SentiXRL: An advanced large language Model Framework for Multilingual Fine-Grained Emotion Classification in Complex Text Environment

Anonymous COLING 2025 submission

Abstract

With strong expressive capabilities in Large Language Models(LLMs), generative models effectively capture sentiment structures and deep semantics, however, challenges remain in fine-grained sentiment classification across multi-lingual and complex contexts. To address this, we propose the Sentiment Cross-Lingual Recognition and Logic Framework (SentiXRL), which incorporates two modules,an emotion retrieval enhancement module to improve sentiment classification accuracy in complex contexts through historical dialogue and logical reasoning,and a self-circulating analysis negotiation mechanism (SANM)to facilitates autonomous decision-making within a single model for classification tasks.We have validated SentiXRL's superiority on multiple standard datasets, outperforming existing models on CPED and CH-SIMS,and achieving overall better performance on MELD,Emorynlp and IEMOCAP. Notably, we unified labels across several fine-grained sentiment annotation datasets and conducted category confusion experiments, revealing challenges and impacts of class imbalance in standard datasets.

1 Introduction

Currently, utilizing large language models (LLMs) for text classification tasks is a prominent research focus(Brown et al., 2020b). Specifically, text sentiment classification has garnered widespread attention due to its significance in understanding the nuances of human communication. Generative models, with their powerful expressive capabilities, can effectively capture the structure and deep semantics of emotional texts, thereby demonstrating outstanding performance in sentiment recognition and classification tasks,which is a solid foundation for other tasks such as roleplay, dialogue generation, and targeted content recommendations.Additionally, instruction fine-tuning of LLMs has proven their exceptional adaptability to various tasks(Ouyang et al., 2022). However, for more complex tasks such as fine-grained sentiment recognition, efficient processing frameworks are often required. In the context of multilingual communication and cultural differences, the complexity of multilingual understanding and response poses higher demands on the generalization capability of LLMs. The differing grammatical and syntactic features across languages, along with the limitations of traditional algorithms that focus on structured and short dialogue scenarios while overlooking more personalized user expressions, are among the many challenges that LLMs currently face.

Our goal is to design an efficient framework for fine-grained emotion classification tasks for LLMs in multilingual and complex text environment. To this end, we design the SentiXRL cross-lingual emotion recognition framework, which enables fine-grained emotion recognition in more complex textual environment across multiple languages. Our architecture primarily includes an efficient emotion retrieval enhancement module, which connects contextual information through historical dialogues and implicit inference while performing emotion reasoning. Additionally, we design a Self-Analytical Negotiation Mechanism (SANM) to help LLMs perform emotion verification and logical reasoning, thereby improving emotion classification capabilities in complex texts and contexts.

We validate our approach on several standard benchmark datasets, surpassing existing SOTA on most benchmarks. Furthermore, we construct the largest fine-grained sentiment annotation dataset to date and conduct category confusion experiments , verifying the impact of category imbalance on LLMs.

Finally, our contributions are summarized as follows:

- We propose a novel framework specifically designed for the task of cross-lingual fine-grained emotion recognition in large language models.

- A novel Self-Analytical Negotiation Mechanism (SANM) is introduced, enhancing emotion recognition accuracy in complex environments through logical reasoning and emotion verification.
- SentiXRL outperforms most previous models on five standard Emotion Recognition in Conversations(ERC) benchmarks and achieves comprehensive single-modal state-of-the-art on two emotion analysis datasets.
- To address category imbalance in mainstream datasets, we standardized label mapping across multiple fine-grained emotion annotation datasets and conducted category confusion experiments. Additionally, ablation studies on the ERC datasets highlight the advantages of the SANM module.

2 Related Work

Dialogue emotion recognition has evolved from traditional machine learning methods such as SVM, which focused primarily on general textual sentiment, to deep learning approaches. Notable among these are mainstream discriminative architectures like RNN, GNN, and LSTM (Poria et al., 2017a), which capture complex inter-sentence dependencies, or Transformers (Liu et al., 2023) that effectively capture contextual information. These advancements have significantly enhanced the accuracy of analysis. The emergence of multimodal fusion (e.g., combining speech or facial emotions) has enabled these discriminative models to comprehensively understand and recognize emotional states in dialogues. However, the integration of more modalities introduces limitations in application scope and complexity in data collection. Consequently, some researchers have begun incorporating dialogue modeling and situational interactions (Lei et al., 2024) or attempting to infuse common-sense information into emotion recognition tasks (Yi et al., 2022; Li et al., 2021).

2.1 Logical Reasoning in Text Emotion Recognition

In our view, both common-sense information and other modalities serve as supplementary information external to the dialogue itself. These types of information cannot fully cover all dialogue scenarios or operate in constrained environment. Therefore, enabling models to reason and validate is the true solution for text emotion recognition tasks. Moreover, mainstream discriminative models suffer from complex system design and overfitting to specific datasets or dialogue patterns. Thus, generative architectures based on large language models (LLMs) have emerged as a novel approach to addressing these issues. The successful application and emergent capabilities of LLMs (Zhao et al., 2023) have demonstrated their excellent performance in natural language reasoning tasks. Research has shown that LLMs can follow contextual information (Brown et al., 2020a) and comprehend natural language instructions (Mishra et al., 2022; Chung et al., 2022). However, LLMs still underperform in reasoning tasks compared to smaller models (e.g., fine-tuned BERT) (Lee et al., 2023), presenting challenges for the application of LLM-based logical reasoning in text emotion recognition tasks.

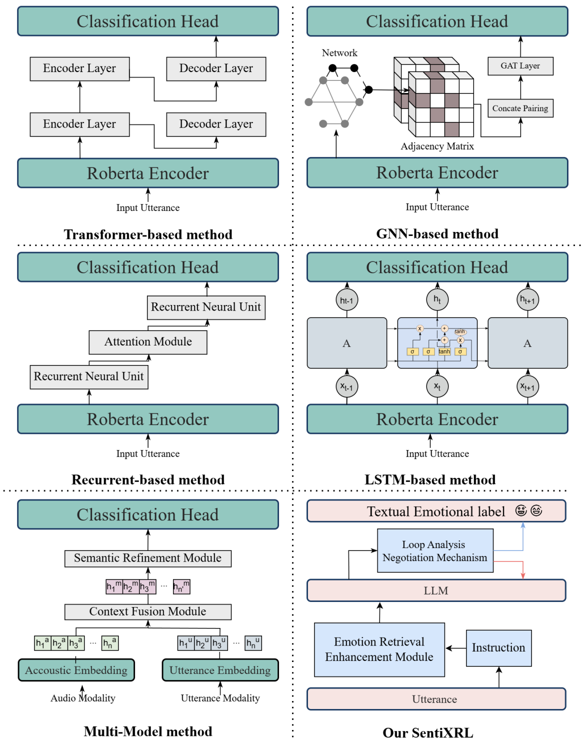

Figure 1: The illustration of different paradigms for ERC

3 Methodology

This chapter provides an in-depth overview of the novel SentiXRL architecture, detailing its emotion retrieval enhancement module, self-circulating analysis negotiation mechanism, and emotion analysis tasks. It also thoroughly explains the experimental training and inference processes.

3.1 Emotion Retrieval Enhancement Module

To better leverage the reasoning capabilities of large language models (LLMs), we restructure the ERC task into a sequence format, fine-tuning the LLM. To adapt the LLM to the specific emotion recognition task at hand, we design an efficient emotion retrieval module. As shown in Figure 2, this module consists of instructions, a history window, label statements, and emotional deduction.

Instructions I: Define the specific task content and standardized format.

History Window H: Represents the round of historical dialogue information used to connect the previous sequence of words, specifically in the form of:

$$H = [h_1, h_2, \ldots, h_m] \tag{1}$$

Labels L: Restrict the model's output range, allowing it to output label categories $ld \in D$ within the label domain D[1].

Emotional Deduction E: Utilize the reasoning capabilities of the generative model to infer possible scenarios S, characters P, and relationships R based on historical dialogue and the current statement. Thus, $E = (S, P, R)$.

Therefore, the task of this module can simplify the processing of the input statement u_i as follows:

$$T_i = [I_{u_i}, H_{u_i}, ld_{u_i}, E_{u_i}] \tag{2}$$

3.2 Self-circular Analysis Negotiation Mechanism

Due to the generative nature of LLM in sentiment analysis tasks, despite fine-tuning efforts and instruction-based constructions, there is still no guarantee that the output belongs to the specified sentiment category, especially for fine-grained sentiment classification tasks. To address the need for correction and supervision, and to mitigate the potential inaccuracies and lack of specificity in individual LLM outputs, WE propose a cyclic verification analysis negotiation mechanism to assist in completing sentiment analysis tasks. This mechanism differs from multi-LLM negotiation strategies and the supervised learning strategy using contextual learning (ICL) paradigms.

The core of this strategy is a generative-discriminative architecture. Unlike conventional approaches requiring an additional discriminative model for supervision, the proposed approach leverages the original LLM model. Given the effectiveness of full parameter training of base models in fine-tuning tasks, there is no need for a separate discriminator. Instead, a new module framework is designed for the original model, integrating tasks of inference and output credibility. This allows for verification of classification objectives and explanatory derivation following structured reasoning chains after preliminary classification tasks.

3.2.1 Inference Generator and Deduction Discriminator

Compared to introducing a deduction discriminator, utilizing the distributional inference of two large language models makes the single LLM analysis negotiation framework more convenient and efficient in handling subtasks. The key lies in implementing an alternating role mechanism where a single LLM can function both as a generator and a discriminator. This paper defines distinct task templates for the generator and discriminator to ensure that the LLM comprehends its current role and task accurately. The generator judges the sentiment of text and generates a chain of distributional inferences and sentiment decisions, while the discriminator evaluates the generator's output and provides explanations.

3.2.2 Role Alternation and Consensus Mechanism

In each interaction round, the role of the LLM is clearly defined, and operations are performed according to different task templates through multi-round interaction processes. If the LLM reaches the same sentiment decision in two consecutive rounds, consensus is achieved. In case of disagreement, multiple attempts are made within the maximum round limit. If consensus is achieved, the process concludes; otherwise, an outlier is directly outputted if no valid sentiment decision is reached within the maximum round limit.

3.2.3 Mathematical Modeling

The self-circular analysis negotiation mechanism can also be represented mathematically by mod-

[1]In traditional sentiment analysis, sentiment primarily includes the categories positive, negative, and neutral, and many sentiment datasets follow this convention. In fine-grained datasets, emotion categories can be mapped to sentiment using emotion dictionaries. However, the surprise shown in Figure 2 needs to be categorized based on the text content. The emotion categories appearing in Figure 2 are those used in the datasets listed below.

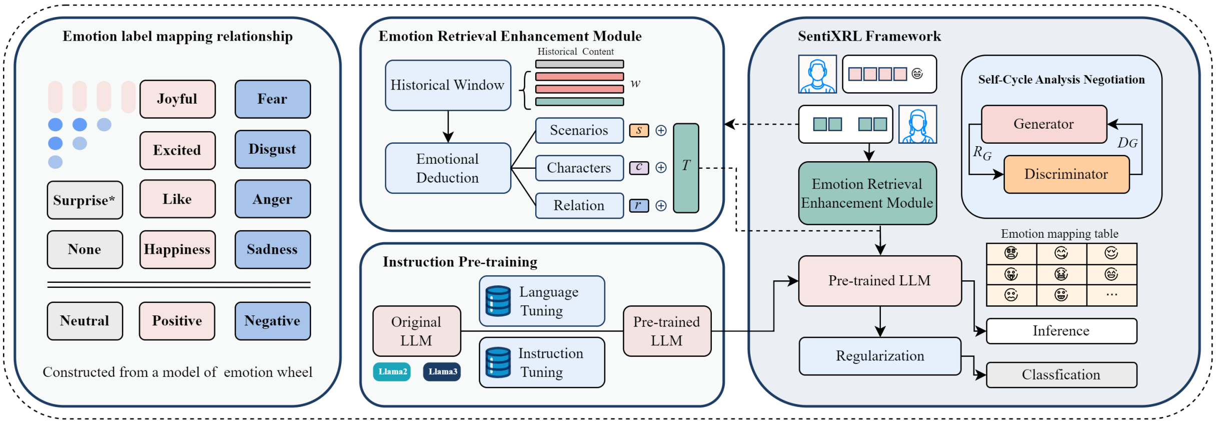

Figure 2: Overview of SentiXRL Framework

eling the interaction between the generator and discriminator as an iterative process, defined as follows:

- G: Generator
- D: Discriminator
- T: Sentiment analysis text
- R_G: Generator response (including sentiment analysis)
- R_D: Discriminator response (evaluating generator response)
- E: Final sentiment analysis
- N: Maximum number of cycles

The generator G generates a response R_G based on the sentiment analysis text T:

$$R_G^{(n)} = G(T) \tag{3}$$

The discriminator D generates a response R_D based on $R_G^{(n)}$:

$$R_D^{(n)} = D(R_G^{(n)}) \tag{4}$$

If $R_D^{(n)}$ correctly evaluates the inference, $E = R_G^{(n)}$; otherwise, $n = n + 1$, repeating the above steps until $n = N$ or $R_D^{(n)}$ correctly evaluates the inference.

The self-circular analysis negotiation mechanism harnesses the capabilities of a single LLM across multiple attempts to achieve inference and decision-making. This method, compared to integrating a new inference generator utilizing the collective capabilities of two LLMs for the entire decision-making process, only requires a decision-making framework, enabling multi-LLM functionality. Moreover, this approach is applicable to other subtasks or LLM-based tasks, requiring only the redesign of the inference framework without the need for separate training of new discriminators.

Due to the imbalanced class distribution in the standard baseline dataset used, we adopt Focal Loss as the loss function for subsequent main tasks. This approach increases the weight of hard-to-classify samples, thereby mitigating the impact of class imbalance on the results:

$$L = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \tag{5}$$

- p_t: The model's predicted probability for the actual class
- α_t: The weighting factor to balance positive and negative samples. For a sample of class t, if $y = 1$, then $\alpha > 0$
- γ: A modulating factor to reduce the weight of easy-to-classify samples

The core idea of Focal Loss is to introduce the modulating factor $(1-p_t)^{\gamma}$, which reduces the loss contribution from samples that the model already predicts with high confidence, while increasing the loss contribution from samples that the model predicts with less confidence. This adjustment encourages the model to focus more on the hard-to-classify samples. In our experiments, the default parameters are set as: $\alpha = 0.25, \gamma = 2.0$.

4 Experiments

Initially, we considered using cross-lingual sentiment analysis datasets, such as XED(Öhman et al.,

2020) or NaijaSenti(Muhammad et al., 2022), for our experiments. These datasets include texts from various language categories and offer fine-grained sentiment annotations. However, we ultimately decided against using them for the following reasons:

1)For the current task of dialog sentiment analysis, the most widely used sentiment datasets are primarily in English and Chinese. Many state-of-the-art methods and models have been validated on these datasets, making them more representative in testing. In contrast, multilingual datasets are still rarely used in this research area.

2)Taking XED as an example, most cross-lingual sentiment datasets are originally annotated in English, with annotations for other languages created through projection or translation. While this ensures linguistic diversity, the cultural and linguistic specificity of the original language limits the applicability of these annotations to other languages. This misalignment means that the dataset may not align with the usage habits of the target language communities, making it less representative for languages other than the one in which it was originally annotated.

Thus, we opted to focus on English and Chinese for the cross-lingual sentiment analysis task, using data collected from movies or personal daily content for training and testing, allowing the model to better adapt to real-world scenarios.

Firstly, for English emotion classification tasks, we selected several challenging datasets: MELD (Poria et al., 2018), EmoryNLP (Zahiri and Choi, 2018), and IEMOCAP (Busso et al., 2008). IEMOCAP is an interactive emotional dyadic motion capture database that covers emotional exchanges in daily life. The MELD and EmoryNLP datasets are derived from dialogues in the TV show **Friends**, providing contextual information and fine-grained emotion labels. Particularly, EmoryNLP also includes emotion annotations for long dialogue sequences. However, it is worth noting that all datasets have imbalanced emotion distributions.

For Chinese emotion recognition tasks, we chose CPED (Chen et al., 2022) and CH-SIMS (Liu et al., 2022) as standard baseline datasets to evaluate the effectiveness of SentiXRL. The CPED dataset includes personal characteristics, various dialogue behaviors, and scenarios, while CH-SIMS offers richer character backgrounds and spans across different ages. Therefore, these datasets pose greater challenges. Although some of these datasets are multimodal, our study currently focuses solely on the emotion categories and textual modality of the data.

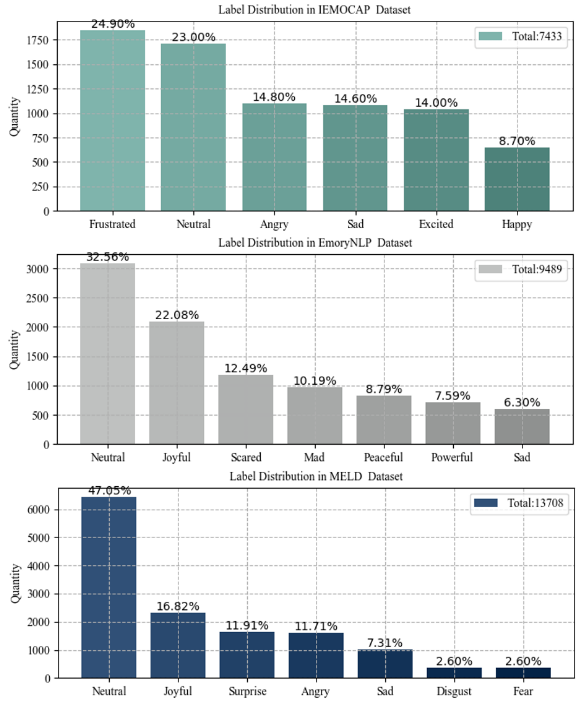

Figure 3: Label Distribution in MELD,EmoryNLP and IEMOCAP Dataset

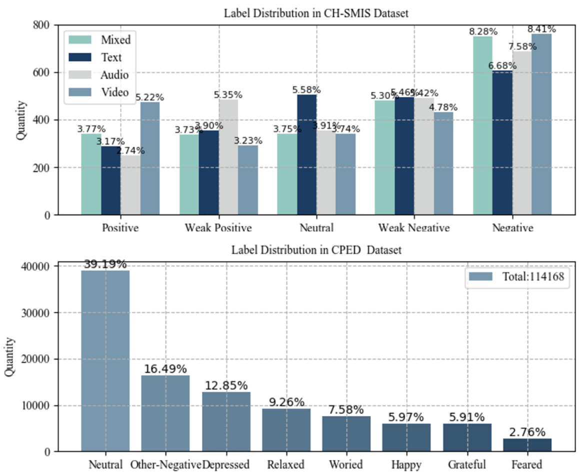

Figure 4: Label Distribution in CH-SIMS and CPED

We compared several single-text modality baselines with SentiXRL, conducting experiments on the Llama2-7B (L2) and Llama3-8B (L3) models. For Chinese emotion recognition tasks, the baseline models included MMML (Wu et al.), ALMT (Zhang et al., 2023a), bcLSTM (Poria et al., 2017b), DialogXL (Shen et al., 2021), and BERT-AVG-MLP (Chen et al., 2022). For English emotion recognition tasks, the baseline models consisted of SPCL+CL (Song et al., 2022), SACL (Hu et al., 2023), EmotionIC (Yingjian et al., 2023), DualGATs (Zhang et al., 2023c), and InstructERC (Lei et al., 2024). For more detailed information on the baseline models and their implementations, please refer to the appendix.

Table 1: Results on two Chinese Benchmarks

Dataset	CH-SIMS		CPED	
Models	**F1**	**Accuracy**	**Macro-F1**	**Accuracy**
Discriminant Models				
MMML	82.9	-	-	-
ALMT	81.57	-	-	-
bcLSTM	-	-	49.65	45.40
DialogXL	-	-	51.24	46.96
BERT-AVG-MLP	-	-	51.50	48.02
Generative Models				
SentiXRL(L_2)	76.51	83.15	32.96	47.00
SentiXRL(L_3)	**82.83**	**84.20**	45.31	**50.70**

Table 2: Results on three English Benchmarks

Dataset Models	IEMOCAP W-F1	MELD W-F1	Emorynlp W-F1	Average W-F1
Discriminant Models				
SPCL+CL	**69.74**	66.35	40.25	58.78
SACL	69.22	66.45	39.65	58.44
EmotionIC	69.61	66.40	40.01	58.67
DualGATs	67.68	**66.90**	**40.29**	58.29
Generative Models				
InstructERC	**71.39**	**69.15**	41.37	60.64
SentiXRL(L_2)	70.52	67.33	40.37	59.41
SentiXRL(L_3)	71.11	68.72	**42.51**	**60.78**

4.1 Main Results

Table 1 and Table 2 respectively present the comparative results of the SentiXRL model against other models on Chinese and English benchmark datasets. The experimental results indicate that our method significantly outperforms existing discriminative models and surpasses the current SOTA models in most benchmarks. Specifically, in the Chinese sentiment classification task, our accuracy on the CPED dataset shows an improvement of 5.6% over the existing SOTA, and the F1 score on the CH-SIMS dataset increases by 1.55%. Similarly, in the English sentiment classification benchmarks, SentiXRL achieves the highest individual performance on the more challenging EmoryNLP dataset and surpasses the existing SOTA in the average Weighted-F1 score across three datasets. The experimental results demonstrate that SentiXRL excels in both Chinese and English linguistic environment, validating the model's compatibility and adaptability in multilingual contexts.

4.2 Ablation Study

The ablation study results indicate that removing any component leads to a decline in relevant metrics, demonstrating that each part of the SentiXRL model is essential. Notably, the performance significantly drops when the ANM cyclic negotiation mechanism is removed, further proving the importance of this module. Additionally, multiple experiments have shown that this mechanism effectively helps the LLM focus on the current task. Even without fine-tuning, the model's task execution capability is relatively well improved.

4.3 Category Impact Verification

To validate the impact of data categories on the results of SentiXRL, we conduct an experiment addressing the inconsistency of emotion labels in most textual sentiment datasets. In this experiment, we collect high-quality, fine-grained Chinese textual sentiment classification datasets that are currently open-source on the Chinese internet. We standardize the emotion labels across all datasets (the mapping rules are shown in Figure 2) and perform data processing and cleaning. Detailed information on data processing and dataset composition can be found in Appendix B.

In the experiment, we design two classification methods: random data mixing and equal category mixing. These methods are intended to explore the impact of different data mixing strategies on the model. We hypothesize that due to the imbalance of categories in various datasets, the independent features of smaller datasets or those with fewer categories may be overshadowed by larger datasets. Therefore, equal category sampling can better highlight the characteristics and impacts of datasets with fewer categories. Both methods used the same data scale and hyperparameter settings.

Figure 5 presents the results of category validation. The experimental results indicate that, after 12K steps, the random mixing method surpasses the equal mixing method in accuracy. This demonstrates SentiXRL's advantage in recognizing certain emotional categories. For more complex emotional categories, such as surprise, the limited textual content makes recognition relatively more challenging,

Table 3: The ablation results of Llama2 and Llama3 on five benchmarks

Dataset	IEMOCAP, MELD, EmoryNLP, Average				CH-SIMS		CPED	
Models	**Weighted-F1**	**Weighted-F1**	**Weighted-F1**	**Weighted-F1**	**F1**	**Accuracy**	**Macro-F1**	**Accuracy**
				Zero-shot+SentiXRL				
w/o $_{SANM}+L_2$	-	-	-	-	37.6	26.3	11.2	10.5
L_2	41.72	32.14	27.13	33.66	52.5	49.9	21.7	27.8
w/o $_{SANM}+L_3$	-	-	-	-	30.8	29.5	13.0	28.3
L_3	44.85	31.67	27.96	34.83	58.7	56.0	36.8	36.4
				LoRA+Backbone				
w/o $_{SANM}+L_2$	-	-	-	-	59.3	59.8	24.3	42.7
L_2	53.27	37.69	29.51	40.16	61.2	63.6	26.8	44.5
w/o $_{SANM}+L_3$	-	-	-	-	65.0	68.3	37.5	47.7
L_3	55.01	38.08	30.28	41.12	67.5	71.0	39.8	49.2
				LoRA+SentiXRL				
w/o $_{SANM}+L_2$	70.52	67.33	40.37	59.41	76.5	83.2	33.0	47.0
w/o $_{SANM}+L_3$	**71.11**	**68.72**	**42.51**	**60.78**	**82.8**	**84.2**	**45.3**	**50.7**

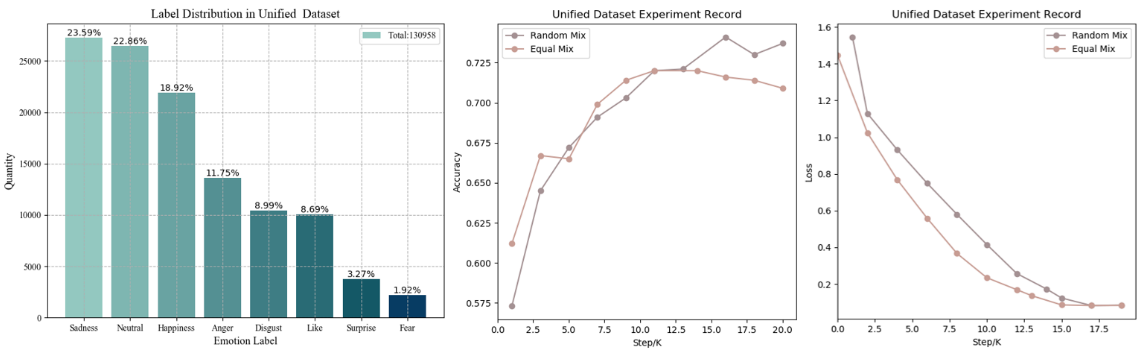

Figure 5: Experimental results of SentiXRL(L_3) on Unified Dataset

but this result aligns with expectations. Additionally, the loss graph shows that the dataset with equal mixing converges more rapidly. Due to the balanced distribution of categories, the model can learn the feature mapping relationships for all categories more quickly. Overall, data category impact on SentiXRL's accuracy is minimal.

4.4 Model performance validation in complex text environments

To further validate the generalization capability of the SentiXRL model, this study selected two additional datasets, Twitter2015(Liu et al., 2015) and Twitter2017(Rosenthal et al., 2017), used for sentiment analysis and assessing information veracity. In contrast to the datasets used in the main experimental section, Twitter2015 and Twitter2017 feature longer text lengths. Unlike CPED, MELD, and EmoryNLP datasets sourced from scripted and emotional short texts from various film and TV show dialogues, the Twitter datasets originate from social media, where data is more heterogeneous and topics are less defined, thereby posing a greater challenge to the model's recognition abilities and subjecting its performance to more rigorous testing in complex textual and linguistic environments. We conducted comparative experiments and evaluated SentiXRL's performance on these datasets using Macro-F1 and Accuracy metrics.

The experimental results indicate that SentiXRL outperforms existing models in single-modal text analysis, particularly on the Twitter2015 and Twitter2017 datasets[2]. It achieved improvements of 2.6% and 2.87% in Accuracy, and 3.21% and 4.63% in Macro-F1 scores, respectively. Additionally, SentiXRL's performance in multimodal models also ranks among the top. Therefore, these

[2]The abbreviations '15' and '17' in the 'Acc' and 'Mac-F1' columns refer to Twitter 2015 and Twitter 2017, respectively.

Table 4: The Main results on Twitter Benchmark

Models	Venue	Acc(15)	Mac-F1	Acc(17)	Mac-F1
		Text Only			
AE-LSTM(Wang et al., 2016)	EMNLP 2016	70.30	63.43	61.67	57.97
MemNet(Tai et al., 2017)	EMNLP 2016	70.11	61.76	64.18	60.90
RAM(Zhang et al., 2023b)	EMNLP 2017	70.68	63.05	64.42	61.01
MGAN(Hoang et al., 2018)	EMNLP 2018	71.17	64.21	64.75	61.46
BERT(Devlin et al., 2019)	NAACL 2019	74.15	68.86	68.15	65.23
		Image and Text			
MIMN(Xu et al., 2019)	AAAI 2019	71.84	65.59	65.88	62.99
ESAFN(Yu et al., 2020)	TASLP 2019	73.38	67.37	67.83	64.22
VilBERT(Lu et al., 2019)	NeurIPS 2019	73.69	69.53	67.86	64.93
TomRoBERTa(Yu and Jiang, 2019)	IJCAI	77.46	72.95	71.12	69.49
ModalNet-Bert(Zhang et al., 2021)	WWW 2021	76.71	70.93	69.55	67.28
EF-CapRoBERTa(Khan and Fu, 2021)	ACM	78.19	73.51	71.14	68.74
FITE(Yang et al., 2022)	EMNLP 2022	78.49	73.90	70.90	68.70
HIMT(Yu et al., 2023)	TAC 2022	78.14	73.68	71.14	69.16
ITM(Yu et al., 2022)	IJCAI 2022	78.27	74.19	72.61	71.97
		Text Only			
Ours	-	77.28	70.93	70.84	69.12

results confirm that SentiXRL performs exceptionally well across both standard datasets and datasets characterized by significant information noise and unclear themes.

5 Conclusion

We introduce the SentiXRL model — an advanced large language model framework for multilingual fine-grained emotion classification in complex text environment,and discuss the Emotion Retrieval Enhancement Module and Self-circular Analysis Negotiation Mechanism within this architecture. This study utilized bilingual Chinese-English datasets and validated the model's effectiveness through comparative experiments and ablation studies. Experimental results demonstrate that SentiXRL achieved improved accuracy across multiple standard datasets, showcasing its superior performance in emotion recognition tasks and efficient fine-grained emotion classification in multilingual settings. Compared to traditional discriminative models, SentiXRL is capable of generating richer and more accurate emotion category labels. The Self-circular Analysis Negotiation Mechanism (SANM) further enhances model stability and accuracy through alternating roles of generator and discriminator, facilitating self-supervision and validation. The effectiveness of SentiXRL suggests new directions for future research in emotion recognition.

6 Limitation

SentiXRL currently focuses solely on unimodal textual information, and due to limitations in the experimental environment, this study was conducted using a pre-trained model with a maximum of 8 billion parameters. In the future, we will explore the potential for multimodal research and extend our investigations to include testing on Arabic,Hindi and Spanish. These languages, being among the most widely spoken after English and Chinese, will help demonstrate the model's effectiveness across a broader range of linguistic contexts.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack

Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. 2022. Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. 2018. Mgan: Training generative adversarial nets with multiple generators.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. *arXiv preprint arXiv:2306.01505*.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? *Preprint*, arXiv:2305.13788.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2024. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *Preprint*, arXiv:2309.11911.

JiangNan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Conference on Empirical Methods in Natural Language Processing*.

Xiao Liu, Jian Zhang, Heng Zhang, Fuzhao Xue, and Yang You. 2023. Hierarchical dialogue understanding with special tokens and turn-level attention. *Preprint*, arXiv:2305.00262.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1867–1870.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022. Make acoustic and visual cues matter: Ch-sims v2.0 dataset and av-mixup consistent module. *Preprint*, arXiv:2209.02604.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Preprint*, arXiv:1908.02265.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, et al. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *arXiv preprint arXiv:2201.08277*.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. Xed: A multilingual dataset for sentiment analysis and emotion detection. *arXiv preprint arXiv:2011.01612*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, and Louis Philippe Morency.

2017a. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. *arXiv preprint arXiv:2210.08713*.

Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. 2017. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Zehui Wu, Ziwei Gong, Jaywon Koo, and Julia Hirschberg. Multimodal multi-loss fusion network for sentiment analysis.

Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 371–378.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3324–3335, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jingjie Yi, Deqing Yang, Siyu Yuan, Caiyan Cao, Zhiyao Zhang, and Yanghua Xiao. 2022. Contextual information and commonsense based prompt for emotion recognition in conversation. *Preprint*, arXiv:2207.13254.

Liu Yingjian, Li Jiang, Wang Xiaoping, and Zeng Zhigang. 2023. Emotionic: Emotional inertia and contagion-driven dependency modelling for emotion recognition in conversation. *arXiv preprint arXiv:2303.11117*.

Jianfei Yu, Kai Chen, and Rui Xia. 2023. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):1966–1978.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. In *Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19*.

Jianfei Yu, Jing Jiang, and Rui Xia. 2020. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.

Jianfei Yu, Jieming Wang, Rui Xia, and Junjie Li. 2022. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4482–4488. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023a. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. *arXiv preprint arXiv:2310.05804*.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. 2023b. Recognize anything: A strong image tagging model. *Preprint*, arXiv:2306.03514.

Yupeng Zhang, Shensi Wang, Peiguang Li, Guanting Dong, Sirui Wang, Yunsen Xian, Zhoujun Li, and Hongzhi Zhang. 2023c. Pay attention to implicit attribute values: A multi-modal generative framework for AVE task. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13139–13151, Toronto, Canada. Association for Computational Linguistics.

Zhe Zhang, Zhu Wang, Xiaona Li, Nannan Liu, Bin Guo, and Zhiwen Yu. 2021. Modalnet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network. *World Wide Web*, (6).

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Model Fine-Tuning

Due to the relatively low proportion of Chinese training data in the original Llama model, we conduct fine-tuning on Chinese instructions to enhance its capability in understanding and expressing Chinese. Before conducting experiments, we fine-tune the model using approximately 3.5 million samples from the BELLE dataset and the moss-003 dataset released by Fudan University's MOSS team. During fine-tuning, we employ the Stanford Alpaca template, an effective training method designed to improve the model's ability to understand and execute instructions more effectively throughout the training process.

B Introduction to Baseline Model

Here is the detailed introduction of the baseline model in the supplementary experimental section.

- MMML

  MMML (Multimodal Multi-loss Fusion Network) proposes a multimodal multi-loss fusion network that compares different fusion methods and evaluates the impact of multi-loss training in multimodal fusion networks. It enhances model performance significantly by integrating contextual information.

- ALMT

  ALMT (Adaptive Language-guided Multimodal Transformer) introduces an Adaptive Hyper-modality Learning (AHL) module to learn representations that suppress irrelevant/conflicting information from visual and audio features under the guidance of language features at different scales. By effectively suppressing redundant information in visual and audio modalities, ALMT improves performance significantly across several popular datasets.

- bcLSTM

  bcLSTM designs a model based on Long Short-Term Memory (LSTM) networks, allowing the capturing of contextual information within the same video segment. By recognizing relationships between environments and characters, it aids in better character emotion classification. This method demonstrates robust generalization capabilities.

- DialogXL

  DialogXL is a pre-trained language model specifically designed for dialogue emotion recognition tasks. By enhancing recursive mechanisms and self-attention mechanisms and improving memory capabilities, DialogXL effectively improves emotion recognition performance. DialogXL modifies XLNet's recursive mechanism from paragraph-level to utterance-level to better simulate dialogue data, and introduces dialogue-aware self-attention to capture useful dependencies within and between speakers.

- CPED(BERT+AVG+MLP)

  The CPED paper introduces the BERT+AVG+MLP model for emotion recognition in dialogues (ERC). This model combines BERT's pre-trained language model with average pooling (AVG) of BERT's hidden layer outputs, passing this output through a multi-layer perceptron (MLP) to predict emotion labels.

- SPCL+CL

  This approach involves a novel Supervised Prototypical Contrastive Learning (SPCL) loss function for emotion recognition in dialogues, outperforming traditional supervised contrastive learning losses. SPCL performs well on imbalanced data categories, is insensitive to training batch size, and reduces computational resource requirements.

Table 5: Fine-tuning hyperparameter settings for Chinese Instruction

	Full parameters	**Full parameters**
Base Model	Meta-Llama-2-7B-Instruct	Meta-Llama-3-8B-Instruct
Epochs	1	2
Learning Rate	2e-4	3e-6
LR Scheduler Type	constant	cosine
Context Length	2K	8K
Attention Heads	32	32
Key Value Heads	32	8
Warmup Ratio	0.01	0.1

- SACL

  SACL (Supervised Adversarial Contrastive Learning) is a supervised adversarial contrastive learning framework. It trains adversarial examples by contrasting perceptual adversarial training to generate worst-case samples. It uses joint category expansion contrastive learning objectives to extract structured representations, effectively leveraging label-level feature consistency and preserving fine-grained intra-class features. To mitigate the negative impact of adversarial perturbations on context-dependent data, the framework includes a context adversarial training strategy to learn more diverse features from contexts, enhancing the model's contextual robustness.

- EmotionIC

  EmotionIC models emotional dependencies in dialogues based on emotional inertia and contagion. Compared to previous ERC models, EmotionIC provides a more comprehensive modeling of dialogues at both feature extraction and classification levels. The model attempts to integrate the advantages of attention-based and recursive-based approaches at the feature extraction level.

C Unified Dataset

Given the standard baseline datasets chosen, such as CPED, MELD, and EmoryNLP, sourced from dialogue lines of TV show characters, most recorded text sentiments tend towards positive emotions, resulting in an overall imbalanced category distribution. Our unified dataset experiment aims to investigate the impact of this category bias on SentiXRL. Our approach to validation involves gathering a larger-scale, fine-grained emotional data set, aiming not only to balance category proportions but also to mitigate the influence of scripted dialogue styles and situational contexts typical in TV dramas.

Table 6: Twitter Dataset Category Weighting Table

	Neutral	**Positive**	**Negative**
Twitter 2015	0.563	1.139	2.884
Twitter 2017	0.723	0.786	2.899

D Overhead and Computational Efficiency Comparison

The Llama2 and Llama3 models we used were trained and tested on a single L20 GPU, which indeed requires certain hardware specifications. However, by deploying the model within the SentiXRL framework, we found that on a single L20 GPU, text with a length of less than 500 characters, under the self-circulating analysis and negotiation mechanism (SANM) with $Max\_round = 3$, has an average processing time of 1.8s (based on an average of 1000 sample data). This is not significantly different from the 1.4s for zero-shot deployment of the model alone. Therefore, the additional computational load brought by the SentiXRL framework is within an acceptable range. The computational capacity of large models is indeed highly dependent on the hardware environment, which imposes limitations on deployment.

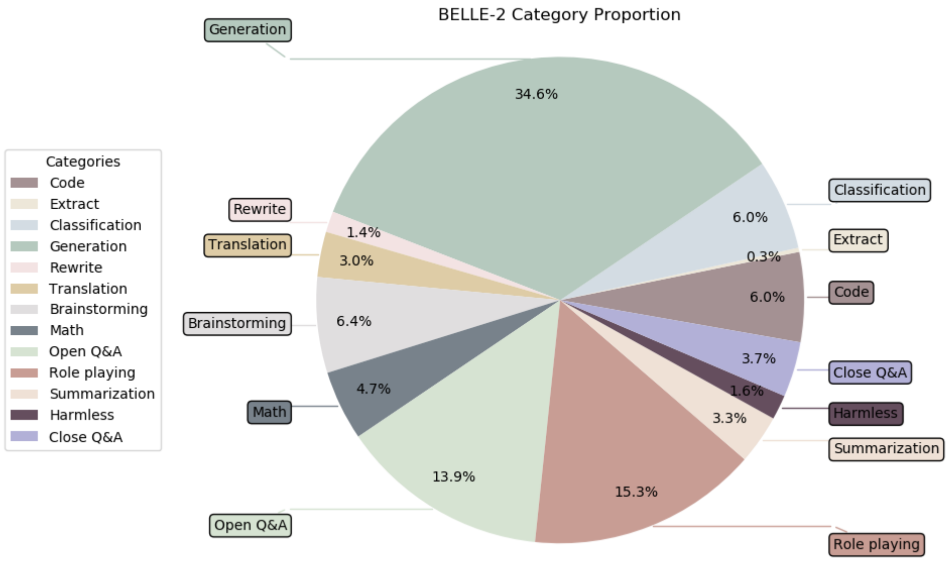

Figure 6: BELLE Data Structure Distribution

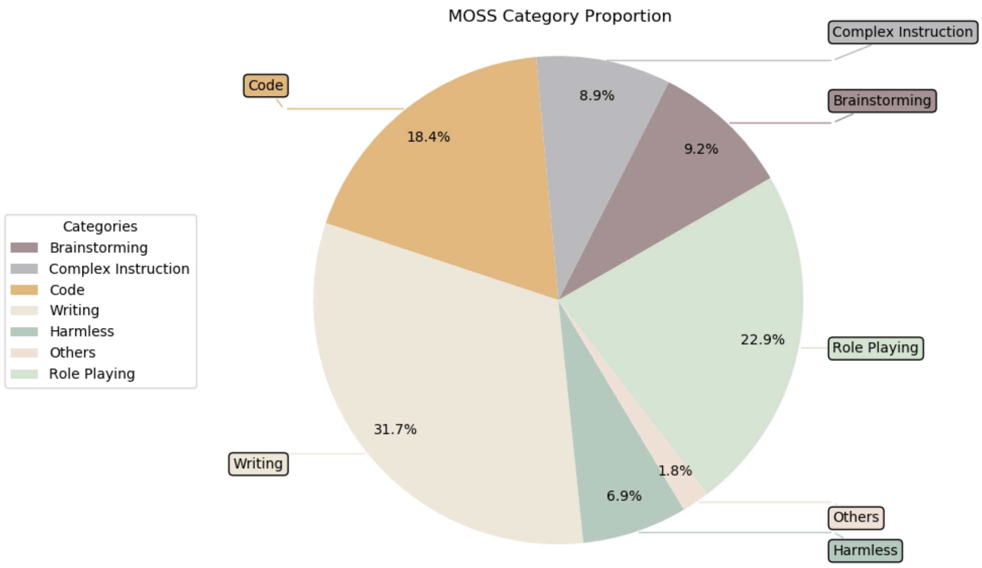

Figure 7: Moss Data Structure Distribution

Table 7: Introduction to Unified Datasets

Dataset	Emotion Categories	Data Scale
OCEMOTION	sadness, happiness, disgust, anger, like, surprise, fear	35693
Chinese Caption Sentiment Dataset	neutral, happiness, sadness, disgust, anger, surprise, fear	6583
Smp2020WECT	neutral, happy, angry, sad, fear, surprise	34768
Smp2020EWECT Covid	neutral, happy, angry, sad, fear, surprise	13606
Emotion Corpus Microblog	happiness, sadness, disgust, like, fear, surprise, anger	39661
NLPCC2014 (whole sentence)	happiness, sadness, disgust, like, fear, surprise, anger	47283
NLPCC2014	happiness, sadness, disgust, like, fear, surprise, anger	24715
NLPCC2013 (whole sentence)	happiness, sadness, disgust, like, fear, surprise, anger	10274
NLPCC2013	happiness, sadness, disgust, like, fear, surprise, anger	7915