Cross-lingual Multimodal Sentiment Analysis for Low-Resource Languages via Language Family Disentanglement and Rethinking Transfer

Long Chen [1†], Shuoyu Guan [1†*], Xiaohua Huang[2], Wen-Jing Wang[1], Cai Xu[3], Ziyu Guan[3], Wei Zhao[3]

[1]Shaanxi Key Laboratory of Information Communication Network and Security,
Xi'an University of Posts and Telecommunications, Xi'an, China
[2]School of Information Science and Technology, Northwest University, Xi'an, China
[3]School of Computer Science and Technology, Xidian University, Xi'an, China

Correspondence: {chenlong@, shuoyuguan@stu., wjing@}xupt.edu.cn,
huangxiaohua@stumail.nwu.edu.cn, {cxu@, zyguan@, ywzhao@mail.}xidian.edu.cn

Abstract

Existing multimodal sentiment analysis (MSA) methods have achieved significant success, leveraging cross-modal large-scale models (LLMs) and extensive pre-training data. However, these methods struggle to handle MSA tasks in low-resource languages. While multilingual LLMs enable cross-lingual transfer, they are limited to textual data and cannot address multimodal scenarios. To achieve MSA in low-resource languages, we propose a novel transfer learning framework named *Language Family Disentanglement and Rethinking Transfer (LFD-RT)*. During pre-training, we establish cross-lingual and cross-modal alignments, followed by a *language family disentanglement* module that enhances the sharing of language universals within families while reducing noise from cross-family alignments. We propose a *rethinking strategy* for unsupervised fine-tuning that adapts the pre-trained model to MSA tasks in low-resource languages. Experimental results demonstrate the superiority of our method and its strong language-transfer capability on target low-resource languages. Code and models are available at https://github.com/ShuoyuGuan/LFD-RT.

1 Introduction

Sentiment analysis aims to detect the sentiment orientations of opinion data. It can be applied to various domains, including market analysis, social media monitoring, and finance and investment. Current advanced social platforms offer diverse multimodal opinion data, thereby extending sentiment analysis beyond its traditional reliance on textual data to encompass multimodal sentiment analysis (MSA). Early research in multimodal sentiment analysis primarily centers on the fusion of multimodal representations, with the progression ranging from basic feature-level fusion (Poria et al., 2018; Zadeh et al., 2018) to more sophisticated decision-level (Chen and Yang, 2020) and hybrid approaches (Han et al., 2021). Later, the research roadmap was dispersed into various fine-grained tasks such as multimodal pre-training models (Lu et al., 2019), modality-invariant representation learning (Hazarika et al., 2020), sentiment analysis under missing uncertain modalities (Zeng et al., 2022) and few-shot learning (Yang et al., 2023). However, none of these works can be utilized for multimodal sentiment analysis in low-resource languages. Multilingual models can handle this problem. Early studies focused on leveraging multilingual LLMs (e.g., mBERT (Pires, 2019), XLM (Conneau and Lample, 2019)) and adversarial training (Dong et al., 2020) for cross-lingual transfer. Subsequently, multi-view-based methods (Fei and Li, 2020) and contrastive learning(Lin et al., 2023) have emerged for MSA. However, these multilingual methods were designed solely for textual data. Thakkar et al. (Thakkar et al., 2024) first attempted to conduct the multimodal and multilingual sentiment analysis task. Nonetheless, the proposed model is overly simplistic and fails to account for the cross-linguistic alignment and cross-lingual noise resistance.

In summary, the key problem we aim to address in this paper is that current multimodal methods struggle to adapt to data in low-resource languages due to their dependence on large-scale training data that is typically available from high-resource languages. Multilingual models can address the challenge of data scarcity; however, they are primarily applied to single-modal textual data.

Inspired by the disentanglement technique in (Ge et al., 2024) and the language family concept in linguistics, we propose a transfer learning framework called Language Family Disentanglement and Rethinking Transfer. During the pre-training phase, we conduct cross-lingual and cross-modal alignments. For aligned cross-lingual features, we

† Equal contribution
* Corresponding author

Findings of the Association for Computational Linguistics: ACL 2025, pages 6513–6522
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

develop a module for language family disentanglement to share language universals within the same family, while reducing the noise generated by cross-family alignments. Then, we use a fusion operation to perform a linguistic-to-modal transfer. We further establish a masked language task to recalibrate the fused representation and conduct the sentiment semantic learning task using double classifiers. During the fine-tuning stage, we propose a novel rethinking strategy to achieve unsupervised fine-tuning of the pre-trained classifiers. The two classifiers conduct a double check on the predictions from the input data. If they reach an agreement, the predicted label serves as a pseudo-label for prediction and backpropagation. If there is inconsistency, the data are sent back as input to conduct a masked prediction task and update the model parameters. In this manner, we finally achieve unsupervised fine-tuning, enabling the pre-trained model to adapt to downstream MSA in low-resource languages. The experimental results demonstrate the efficacy of our methodology and its robust cross-linguistic transferability to low-resource target languages.

2 Related Works

As most of the existing work has been reviewed in the Introduction, we briefly supplement some of the latest representative research.

Multimodal Sentiment Analysis. Li et al. (Li et al., 2024) introduced sample-level contrast and category-guided prototype extraction to handle uncertain missing modalities. Zhu et al. (Zhu et al., 2024) fused video information into text semantics via a text-based cross-modal approach, while Zhao et al. (Zhao et al., 2025) utilized pseudo-data generation for MSA.

Multilingual Sentiment Analysis. Kanayama et al. (Kanayama et al., 2024) enhanced generative methods by incorporating sentiment extractors and external polarity data. Miah et al. (Miah et al., 2024) performed sentiment analysis by translating foreign languages into English.

3 Methodology

3.1 Problem Definition

Given a text-image tuple (T, I), our goal is to develop a model that can accurately predict the sentiment label y. This model supports diverse sentiment classification tasks, including but not limited to the Ekman Emotion Classification Task (Anger, Disgust, Fear, Joy, Sadness, Surprise, Neutral).

3.2 Overall Framework

As shown in Figure 1, we employ a pre-training and unsupervised fine-tuning paradigm as our transfer learning framework. In the pre-training phase, we first perform *cross-linguistic and cross-modal alignments*. Then, we use *language family disentanglement* to mitigate noises induced by linguistic characteristics in cross-lingual alignment, such as lexical, syntactic, and morphological differences. Finally, we train multiple sentiment classifiers on aligned disentangled representations to capture sentiment semantics. After pre-training, the entire pre-trained model is transferred. During the unsupervised fine-tuning stage, we design a *rethinking fine-tuning* strategy to adapt the multiple sentiment classifiers without supervised signals while keeping the other components of the pre-trained model frozen.

3.3 Pre-training

We utilize two data sets for pre-training. One set, denoted by $D_{cm} = \{T_i, I_i, \mathbf{y}_i\}_{1 \le i \le N}$, is employed for cross-modal process. Here, T_i represents an opinion text in the anchor language (a high-resource language, such as English), I_i denotes its corresponding image, and $\mathbf{y}_i$ is the ground truth label for (T_i, I_i). Another set, denoted as $D_{cl} = \{\hat{T}_j, T_j^{ts}\}_{1 \le j \le M}$, is used for the cross-lingual computations. In this set, $\hat{T}_j$ is the anchor text, and T_j^{ts} represents its translation in another low-resource language. Note that T_i and $\hat{T}_j$ are different; the former are opinion texts while the latter is not.

Cross-lingual Alignment. This component is designed for cross-lingual transfer. We employ a pre-trained Multilingual BERT (mBERT) as the encoder:

$$\begin{aligned} \mathbf{x}_j^{an} &= mBERT(\hat{T}_j) \\ \mathbf{x}_j^{ts} &= mBERT(T_j^{ts}) \end{aligned} \tag{1}$$

A contrastive learning strategy is then applied to the feature vectors, enabling the model to explore the similarities and differences between texts in diverse languages, thereby achieving cross-lingual alignment. Considering that cross-lingual knowledge is transferred to cross-modal representations in the following module, we employ unified contrastive

https://huggingface.co/google-bert/bert-base-multilingual-cased

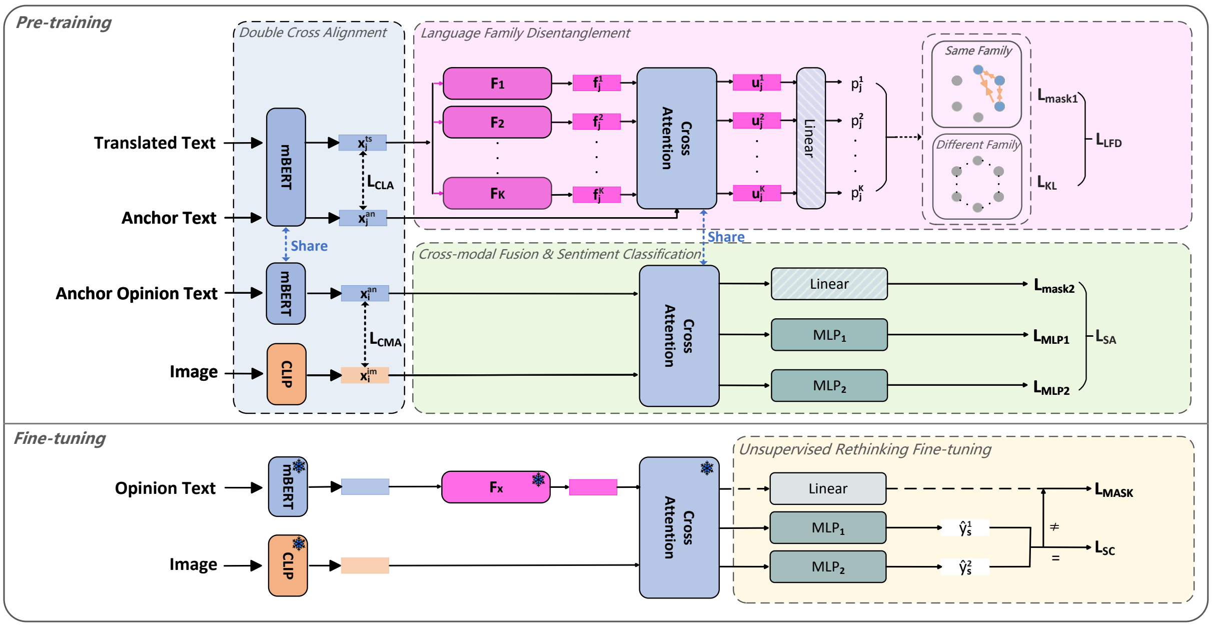

Figure 1: The overall framework.

loss (Li et al., 2023b) for cross-lingual alignment because it can simultaneously align cross-lingual texts and image-text pairs. The loss function is:

$$\begin{aligned} \mathcal{L}_{\text{CLA}} &= \mathcal{L}_{ucl}(\mathbf{x}_j^{an}, \mathbf{x}_j^{ts}) = \\ -\mathbb{E}_{\sim D_{cl}}&[\log \frac{\exp(s(\mathbf{x}_j^{an}, \mathbf{x}_j^{ts}))/\tau)}{\sum_{j=1}^{M} \exp(s(\mathbf{x}_j^{an}, \mathbf{x}_{\neg j}^{ts})}) / \tau) \\ &+ \log \frac{\exp(s(\mathbf{x}_j^{an}, \mathbf{x}_j^{ts}))/\tau)}{\sum_{j=1}^{M} \exp(s(\mathbf{x}_j^{an}, \mathbf{x}_{\neg j}^{ts})/\tau)}] \end{aligned} \quad (2)$$

where τ is a temperature hyperparameter.

Cross-modal Alignment. This component performs a cross-modal alignment using a method similar to the cross-lingual alignment:

$$\begin{aligned} \mathbf{x}_i^{an} &= mBERT(T_i) \\ \mathbf{x}_i^{im} &= CLIP(I_i) \end{aligned} \quad (3)$$

$$\mathcal{L}_{\text{CMA}} = \mathcal{L}_{ucl}(\mathbf{x}_i^{an}, \mathbf{x}_i^{im}) \quad (4)$$

The only difference is that we use a CLIP (Radford et al., 2021) as the encoder for the input images.

Language Family Disentanglement. After cross-lingual alignment, the aligned features are fed into a language family disentanglement module. We use this module to mitigate the negative impacts of noises induced by linguistic characteristics while maximizing the retention of beneficial transfer knowledge. Cross-lingual transfer learning involves leveraging the structural, syntactic, semantic, and cultural similarities and differences between languages. Closely related languages, such as English and Frisian, exhibit significant overlaps in grammatical structures, vocabulary semantics, and linguistic features. For instance, both languages share similar noun inflections for gender, number, and case, as well as comparable verb tense and voice systems. These shared linguistic properties can serve as inductive biases, facilitating more effective cross-lingual knowledge transfer by reducing the domain gap between source and target languages. However, for distantly related languages, such as English and Chinese, these similarities diminish, and differences become more pronounced, introducing noise that can hinder transfer performance. The challenge, therefore, lies in designing models that can effectively capture and preserve shared linguistic universals while minimizing the interference caused by language-specific variations.

Coincidentally, language families serve as a robust indicator of linguistic similarities and differences. Languages within the same family share common origins and exhibit similar syntactic, morphological, and phonological patterns, and vice versa. The concept of a language family offers a valuable strategy, enabling the retention of relevant features within the same family while filtering out noise from different families. To this end, we design a language detangler as shown in Figure 1. We first use 13 family disentanglers $f_{k:1\leq k\leq K}$: $\mathbf{x}_j^{ts} \rightarrow \mathbf{f}_j^k$ to generate the family-related features . The disentangler function is:

$$\mathbf{f}_j^k = \mathbf{W}^k \mathbf{x}_j^{ts} \tag{5}$$

where $\mathbf{W}^{k:1\leq k\leq K}$ are parameter matrices for K family disentanglers. Then, family-related features $\mathbf{F}_j = [\mathbf{f}_j^1 \; \mathbf{f}_j^2 \; \cdots \; \mathbf{f}_j^K]$ are fed into the cross-attention mechanism to calculate the correlations with features of the aligned anchor language features as follows:

$$\begin{aligned}\mathbf{Q}_j &= \mathbf{W}_q \mathbf{F}_j \\ \mathbf{K}_j &= \mathbf{W}_k \mathbf{X}_j^{an} \\ \mathbf{V}_j &= \mathbf{W}_v \mathbf{X}_j^{an} \\ \mathbf{U}_j &= \mathrm{softmax}\left(\frac{\mathbf{Q}_j \mathbf{K}_j^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{V}_j \end{aligned} \tag{6}$$

where $\mathbf{W}_q$,$\mathbf{W}_k$ and $\mathbf{W}_v$ are parameter matrices. $\mathbf{U} = [\mathbf{u}_j^1 \; \mathbf{u}_j^2 \; \cdots \; \mathbf{u}_j^K]$ are anchor-correlation features, and we input them into linear projection to obtain the final predictions $\mathbf{P}_j = [\mathbf{p}_j^1 \; \mathbf{p}_j^2 \; \cdots \; \mathbf{p}_j^K]$ as following:

$$\mathbf{P}_j = \mathbf{W}_l \mathbf{U}_j \tag{7}$$

We implement a masked language task to ensure that samples within the same language family (i.e., anchor family) are more closely grouped, coupled with a KL divergence regularization task to promote a uniform distribution between samples from different families. The loss function for the masked language task is:

$$\begin{aligned}&\mathcal{L}_{mask_1} = \mathcal{L}_{MASK}(T_j^{tr}, \mathbf{p}_j) = \\ &-\frac{1}{M}\sum_{j=1}^{M}\sum_{p=1}^{r|T_j^{tr}|} \mathbf{w}_p^{T_j^{tr}} \log\left(\mathbf{p}_{jp}^{k=anchor}\right)\end{aligned} \tag{8}$$

where $\mathbf{w}_p^{T_j^{tr}}$ represents the vocabulary probability distribution for masked words in position p, and $\mathbf{p}_{jp}^{k=anchor}$ signifies that the predicted word probability in position p of the translated text belongs to the same family as the anchor language. r represents the masking probability, and $|T_j^{tr}|$ denotes the total number of words in the text T_j^{tr}. We employ the same masking strategy as BERT (Kenton and Toutanova, 2019). The KL divergence regularization loss is:

$$\mathcal{L}_{KL} = \frac{1}{M}\sum_{j=1}^{M} P_j^{k\neq anchor} \log \frac{P_j^{k\neq anchor}}{P_{uniform}} \tag{9}$$

where $P_j^{k\neq anchor}$ denotes the probability distribution of translated texts originating from language families distinct from that of the anchor language, and P_{uni} signifies a uniform distribution. We regulate the samples from other language families to achieve a uniform distribution with a high degree of chaos, indicating significant information disorder. This is in line with the objective of alleviating the noise effects of other language families. Combining the two tasks mentioned above, the language family disentanglement loss function is:

$$\mathcal{L}_{LFD} = (\alpha_1 \mathcal{L}_{mask_1} + \alpha_2 \mathcal{L}_{KL}) \tag{10}$$

where α_1 and α_2 are hyperparameters.

Cross-modal Fusion. In the cross-modal interaction pipeline, we employ a cross-attention mechanism, enhanced through language family disentanglement, to process aligned multimodal representations. This approach promotes knowledge transfer from linguistic to visual modalities, while enabling cross-modal fusion. The computations are:

$$\begin{aligned}\mathbf{q}_i &= \mathbf{W}_q \mathbf{x}_i^{im} \\ \mathbf{k}_i &= \mathbf{W}_k \mathbf{x}_i^{an} \\ \mathbf{v}_i &= \mathbf{W}_v \mathbf{x}_i^{an} \\ \mathbf{z}_i &= \mathrm{softmax}\left(\frac{\mathbf{q}_i \mathbf{k}_i^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{v}_i\end{aligned} \tag{11}$$

Note that the parameter matrices $\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v$ are shared by the cross-attention of the language family disentanglement module. Next, we fed fused representation into two multilayer perceptions (MLP_1 and MLP_2) for predicting the sentiment orientations:

$$\mathbf{o}_i^{1\vee 2} = MLP_{1\vee 2}(\mathbf{z}_i) \tag{12}$$

$$\begin{aligned}\mathcal{L}_{MLP_{1\vee 2}} &= \mathcal{L}_{CE}(\mathbf{y}_i, \mathbf{o}_i^{1\vee 2}) \\ &= -\frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i \log\left(\mathbf{o}_i^{1\vee 2}\right)\end{aligned} \tag{13}$$

Both perceptions have one hidden layer, utilize ReLU as the activation function, and share the same cross-entropy loss function. To mitigate the semantic shift caused by the fusion operation, we also incorporate a masked language loss to recalibrate the semantics of the fused representation. We employ another projection layer along with the same masked language strategy used in the language family disentanglement module. The computational equations are:

$$\mathbf{P}_i = \mathbf{W}_{1'} \mathbf{Z}_i \tag{14}$$

K = 13 is the total number of language families covering all the languages in our dataset.

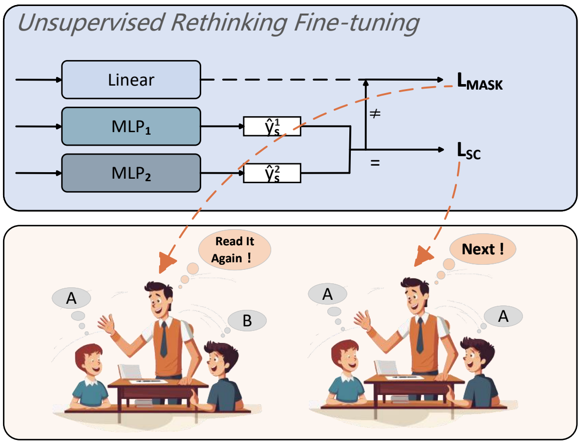

Figure 2: An intuitive explanation of rethinking unsupervised fine-tuning.

$$\mathcal{L}_{mask_2} = \mathcal{L}_{MASK}(T_i, \mathbf{p}_i) \tag{15}$$

Thus, the combined loss function for sentiment analysis is:

$$\mathcal{L}_{SA} = (\beta_1 \mathcal{L}_{MLP_1} + \beta_2 \mathcal{L}_{MLP_2} + \beta_3 \mathcal{L}_{mask_2}) \tag{16}$$

where β_1 and β_2 are weight coefficients. By combining the loss functions from the above modules, we arrive at the final optimization loss for the pre-training part:

$$Loss_{PT} = (\mathcal{L}_{\text{Align}} + \mathcal{L}_{FCD} + \mathcal{L}_{SA}) \tag{17}$$

Note that we only use the labeled multimodal data in the high-resource language for pre-training.

3.4 Unsupervised Rethinking Fine-tuning

Table 1: MSA data statistics

Meld			**CH SIMS**		
Train	**Val**	**Test**	**Train**	**Val**	**Test**
5548	2710	2511	1462	665	655

Dataset	**Languages**				
Meld	**en**	eo	es	et	eu
	fi	fr	he	hu	id
	io	ka	ar	mg	ml
	oc	sk	sw	ta	te
	tg	tl	tr		
CH SIMS	**zh**	ja	jv	ko	tt

Table 2: Languages in Meld and CH SIMS dataset.

Fine-tuning is a crucial step in adapting the pre-trained model for low-resource languages. In real-world scenarios, we commonly encounter the challenge of limited data resources in minority languages. To address this challenge, we design an unsupervised learning method called rethinking fine-tuning. This method draws inspiration from the answer consistency check with multiperson cross-validation, as illustrated in Figure 2. Specifically, during the fine-tuning phase, the pre-trained models are used as initialization. We fine-tune the two MLPs and fixed the other components of the model. The dataset for fine-tuning is denoted as $D_{low} = \{T_s^{low}, I_s^{low}\}_{1 \leq s \leq S}$, where T_s^{low} is the opinion text in the target low-resource language, and its corresponding image is represented as I_s^{low}. The forward computation process is identical to that of the pre-training (note that input text is processed by the disentangler of its corresponding language family, i.e., F_x). For the two pre-trained MLP classifiers, we develop an unsupervised rethinking strategy to tune their parameters: If both classifiers yield the same label, we deem this label a pseudo-label for backpropagation; if they produce different labels, it indicates a divergence in their understanding of the sentiment semantics of the input data. In this case, the input data pairs are returned and used as model inputs for masked language training. This approach is the same as the training method that addresses the semantic shift problem during the pretraining process. The loss functions for rethinking fine-tuning are:

$$Loss_{FT} = \begin{cases} \text{if } \hat{y}_s^1 = \hat{y}_s^2 = \hat{y}_s : \\ \quad \mathcal{L}_{SC} = \lambda_1 \mathcal{L}_{MLP_1} + \lambda_2 \mathcal{L}_{MLP_2}, \\ \text{if } \hat{y}_s^1 \neq \hat{y}_s^2 : \\ \quad \mathcal{L}_{MASK}(T_s^{low}, \mathbf{p}_s), \end{cases} \tag{18}$$

where $\mathcal{L}_{MLP_{1\vee 2}} = \mathcal{L}_{CE}(\hat{y}_s, \mathbf{o}_s^{1\vee 2})$. $\hat{y}_s^1$ and $\hat{y}_s^2$ are the predicted one-hot labels of the MLP_1 and MLP_2 respectively. $\hat{y}_s$ is the pseudo-label. The objectives of the masked language task in the pre-training and fine-tuning stages differ; the former aims to solve the semantic shift problem of fused features, whereas the latter aims to rethink the sentiment semantics of the input data. Therefore, we reconstruct a new linear projection layer to generate $\mathbf{p}_s$ for the masked language task in the fine-tuning stage. By employing a rethinking strategy, we conduct an unsupervised fine-tuning. We use double MLP checkers because the experimental results indicate that adding more checkers does not significantly enhance performance; rather, it increases the model complexity and the risk of conflicting outcomes.

3.5 Inference

During the inference phase, the final predicted label is determined through agreement between the two classifiers. In cases of disagreement, the class with the highest predicted probability is assigned as the output.

4 Experiments

4.1 Datasets and Language Families

We employ two MSA datasets with different granularities for evaluation: 1) a fine-grained 7-level Ekman MSA dataset called **Meld**. Collected from the TV series "Friends", it contains 1,433 conversations, 13,708 utterances, and 304 speakers. Each utterance is labeled as anger, disgust, sadness, joy, surprise, fear, or neutral. 2) A coarse-grained 5-level MSA dataset called **CH SIMS**: A Chinese multimodal sentiment analysis dataset with 2,281 video clips. Each utterance is labeled as positive, weakly positive, negative, weakly negative, or neutral. The video-text pairs are used for training. We adhere to the data splitting procedure outlined in Table 1. To create the validation and test sets, we translate English and Chinese texts into 22 and 4 other languages, respectively, using Google Translate. The complete set of 28 language abbreviations that follow the ISO 639 standard is presented in Table 2. The bold abbreviations are English (en) and Chinese (zh), which serve as the anchor languages. Note that our primary emphasis is on multilingual learning. Therefore, we exclude the audio data, as it pertains to a single language, and instead focus solely on texts and images.

For the cross-lingual task in the pretraining stage, we select 124,037 text pairs from **WikiMatrix** . The corpus is a large-scale multilingual parallel dataset released by Facebook, comprising 135 million parallel sentences. The collected pairs feature 28 languages, including Esperanto, Tamil, and others, with English serving as the anchor language. Similarly, we also collect 26,965 text pairs in Chinese (anchor language) and other 4 languages from **WikiMatrix**. The two gathered datasets are used for pretraining in the **Meld** and **CH SIMS** tasks. Table 3 illustrates the language family affiliations of all languages used in the experiments.

Language	Family	Language	Family
eo	-	ko	KR
es	IE	ar	AA
et	U	mg	AN
eu	-	ml	Drav.
fi	U	oc	IE
fr	IE	sk	IE
gl	IE	sw	NC
he	AA	ta	Drav.
hu	U	te	Drav.
id	AN	tg	IE
io	-	tl	AN
ja	JR	tr	Trk.
jv	AN	zh	ST
ka	K		

Table 3: Language-family affiliation.

4.2 Implementation details

We optimize the hyperparameters on the validation set through a sequential tuning strategy, where each hyperparameter is individually adjusted while the others remain fixed. The final configuration is as follows: the temperature hyperparameter in the contrastive loss is set to $\tau = 0.07$, and the masking probability is $r = 15\%$. The weighted hyperparameters of the loss functions are configured as: $\alpha_1 = 0.3$, $\alpha_2 = 0.1$, $\beta_1 = \beta_2 = 1$, $\beta_3 = 0.3$, and $\lambda_1 = \lambda_1 = 0.5$. We use the Adam optimizer with a learning rate of 5×10^{-6}.

All experiments were performed on an NVIDIA RTX 3090 GPU, which required 20 hours of training time. Our hyperparameter selection was guided by preliminary grid search and empirical validation: temperature $\tau \in (0, 0.1)$ with 0.01 increments, and task weights $\alpha_1, \alpha_2 \in (0, 1)$ with 0.05 increments. The parameter sensitivity analysis revealed an inverted U-shaped performance curve, where excessive α_1 (> 0.3) caused sentiment-task dominance degradation, while $\alpha_2 > 0.1$ induced random masking behavior. These observations align with the theoretical expectations. Optimal parameters were determined as $\tau = 0.07$, $\alpha_1 = 0.3$, $\alpha_2 = 0.1$, with the masking ratio fixed at BERT's standard 15%.

4.3 Baselines

Large language models (LLMs) have exhibited remarkable proficiency in MSA tasks, prompting us to adopt the top-performing LLMs as our baseline methodologies. The baselines can be categorized into three groups: 1) multilingual baselines,

https://github.com/facebookresearch/LASER/tree /main/tasks/WikiMatrix

https://www.ethnologue.com/

including **mBERT** (Pires, 2019), **XLM-R** (Conneau, 2019), and **Llama 3.2** (Dubey et al., 2024); 2) multimodal baselines, comprising **CLIP** (Radford et al., 2021), **BLIP-2** (Li et al., 2023a) and **GPT-4o (2024)**; and 3) hybrid baselines combining multilingual and multimodal LLMs. **L+C**: We use Llama 3.2 and CLIP as text and image encoders, respectively. We then employ concatenation for multimodal feature fusion. We fix CLIP and use Lora to fine-tune Llama. **M2SA** (Thakkar et al., 2024): This is the current state-of-the-art method for multimodal and multilingual sentiment analysis. It employs XLMR-SM and CLIP encoders along with concatenation for feature fusion.

To ensure fairness, we use the same labeled MSA training data for both the multimodal and hybrid baselines. We remove images and rely solely on the labeled texts to train the multilingual baselines. We use accuracy and the weighted average F1 score as evaluation metrics.

4.4 Result Analysis

Main comparisons. Table 4 shows the comparison results on Meld. For the overall classification results, multilingual baselines perform worse than the multimodal baselines. This indicates that additional modal data, such as visual modality, can enhance the performance of single-modal models using textual data. L + C and M2SA perform significantly worse than GPT-4o on the F1 measure, demonstrating that merely combining multilingual and multimodal LLMs does not yield performance gains. Conversely, this may lead to negative effects due to the noise introduced by cross-language fusion and the semantic shift resulting from linguistic-modality fusion. Our method, which benefits from custom-designed modules for cross-lingual noise and semantic shift, performs better than the baseline methods. The slightly lower performance of our method compared to GPT-4o in handling emotions, such as Disgust and Fear, may be attributed to the relatively small number of pre-training samples for these categories. This data imbalance likely confers GPT-4o with a comparative advantage in addressing such scenarios, as its extensive pre-training on diverse datasets enhances its robustness to minority categories.

Cross-lingual performances. Table 5 presents the F1 scores across different languages. English, Spanish, and French belong to the Indo-European language family. The transfer performance from English to French is superior to that from English to Spanish, which can be attributed to the closer linguistic relationship between English and French, characterized by greater similarities in language structure and vocabulary. Notably, the model demonstrates exceptional transfer performance for several low-resource languages, including Occitan and Tajik. When combined with its robust average performance across 23 languages, these results strongly validate the efficacy of language family disentanglement in enhancing cross-lingual transfer capabilities. Furthermore, the model's consistent performance across various low-resource languages from diverse language families underscores its resilience to cross-lingual noise and highlights its strong generalization capabilities beyond linguistic boundaries.

Generalization ability. Table 6 shows the model's overall and cross-language performance on the CH-SIMS dataset. The overall trends align with Table 4, with our method outperforming baselines. The balanced dataset enables the Llama and CLIP combination to fully use their complementary strengths. The model achieves higher F1 scores for pictographic languages (Chinese, Korean, and Japanese), while performance is slightly lower for alphabetic languages like Javanese and Tatar. Nonetheless, the results underscore the model's robust capability in cross-lingual sentiment analysis.

4.5 Ablation Study

To evaluate the contribution of each component in LFD-RT, we conduct an ablation study, with results shown in Table 7. Removing the image modality (w/o CLIP) or removing the text modality (w/o mBERT) results in the most significant performance degradation, highlighting the importance of multimodal information for performance improvement. Eliminating the language family disentangler (w/o LFD) or double cross-alignments (w/o DSA) also leads to notable performance drops, demonstrating their equal importance. Additionally, removing the rethinking fine-tuning module (w/o RFT) degrades the model performance, as this module enables self-adaption to downstream MSA tasks in low-resource languages.

5 Conclusion

In this paper, we propose a multilingual multimodal sentiment analysis framework leveraging language family disentanglement and rethinking transfer. During pre-training, we first perform

	Meld															
Method	**Anger**		**Disgust**		**Fear**		**Joy**		**Neutral**		**Sadness**		**Surprise**		**All**	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
mBERT	15.7	19.1	5.8	8.9	10.9	14.1	28.0	26.3	75.6	65.1	12.0	11.6	21.5	22.4	46.5	32.9
XLM-R	16.3	19.7	5.4	10.1	6.7	9.4	21.8	15.3	80.7	65.5	5.2	8.3	25.0	24.0	47.7	36.3
Llama 3.2	18.1	24.5	20.2	15.5	11.6	8.7	48.2	43.4	73.1	69.4	10.3	13.1	33.6	34.7	50.3	46.9
CLIP	12.0	17.8	7.2	12.4	9.7	11.3	30.3	35.3	81.4	68.3	1.1	2.3	22.0	29.8	48.6	42.5
BLIP2	15.3	16.1	8.4	10.5	9.3	10.7	29.1	32.1	80.5	66.5	10.3	12.8	21.7	27.4	48.0	44.7
GPT-4o	21.2	**28.1**	11.7	16.5	**13.0**	15.5	25.7	32.5	81.8	66.7	14.2	19.4	24.1	28.5	50.8	48.6
L + C	20.6	25.2	17.3	20.8	11.5	13.1	51.0	40.0	81.1	70.1	7.4	10.6	34.6	35.2	51.6	44.1
M2SA	20.1	23.8	21.3	17.5	12.4	**19.6**	38.6	43.8	82.3	70.7	14.6	21.1	24.7	27.0	53.5	43.8
LFD-RT	**21.7**	27.7	**22.0**	**24.1**	12.1	17.2	49.1	**45.3**	**83.7**	**71.1**	**15.2**	**22.4**	**35.1**	**37.8**	**54.7**	**49.1**

Table 4: Comparison results on Meld.

	Meld							
Method	**English**	**Spanish**	**French**	**Occitan**	**Tamil**	**Luxembourgish**	**Tajik**	**Avg.**
mBERT	42.3	33.8	33.1	32.2	34.3	35.1	32.4	32.9
XLM-R	42.6	35.9	34.2	30.7	32.6	36.0	33.9	36.3
Llama	56.1	44.6	55.4	44.8	36.0	25.7	43.9	46.9
CLIP	51.8	41.5	47.7	40.9	39.7	39.4	43.6	42.5
BLIP2	52.2	40.1	43.5	40.6	36.1	37.1	42.8	44.7
GPT-4o	55.6	34.9	**61.5**	35.7	41.3	43.8	44.0	48.6
L + C	44.3	40.0	41.7	39.4	38.2	44.0	46.5	44.1
M2SA	43.7	44.9	41.2	40.0	40.4	41.9	41.3	43.8
LFD-RT	**57.0**	**46.6**	52.7	**46.1**	**43.5**	**44.6**	**47.3**	**49.1**

Table 5: Comparison results on different languages.

CH SIMS		
Method	**Acc.**	**F1**
mBERT	34.3	32.1
XLM-R	34.8	30.9
Llama 3.2	35.1	32.2
CLIP	31.1	24.0
BLIP2	37.4	32.1
GPT-4o	21.5	23.8
L + C	38.8	34.1
M2SA	37.1	33.7
LFD-RT	**40.1**	**34.3**

	CH SIMS					
Method	**Chinese**	**Japanese**	**Korean**	**Javanese**	**Tatar**	**Avg.**
mBERT	36.9	40.0	28.3	25.2	25.5	32.1
XLM-R	38.1	38.6	29.7	23.5	24.7	30.9
Llama	37.8	39.0	30.6	28.9	24.8	32.2
CLIP	30.2	34.4	16.4	18.7	17.7	24.0
BLIP2	37.6	39.2	28.3	28.6	26.8	32.1
GPT-4o	32.0	24.3	26.7	18.3	17.7	23.8
L + C	40.3	39.7	**35.2**	27.4	**28.1**	34.1
M2SA	39.8	**41.0**	32.7	28.9	26.3	33.7
LFD-RT	**41.0**	40.3	33.3	**29.7**	27.0	**34.3**

Table 6: Comparison results on CH SIMS.

cross-lingual and cross-modal alignment, followed by disentangling language families to facilitate the sharing of language universals within the same family while minimizing noise from cross-lingual alignment across different families. To address low-resource language challenges, we introduce a novel double-check rethinking strategy for unsupervised fine-tuning. Experimental results demonstrate the superiority of our approach and highlight the effectiveness of language family-based transfer learning.

Limitations

Although our approach demonstrates promising results, it is subject to two primary limitations. First,

Method	**Acc**	**F1**
w/o CLIP	48.1	42.3
w/o mBERT	48.9	43.4
w/o LFD	51.4	47.1
w/o DSA	51.3	47.4
w/o RFT	53.8	47.6
LFD-RT	54.7	49.1

Table 7: Ablation results of DFD-RT on Meld.

constrained by the limited scale of pre-training data, our model's performance on minority-class emotions has hindered its overall effectiveness. Second, we relied solely on coarse-grained language family information, overlooking the fine-grained hierarchical structure of language families, groups, branches, and individual languages. Consequently, the model's ability to differentiate between languages within the same family remains insufficient.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (62203354, 62472340, 6230336, 62133012, 62425605, 62306235), the Key Research and Development Program of Shaanxi (Program Nos. 2024GX-YBXM-122, 22025CY-YBXM-041, 2024CY2-GJHX-15, 2022ZDLGY01-10), the Natural Science Basic Research Program of Shaanxi (Program Nos. 2025JC-QYXQ-040, 2025JC-YBQN-900, 2025CY-YBXM-067), Guangdong Basic and Applied Basic Research Foundation (Program No. 2025A1515010267), the Postdoctoral Research Project Sustentation Fund of Shaanxi (No. 2024BSHSDZZ112), and the Key Scientific Research Projects of the Shaanxi Provincial Education Department (No. 24JR152).

References

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Xin Dong, Yaxin Zhu, Yupeng Zhang, Zuohui Fu, Dongkuan Xu, Sen Yang, and Gerard De Melo. 2020. Leveraging adversarial training in self-learning for cross-lingual text classification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1541–1544.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5759–5771.

Ling Ge, Chunming Hu, Guanghui Ma, Jihong Liu, and Hong Zhang. 2024. Da-net: A disentangled and adaptive network for multi-source cross-lingual transfer learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18047–18055.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192.

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.

Hiroshi Kanayama, Yang Zhao, Ran Iwamoto, and Takuya Ohko. 2024. Incorporating syntax and lexical knowledge to multilingual sentiment classification on large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4810–4817.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. 2024. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12458–12468.

Zejun Li, Zhihao Fan, Jingjing Chen, Qi Zhang, Xuan-Jing Huang, and Zhongyu Wei. 2023b. Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5939–5958.

Nankai Lin, Yingwen Fu, Xiaotian Lin, Dong Zhou, Aimin Yang, and Shengyi Jiang. 2023. Cl-xabsa: Contrastive learning for cross-lingual aspect-based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Md Saef Ullah Miah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and MF Mridha. 2024. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm. *Scientific Reports*, 14(1):9603.

T Pires. 2019. How multilingual is multilingual bert. *arXiv preprint arXiv:1906.01502*.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain. 2018. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25.

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.

Gaurish Thakkar, Sherzod Hakimov, and Marko Tadić. 2024. M2sa: Multimodal and multilingual model for sentiment analysis of tweets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10833–10845.

Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Soujanya Poria. 2023. Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6045–6053.

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.

Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. 2022. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1545–1554.

Zheyu Zhao, Zhongqing Wang, Shichen Li, Hongling Wang, and Guodong Zhou. 2025. Bridging modality gap for effective multimodal sentiment analysis in fashion-related social media. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1813–1823.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Fuji Ren. 2024. Kebr: Knowledge enhanced self-supervised balanced representation for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5732–5741.