Encoder(T)	Bert-base(Chinese/uncased)
	Encoder(V/A)	Transformer Encoder[1]
Single-modal	Learning rate	0.00001(T),0.0001(V/A)
	Batch size	64(T),128(V/A)
	Epochs	150(T),300(V/A)
	Epochs	50(CH,SI),25(SEI)
Fusion stage	Learning rate	0.00001(CH,SI),0.00005(SEI)
	Batch size	32(CH),16(MOSI),4(MOSEI)

Table 6: Parameter settings for the single-modal stage. T represents text, V represents visual, and A represents audio.

	2-class	3-class	5-class
	[-1,0]	[-1,-0.1]	[-1,-0.7]
Sentiment	(0,1]	(-0.1,0.1]	(-0.7,-0.1]
Intensity		(0.1,1]	(-0.1,0.1]
			(0.1,0.7]
			(0.7,1.0]

Table 7: Classification Label Division Method of the CH-SIMS Dataset

modality to obtain better modality representations for the fused input. (Hazarika et al., 2020)

MAG-BERT: Enhancing model performance by applying multimodal adaptation gates at different layers of the BERT backbone. (Rahman et al., 2020)

Self-MM: First, utilize a self-supervised label generation module to obtain unimodal labels, then jointly learn multimodal and unimodal representations based on multimodal labels. (Yu et al., 2021)

ConFEDE: Firstly, decompose unimodal features into Modality-Invariant and Modality-Specific features through feature decomposition. Subsequently, utilize multi-task learning to combine multimodal sentiment analysis, unimodal sentiment analysis, and contrastive learning tasks to optimize model training. (Yang et al., 2023)

B Experiment

B.1 Experiment Setting

All experiments were conducted on an NVIDIA Tesla A100 GPU. Our remaining experimental settings were consistent with the previous state-of-the-art model, ConFEDE. Table 6 presents the parameter settings for the unimodal training and multimodal fusion stages.

B.2 Methods for Multimodal Sentiment Classification Labeling

CH-SIMS: For this dataset, we have defined 2-class, 3-class, and 5-class classification tasks representing three different difficulty levels. The specific divisions are shown in the table 7:

MOSI and **MOSEI**: For these two datasets, we have defined 2-class, 3-class, 5-class, and 7-class classification tasks. The specific division details are shown in the table 8.

	2-class	3-class	5-class	7-class
	[-3,0)	[-3,-0.5)	[-3,-1.5)	[-3,-2.5)
	[0,3]	[-0.5,0.5)	[-1.5,-0.5)	[-2.5,-1.5)
		[0.5,3]	[-0.5,0.5)	[-1.5,-0.5)
Sentiment			[0.5,1.5)	[-0.5,0.5)
Intensity			[1.5,3]	[0.5,1.5)
				[1.5,2.5)
				[2.5,3]

Table 8: Classification Label Division Method of the MOSE and MOSEI Dataset