Fine-Grained Sentiment Labeling with Retrieval-Augmented LLMs: A Computational Approach

Anonymous ACL submission

Abstract

Sentiment analysis has evolved from coarse-grained classification (e.g., positive/negative) to fine-grained dimensional labeling (e.g., valence and arousal), yet accurately scoring emotional intensity remains challenging. While large language models (LLMs) show promise, their direct application suffers from domain adaptation issues and inconsistency in fine-grained annotations. To address this, we propose a retrieval-augmented multi-LLM ensemble framework that combines dynamic knowledge retrieval with weighted model collaboration. Experiments on the Chinese EmoBank corpus demonstrate significant improvements of our proposed model on arousal labeling compared to zero-shot LLM baselines. In this paper, we introduce an augmented generation strategy based on cross-lingual (Chinese-English) retrieval and propose a data-driven weighting mechanism to assign model importance based on task-specific performance, providing a repeatable open source implementation. The results highlight the key role of retrieval-augmented generation in fine-grained sentiment analysis. Our work provides a scalable solution for real-world sentiment analysis applications such as social media, product reviews, and opinion monitoring, and lays a foundation for future extensions of multi-modal affective computing.

1 Introduction

Fine-grained sentiment annotation is a key task in the field of natural language processing, which is widely used in public opinion analysis, market analysis, financial forecasting and other scenarios (Wankhade et al., 2022). Different from traditional positive/negative emotion recognition, fine-grained emotion modeling usually relies on the measurement system of emotional dimensions, especially the two dimensions of valence (positive/negative) and arousal (intensity) (Kensinger, 2004). Compared with polarity labels, the recognition of arousal is more dependent on context understanding and emotional expression details, which faces challenges such as strong subjectivity, cross-cultural differences, and computational complexity.

With the wide application of Large language models (LLMs) in natural language understanding tasks, LLMs show powerful zero-shot and few-shot learning capabilities in the field of sentiment analysis (Krugmann and Hartmann, 2024; Zhang et al., 2023b). Despite the superior performance of LLMs in sentiment classification tasks, there are still shortcomings when LLMs are directly used for fine-grained sentiment labeling, especially in the arousal dimension. This problem mainly stems from the deviation between the pretraining objective of LLMs and the fine-grained emotion recognition task, resulting in the lack of stability and context sensitivity of the model in terms of intensity scores.

To solve the above problems, this paper proposes a sentiment annotation framework based on Retrieval Augmented Generation (RAG) and multi-model collaboration. This method uses the semantic similarity retrieval mechanism to dynamically extract similar expressions from the annotated corpus as reference, and constructs prompts with the text to be annotated to enhance the model's understanding of context and emotional reference points. LLMs (such as LLaMA3, Gemma2, GPT-4o) are introduced to form an "annotation committee", which independently scores the input text. And the optimal weights combination through linear regression for LLMs is trained to form the final annotation result. We conduct a systematic evaluation on the professional Chinese emotion corpus Chinese EmoBank (Lee et al., 2022). The experimental results show that the proposed method achieves high human agreement on both valence and arousal dimensions, and significantly outperforms the existing zero-shot baseline methods.

2 Related Work

Sentiment analysis research has long focused on the identification of sentiment polarity, such as positive, negative, and neutral classification. However, in recent years, researchers have gradually paid attention to more fine-grained emotion modeling methods, especially the valence and arousal label systems widely used in two-dimensional emotion framework (Kensinger, 2004). Compared with traditional labels, this continuous dimension modeling can more accurately capture the emotional intensity and subtle differences, especially suitable for user comments, public opinion analysis and other scenarios that need to identify emotional fluctuations. Although some studies have constructed corresponding sentiment dataset or manually annotated corpus (Buechel and Hahn, 2022; Xu et al., 2022). However, high-quality fine-grained labels rely on a lot of labor costs, which limits their scalability on large-scale data.

With the development of large language models, the GPT series, LLaMA and other models have shown significant context understanding and generalization ability in text classification and generation tasks. Studies have shown that LLMs can achieve comparable or even better performance than supervised models in zero-shot or few-shot settings on common datasets such as IMDb, Yelp, and Twitter (Krugmann and Hartmann, 2024). However, the vast majority of works focus on three-class classification or polarity judgment, and few studies focus on continuous label prediction of dimensions such as arousal. Some studies try to guide the model output scoring through prompt engineering or a small number of examples, but there are problems of unstable output and high subjectivity (Fatemi and Hu, 2023).

To make up for the lack of task adaptability of large models, retrieval-augmented generation (RAG) methods have been introduced into open-domain question answering, code generation and other tasks (Lewis et al., 2020; Siriwardhana et al., 2023). RAG improves the domain adaptability and output interpretability by retrieving relevant background documents before generation and providing them to the LLMs together with the input. In the field of sentiment analysis, the exploration of RAG technology is still in its infancy, and the existing work mostly focuses on positive and negative polarity recognition in financial texts (Zhang et al., 2023a).

On the other hand, multi-model integration strategy has been widely verified in improving annotation consistency. The "wisdom of crowdsourcing" theory states that the integrated results of multiple independent evaluations can significantly reduce individual bias (Welinder et al., 2010). Previous studies have attempted to use multiple LLM voting mechanisms to improve consistency and robustness of model labeling (Sun et al., 2023). In this study, the RAG semantic retrieval mechanism and the multi-model integration strategy are combined to further improve the adaptability and accuracy of the model in the dimensional emotion modeling task.

3 Methodology

3.1 Task definition

The goal of this paper is to label a given Chinese text with fine-grained emotion and output its scores on two dimensions: valence and arousal. Referring to the format of EmoBank (Buechel and Hahn, 2022), we mapped the emotional dimension into a 5-level Likert scoring system, corresponding to different degrees of emotional polarity and intensity. The model outputs a scalar rating on these two continuous emotion dimensions based on the input text content. This task can be modeled as two parallel regression problems, or can be transformed into a multi-classification problem. In this paper, we evaluate the performance of the model in both regression and classification metrics.

3.2 Framework overview

The overall framework is shown in 1, which combines RAG mechanism with multi-model voting fusion strategy. The input text firstly extracts similar sentences and their labels through the semantic retrieval module as the basis for prompt construction, and then combines the input text to form the prompt, which is input into the large language model for scoring, and then weighted and fused with the zero-shot results of the annotation committee to form the final prediction output. Note that when the semantic retrieval module fails to retrieve similar statements, then the weighted scoring is performed only by the results of the annotation committee.

3.3 Semantic retrieval strategy

We apply BGE (BAAI General Embedding) (Xiao et al., 2023) to encode the Chinese and English

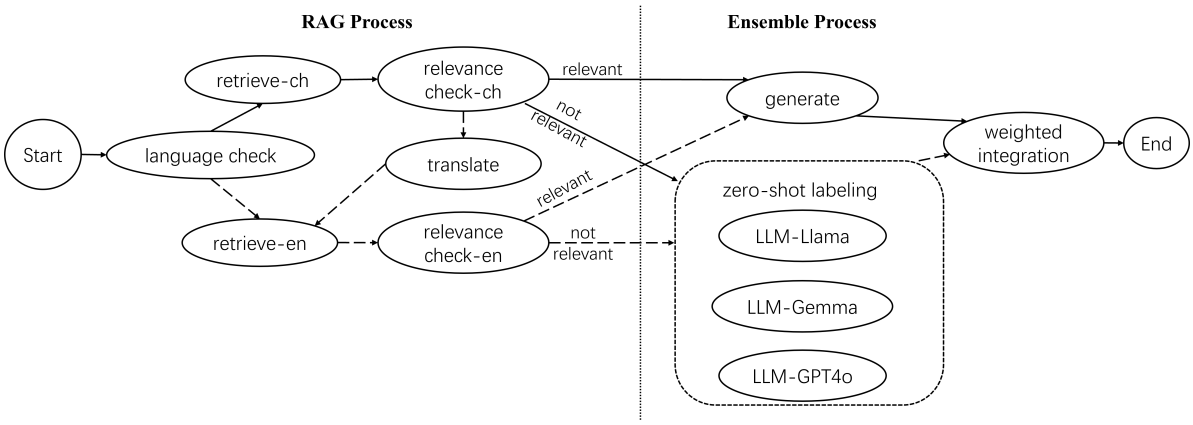

Figure 1: RAG-based Multi-LLM Integrated Sentiment Annotation Framework

Emobank corpus to obtain the semantic vector of each text sentence, and use Faiss (Facebook AI Similarity Search) (Douze et al., 2024) to build the index to achieve efficient similarity search. In the prediction stage, the input text is encoded and the top-5 similar statements are searched in the index, which is sorted and filtered based on cosine similarity. If the high similarity results are not found in the Chinese corpus, the system will translate the input text into English and retrieve it in the English corpus to form cross-language support.

3.4 Multi-model labeling and integration

We choose three mainstream LLM models: LLaMA3, Gemma2 and GPT-4o to form the emotion annotation committee. Among them, GPT-4o is not only responsible for score generation based on similar sentence retrieved, but also a member of the labeling committee, and outputs the predicted values of valence and arousal independently with the other two classical LLMs. To improve the overall performance, linear regression models are used to perform supervised learning on the outputs of each sub-model, and the optimal combination of weights is trained to minimize the prediction error and generate the final ensemble model annotation result.

4 Experiment and Result

4.1 Data and experimental setup

The experiments are carried out on the Chinese EmoBank dataset. This corpus contains a large number of annotated Chinese texts, and each sentence is manually rated in two dimensions: valence and arousal. The original score uses a 9-level Likert scale (Lee et al., 2022). To keep consistent with the annotation level of the English Emobank corpus and improve the stability of the model, we map the Chinese corpus ratings to a 5-level scoring system. The dataset is divided as follows: 80% to construct the external retrieval knowledge base (for RAG semantic enhancement), 10% as the training set for integrated weights learning, and 10% as the test set.

4.2 Model comparison

We set the following comparison models: (1) direct zero-shot labeling for a single LLM model; (2) simple average ensemble of three independent models (LLaMA3, Gemma2, GPT-4o); (3) weighted fusion after introducing Chinese RAG module (C-RAG); (5) expanded to weighted fusion of Chinese-English bilingual RAG module (CE-RAG) and three LLMs.

It can be seen from Table 1 that multi-model ensemble already significantly outperforms single model in the valence labeling task, and further introduction of RAG module has limited performance improvement, indicating that valence mainly relies on text semantics itself. In the arousal labeling task, the RAG module significantly improved the stability and accuracy of the model (MAE decreased by 19.3%), which verified the complementary role of external reference information on emotion intensity perception.

In addition, although we also tried to introduce a Chinese-English mixed retrieval (CE-RAG) module to try to extend the cross-context reference coverage, the experimental results show that its performance improvement is limited. Specifically, valence labeling results are on par with Chinese RAG, with only a slight error decrease in arousal dimension. This phenomenon may be due to the fact that the Chinese retrieval database has covered

Model	Valence MAE	Valence ACC	Arousal MAE	Arousal ACC
LLaMA3	0.62	0.57	0.84	0.33
Gemma2	0.56	0.43	0.75	0.38
GPT-4o	0.57	0.58	0.62	0.46
3LLMs	0.37	**0.68**	0.57	0.42
C-RAG+3LLMs	**0.33**	0.51	0.46	**0.50**
CE-RAG+3LLMs	**0.33**	0.51	**0.45**	**0.50**

Table 1: Performance comparison of different model schemes on valence and arousal labeling tasks

most of the effective expressions, and the English corpus has deviation in cultural semantic mapping, which may weaken the reference value of the reference.

4.3 Model weights distribution

We further analyze the regression weights of each module in the weighted ensemble model with Chinese RAG introduced in the two labeling tasks, and the results are as follows.

Module	Valence Weights	Arousal Weights
LLaMA3	0.145	0.010
Gemma2	0.146	0.168
GPT-4o	0.306	0.212
C-RAG	0.029	0.111
Intercept	1.094	1.399

Table 2: Regression coefficients for each module in the weighted ensemble model

As shown in Table 2, the RAG module has a significantly higher weight in the arousal labeling task than valence (0.111 vs 0.029), indicating that external reference information plays a greater role in intensity judgment. In contrast, valence depends more on the language model's own understanding of emotional polarity, while arousal involves the perception of emotional intensity, and needs to use similar emotional precedents as reference to improve the stability and consistency of prediction.

5 Conclusion

This paper proposes a fine-grained emotion annotation method that combines semantic retrieval-augmented generation and multi-model collaboration mechanism, and constructs an end-to-end annotation framework for two emotional dimensions: valence and arousal. This method uses the BGE semantic embedding and Faiss index to construct a retrieval module, and guides the large language model to generate context-aware scoring results. At the same time, the weighted integration strategy of multiple models is introduced to effectively improve the consistency and robustness of labeling process. Experimental results in the Chinese EmoBank corpus show that the proposed method performs better in the valence dimension and achieves a significant performance improvement in the arousal dimension, which verifies the key role of the retrieval-augmented generation mechanism in the prediction of emotion intensity.

Although the effectiveness of the method is initially verified, there are still problems of insufficient cross-language transfer ability and limited valence improvement. Future research can further explore culturally sensitive semantic retrieval strategies and construct an aligned emotion annotation database covering multiple languages and scenes. In addition, the introduction of speech, image and other modalities to build a multi-modal fusion framework will also provide a new direction for complex emotion recognition tasks.

Limitations

Although the proposed RAG enhanced multi-model emotion labeling method achieves promising experimental results on both valence and arousal dimensions, it still has several limitations. First, the cross-lingual and cross-cultural generalization ability of the model is not sufficient. Although we designed the Chinese-English hybrid retrieval mechanism to expand the semantic reference range, experiments show that cross-lingual reference fail to bring the expected performance improvement in all scenarios due to the possible introduction of semantic drift in the translation process and the cultural differences in emotional expression. Therefore, the subsequent research can try to construct a multilingual aligned sentiment corpus, and introduce the language-aware template and similarity threshold adjustment mechanism in the model design to

improve the retrieval quality.

Second, current emotion modeling still mainly relies on text-based static context. When dealing with sentences with strong subjectivity or complex emotional expression (such as double negation, rhetorical question, and mixed emotional expression), the model prediction results still need to be further explored. To this end, more fine-grained modeling of emotional components (such as intention, tone, contextual emotional consistency) can be considered in the future.

In addition, our method is only verified on the Chinese EmoBank corpus, which fails to cover the differences of emotion distribution in multiple domains, such as social media, financial reviews, and psychological counseling. Subsequently, the model can be extended to multi-scene data, and the domain adaptation mechanism can be introduced to enhance the transferability and generalizability of the model.

Finally, this paper still focuses on single-modal text input. In real applications, emotions are usually accompanied by multi-modal cues such as speech, facial expression or images. Therefore, multi-modal fusion will be an important direction to improve the accuracy of emotion understanding. Combined with image emotion analysis or speech rhythm intensity modeling, the application ability of the proposed method in complex human-computer interaction scenarios can be further expanded.

Acknowledgments

We would like to thank all anonymous reviewers for their valuable comments and suggestions. In addition, we thank the developers of the open source projects used in this paper, such as the BGE embedding model developed by Beijing Zhiyuan Research Institute (BAAI), and the Faiss vector search system developed by Facebook AI Research. Also, we would like to thank the open source community and platforms such as HuggingFace and Ollama for their continuous contributions to the LLaMA, Gemma, GPT series models, which enabled the model development of this research. In particular, the Chinese EmoBank corpus used in this paper was constructed by the Natural Language Processing Laboratory of Yuan Ze University, and we hereby express our gratitude to the original data provider.

References

Sven Buechel and Udo Hahn. 2022. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Sorouralsadat Fatemi and Yuheng Hu. 2023. A comparative analysis of fine-tuned llms and few-shot learning of llms for financial sentiment analysis. *arXiv preprint arXiv:2312.08725*.

Elizabeth A Kensinger. 2004. Remembering emotional experiences: The contribution of valence and arousal. *Reviews in the Neurosciences*, 15(4):241–252.

Jan Ole Krugmann and Jochen Hartmann. 2024. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–18.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. Sentiment analysis through llm negotiations. *arXiv preprint arXiv:2311.01876*.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. 2010. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Xu Xu, Jiayin Li, and Huilin Chen. 2022. Valence and arousal ratings for 11,310 simplified chinese words. *Behavior research methods*, 54(1):26–41.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023a. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 349–356.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023b. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

A Appendix

To improve annotation consistency and model controllability, we design two types of standardized prompt templates, one for context-enhanced input in the RAG generator, and the other for direct scoring tasks with independent LLMs without retrieval. The template structure is unified and clearly indicates the task goal and output format of the model, as shown below.

A.1 Retrieval-Augmented Emotion Labeling Prompt (RAG Generator)

System:
You are an expert in text sentiment analysis.
You are responsible for assigning valence and arousal scores for the text input by the user. And you are using the 5-point Likert scale to assign scores.
Valence is the level of emotional polarity, and it ranges from 1 to 5, in which 1 represents the most negative emotion and 5 represents the most positive emotion.
Arousal is the level of emotional intensity, and it also ranges from 1 to 5, in which 1 represents the most calm emotion and 5 represents the most strong emotion.
Please refer to the documents that have been labeled to assign the valence and arousal scores to the text input by the user, where Valence_Mean and Arousal_Mean represent the valence and arousal scores in the graded documents respectively.
Provide the scores as a JSON output with keys 'Valence' and 'Arousal' and no preamble or explanation.
User:
Graded documents: {document}
User input: {text}

A.2 Zero-shot Emotion Labeling Prompt

System:
You are an expert in text sentiment analysis.
You are responsible for assigning valence and arousal scores for the Chinese text input by the user. And you are using the 5-point Likert scale to assign scores.
Valence is the level of emotional polarity, and it ranges from 1 to 5, in which 1 represents the most negative emotion and 5 represents the most positive emotion.
Arousal is the level of emotional intensity, and it also ranges from 1 to 5, in which 1 represents the most calm emotion and 5 represents the most strong emotion.
Provide the scores as a JSON output with keys 'Valence' and 'Arousal' and no preamble or explanation.
User: {text}