RESEMO: A Benchmark Chinese Dataset for Studying Responsive Emotion from Social Media Content

Bo Hu[1], **Meng Zhang**[1], **Chenfei Xie**[1], **Yuanhe Tian**[2], **Yan Song**[1*], **Zhendong Mao**[1]
[1]University of Science and Technology of China, Hefei, China
[2]University of Washington, Seattle
hubo@ustc.edu.cn, zhangm@mail.ustc.edu.cn, xiechenfei@mail.ustc.edu.cn
yhtian@uw.edu, clksong@gmail.com, zdmao@ustc.edu.cn

Abstract

On social media platforms, users' emotions are triggered when they encounter particular content from other users, where such emotions are different from those that spontaneously emerged, owing to the "responsive" nature. Analyzing the aforementioned responsive emotions from user interactions is a task of significant importance for understanding human cognition, the mechanisms of emotion generation, behavior on the Internet, etc. Performing the task with artificial intelligence generally requires human-annotated data to help train a well-performing system, while existing data resources do not cover this specific area, with none of them focusing on responsive emotion analysis. In this paper, we propose a Chinese dataset named RESEMO for responsive emotion analysis, including 3,813 posts with 68,781 comments collected from Weibo, the largest social media platform in China. RESEMO contains three types of human annotations with respect to responsive emotions, namely, responsive relationship, responsive emotion cause, and responsive emotion category. Moreover, to test this dataset, we build large language model (LLM) baseline methods for responsive relation extraction, responsive emotion cause extraction, and responsive emotion detection, which show the potential of the proposed RESEMO being a benchmark for future studies on responsive emotions.

1 Introduction

With the rise of social media and increasing user activity on online platforms, many researchers focus on social media text processing (Tang et al., 2018; Wang et al., 2019; Hruska and Maresova, 2020; Nie et al., 2020). On social media, people tend to express their sentiments or emotions in their generated content and thus attract much attention from existing studies (Demszky et al., 2020; Chen et al., 2020; Saha et al., 2020; Wang et al., 2021; Qin et al., 2022; Tian et al., 2023a, 2024). Different from emotions delivered in normal scenarios mostly driven by various causal events, interactions on social media play the dominant role in pushing individuals to develop emotions with respect to the content they engage with (Gaind et al., 2019), where such emotions are responsive to the previous user posts and comments. Figure 1 provides an example of a blog post with its comments, where the comment *id:1-3* is semantically responsive to both *id:1-1* and *id:1-2*, and expresses an emotion of "cynicism", with corresponding word-level emotion cause marked in red. Apparently, to analyze such responsive emotions and emotion causes, it is beneficial to mine the responsive relationships conveyed in the context of social interactions.

However, responsive relationships are often implicit and intricate on social media. Users typically browse a blog post and its accompanying comments before composing their own comments. Throughout this process, they form opinions based on the overall impressions gathered from the blog and comments they have browsed. As a result, they may respond to multiple comments or the blog post itself, creating implicit one-to-many responsive relationships. Additionally, there can be a time delay between a user's comment and the post or comments they are responding to due to the asynchronous nature of social media interactions. These unique characteristics present challenges in mining response relationships. Considering that currently there are no existing models or systems capable of addressing these challenges, one should prepare such resources in order to perform research related to this topic. Therefore, collecting data with interactive responsive emotions from social platforms and constructing a dataset with explicit annotations for responsive relationships, as well as emotion categories and causes has become a vital step for later studies on responsive emotion analysis.

*Corresponding author

Findings of the Association for Computational Linguistics ACL 2024, pages 16375–16387

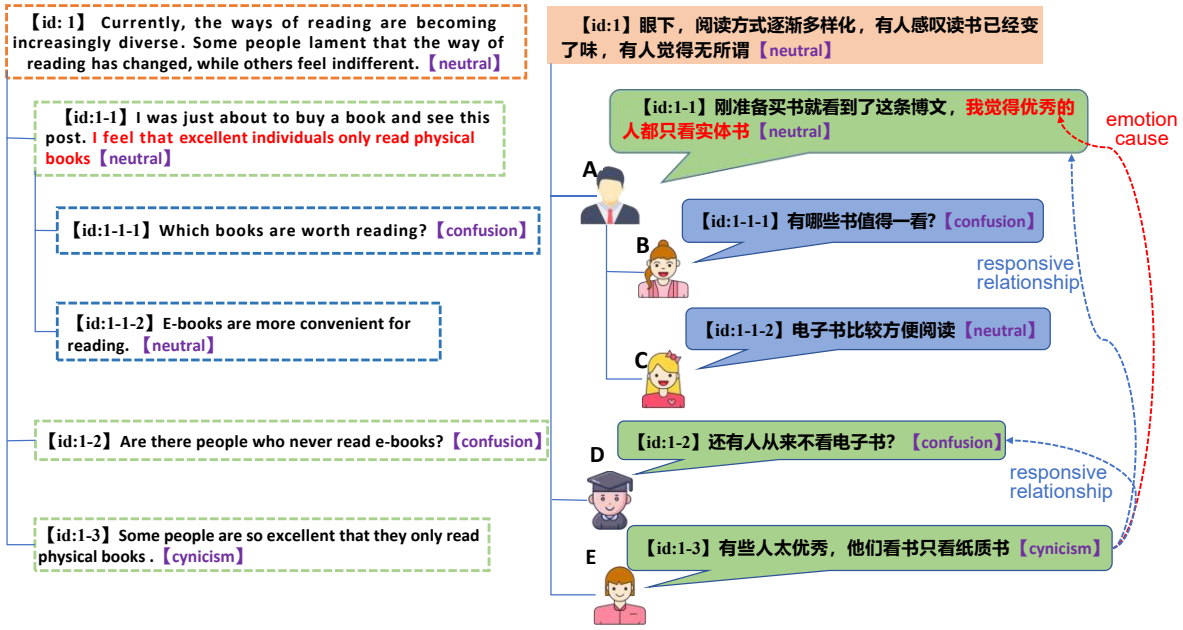

Figure 1: An example of a blog post with its comments. In Weibo, the blog post and accompanying comments are organized in a 3-level tree structure, including the blog post as the root node, and two levels of comments, where the comments directly respond to the blog post from the first-level comments. These first-level comments can also be further replied to, forming the second-level comments. Note that in this social platform, all subsequent interactions and replies within second-level comments are organized on the same level. In the RESEMO dataset, comments with the same parent node are assigned with indices in chronological order from oldest to newest. The dataset includes three types of annotations, i.e. responsive relation, responsive emotion cause and responsive emotion category. For example, given the annotation target comment *id:1-3*, it is semantically responsive to *id:1-1* and *id:1-2*, thus its responsive relationship is *id:1-1* and *id:1-2*, while the responsive emotion cause is marked in red, which acts as the reason behind the conveyed emotion of *id:1-3*. The responsive emotion category of comment *id:1-3* is labeled as *Cynicism*. The dataset only contains Chinese and the English translation is only given for reference.

Despite the existence of datasets (Li et al., 2017; Poria et al., 2018; Demszky et al., 2020; Chen et al., 2022) that provide emotion and emotion cause annotations from conversation corpus or social media posts, they do not fully consider the responsive relationship within the context. As a result, they do not adequately capture the implicit one-to-many and temporal delay characteristics of responsive relationships mentioned earlier. For instance, datasets like (Poria et al., 2018; Chen et al., 2022) gather dialog corpus from TV series, where each utterance mainly serves as a direct response to the previous one, representing a one-to-one and instant responsive relationship. Additionally, (Demszky et al., 2020) collect data from an English social platform, but their annotation only focuses on emotion categories, disregarding the responsive relationships.

In this paper, we introduce a novel Chinese dataset called RESEMO that aims to facilitate an in-depth analysis of responsive emotions in social media text. The raw data for RESEMO is collected from Weibo, the largest social media platform in China. As shown in Figure 1, for a given blog post with accompanying comments, we annotate each comment with three types of labels: responsive relationship, responsive emotion cause, and responsive emotion category. The responsive relationship label indicates which preceding comments or the blog post itself the current comment is responding to, allowing for the possibility of a one-to-many relationship. The responsive emotion cause label identifies a specific word-level text span that acts as the reason behind the conveyed emotion in the current comment. Lastly, the responsive emotion category label classifies the specific type of emotion expressed in the current comment. Using the annotated data, several baseline methods based on large language model (LLM) are employed for tasks such as responsive relationship extraction, responsive emotion cause extraction, and responsive emotion detection. The experimental results demonstrate the potential benefits of mining the responsive relationship, as it proves advantageous for both emotion cause extraction and responsive

emotion detection. Furthermore, these results validate the suitability of RESEMO as a benchmark dataset for future studies on responsive emotions.

2 The RESEMO Dataset

In this section, we first introduce how we collect our data, then we introduce the annotation guidelines, and annotation process and finally discuss properties of our dataset.

2.1 Data Collection

On Weibo, users can share their personal experiences, express their emotions, and engage in discussions. The platform allows users to interact with blog posts by posting comments directly underneath them, resulting in first-level comments. These first-level comments can also be further replied to, forming second-level comments. Note that in this social platform, all subsequent interactions and replies within second-level comments are organized on the same level.

To collect the data for our dataset, we conduct web crawling on posts and comments from selected accounts, which are chosen based on their popularity in various fields, ensuring that their posts receive a significant number of user comments. The topics of the collected posts encompass a wide range of categories, including entertainment, the stock market, digital technology, sports, and daily life. Prior to analysis, we take precautions to remove any personal information such as names, ages, genders, and other privacy-related details. In total, we initially collected 17,915 posts along with 317,975 comments.

We then apply the following post-processing. We first select post blogs with a comment count in the range of $[15, 40]$. This is because a small number of comments suggests limited user engagement, while a large number of comments increases the difficulty of annotation, thereby compromising the quality of the dataset. We then filter out redundant posts and comments to enhance the diversity of our dataset. Finally, our dataset consists of 3,813 posts and 68,781 comments, with 50,738 and 18,043 first- and second-level comments, respectively.

2.2 Annotation Guideline

Our dataset has three types of annotations, including responsive relationship, responsive emotion cause and responsive emotion category. In this section, we provide an overview of the annotation guidelines. For more detailed information, please refer to Appendix C. As shown in Figure 1, for a given blog post with accompanying comments organized chronologically, annotators are instructed to label each comment one by one with all three types of annotations.

2.2.1 Responsive Relationship

The responsive relationship label indicates which preceding comments or the blog post itself the current comment is responding to, allowing for the possibility of a one-to-many relationship. For instance, in Figure 1, comment *id:1-3* is not a direct reply to comments *id:1-1* and *id:1-2*. However, from a semantic perspective, it is evident that comment *id:1-3* serves as a response to both comments *id:1-1* and *id:1-2*. In this scenario, both comments *id:1-1* and *id:1-2* are labeled as the utterances that comment *id:1-3* responds to.

2.2.2 Responsive Emotion Cause

The responsive emotion cause label identifies a specific word-level text span that acts as the reason behind the conveyed emotion in the current comment. Given that a user's responsive emotion stems from interactions with other users, it is anticipated that this emotion cause originates from the blog posts or comments to which the current comment is responding. Therefore, we only need to annotate word-level text spans as emotion causes within sentences with such response relationships. Figure 1 shows that the responsive emotion cause of *id:1-3* lies in *id:1-1*, to which *id:1-3* responds.

2.2.3 Responsive Emotion Category

In this dataset, we need to annotate the emotion category of post blogs and comments. In the literature, Robert Plutchik proposes the famous Wheel of Emotions (Plutchik and Kellerman, 1980), categorizing emotions into eight basic types[1], while the later on studies (Rashkin et al., 2018; Shen et al., 2020) suggest that a finer granularity of emotional annotations can improve in emotional analysis. Therefore, following those previous works, we expand the original eight basic emotions by incorporating another eight emotion categories commonly found on social media platforms and finally have 16 emotion categories, including six positive emotions, nine negative emotions, and one neutral emotion, as elaborated in Table 1 with explanations

[1]The eight basic emotions are anger, fear, sadness, disgust, surprise, anticipation, trust, and joy.

ID	Label	Emotion	Explanations	Examples
1	Positive	Anticipation	A feeling of excitement and expectation.	Hope for a world without war.
2		Gratitude	A feeling of thankfulness.	Thank you for your compliment.
3		Joy	A feeling of happiness and delight.	Wow! I win the lottery!
4		Pride	A feeling of satisfaction and self-respect.	My hometown is truly beautiful!
5		Surprise	A sudden feeling of astonishment or disbelief.	Oh! This is so unbelievable!
6		Trust	A belief and appreciation for others.	The trophy belongs to you!
7	Negative	Anger	A feeling of intense displeasure or hostility.	Shut up!
8		Compassion	A feeling of empathy and care towards others.	This little cat is so pitiful.
9		Confusion	A state of being bewildered or perplexed.	So what's the answer then?
10		Cynicism	A skeptical and distrustful attitude.	Your ability to lie is great.
11		Disappointment	A feeling of dissatisfaction or letdown.	I indeed overestimated you.
12		Disgust	A strong feeling of revulsion or repulsion.	I really dislike people who litter.
13		Fear	A strong emotion of apprehension or dread.	I can't imagine the consequences.
14		Sadness	A deep feeling of sorrow or unhappiness.	I didn't pass the exam again.
15		Shame	A feeling of embarrassment or guilt.	I'm too ashamed to face you.
16	Neutral	Neutral	A state of being neither positive nor negative.	Please keep a healthy habit.

Table 1: An illustration of the 16 responsive emotion categories used in our dataset, along with corresponding explanations and examples.

ID	Tasks	Mertics	Score
1	RE	Kappa	0.5119
2	RR	F1	0.8452
3	REC	ROUGE-1	0.7203
		ROUGE-2	0.6524
		ROUGE-L	0.6539

Table 2: Inter-annotator agreement measurements of different annotation tasks. "RR", "RE" and "REC" stand for responsive relationship, responsive emotion category and responsive emotion cause.

# of Post	3,813
# of First-level Comments	50,738
# of Second-level Comments	18,043
# of Responsive Relations	120,253
# of Emotion Causes	68,781
# of Comments Per Post	27.52
Avg. # of Chars per Post	163.07
Avg. # of Chars per Comments	17.42
Avg. # of Chars per Emotion Cause	87.02
# of Comments with One RR	46,902
# of Comments with Two RR	10,713
# of Comments with Three RR	4,667
# of Comments with More than Three RR	6,499

Table 3: The statistics of RESEMO. "RR" denotes responsive relationship.

and examples. The positive emotions include *Anticipation*, *Gratitude*, *Joy*, *Pride*, *Surprise*, *Trust*. The negative emotions include *Anger*, *Compassion*, *Confusion*, *Cynicism*, *Disappointment*, *Disgust*, *Fear*, *Sadness*, *Shame*, and the neutral emotion is *Neutral*. It is worth noting that when labeling the emotional category, the annotator should take the context of the responsive relationship into consideration. For example, in Figure 1, the comment *id:1-3* itself generally expresses the positive emotion, but in the context of responsive relationships, it has an ironic meaning and express a negative emotion, which should be annotated as *Cynicism*.

2.2.4 Annotation Process

We recruited 10 annotators who are Chinese native speakers, and they all use Weibo frequently. We adopt an open-source annotation software[2] for annotation. We inform the annotators about the purpose of the RESEMO, emphasizing its intended use for scientific research purposes. Throughout the annotation process, we ensure the confidentiality of annotators' information and offer competitive compensation aligned with their workload. To ensure the quality of annotation, we provide annotators with detailed annotation guidelines, including clear instructions for each annotation task and operational guidance for the annotation platform.

To assess the annotation quality, we randomly divide our 10 annotators into 5 groups, and each group has 2 annotators. We then randomly selected 150 posts with accompanying comments for anno-

[2]Label Studio, https://labelstud.io/.

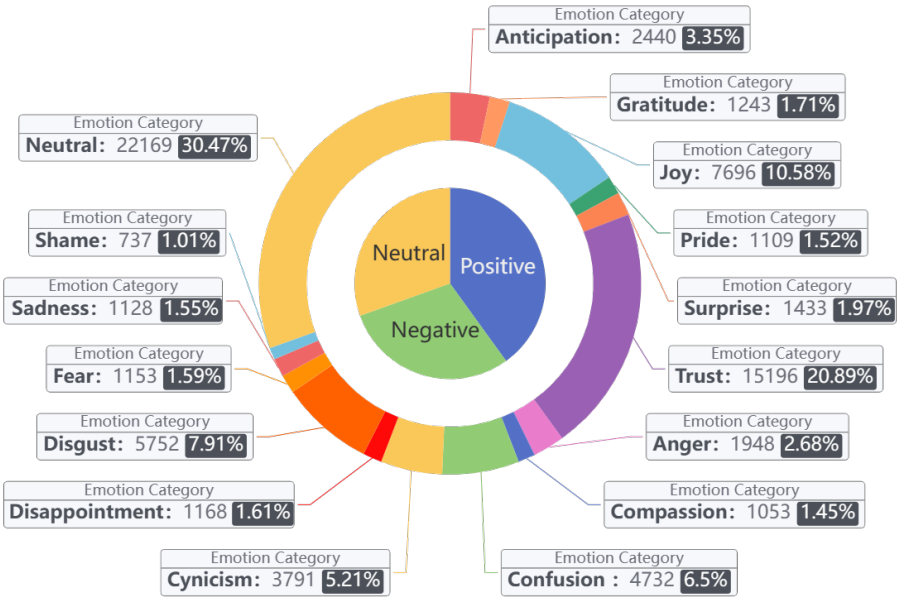

Figure 2: The distribution of 16 emotion categories in our dataset.

tation. Each group is assigned 30 posts, and each member of the group independently annotates the assigned posts. We then compute inter-annotator agreement by different measurements and report the scores in Table 2.

For the annotations of the responsive emotion category, we follow existing studies (Liu et al., 2021; Tian et al., 2022b,a) to employ Cohen's Kappa as the evaluation metric (McHugh, 2012). When compared to the EmotionLines dataset with 7 categories of emotions achieving a Kappa score of 0.34 (Chen et al., 2018), and the MELD dataset also with 7 categories of emotions achieving a Kappa score of 0.43 (Poria et al., 2018), our dataset obtains a higher Kappa score 0.51 with more categories of emotions, which validates the quality of our annotations.

For the annotations of responsive relationship, each comment may have multiple responsive relationships with preceding comments or the blog post. Following the work (Deleger et al., 2012; Brandsen et al., 2020), we treat the labels made by one annotator as the ground truth, and treat the other annotator's labels as the predicted output to calculate the F-score to measure the inter-annotator agreement, which is around 0.85 in our dataset.

For the annotations of responsive emotion cause, the ROUGE scores (Lin, 2004) can be employed to measure the inter-annotator agreement. In particular, ROUGE-1 measures the overlap of unigrams between the two annotators' labels. ROUGE-2 measures the overlap of bigrams between the two annotators' labels. ROUGE-L focuses on the longest common subsequence between the two annotators' labels. The ROUGE scores are reported in Table 2. Based on the above analysis, the results of inter-annotator agreement with different measurements are consistent, which validates the annotation quality of our dataset.

2.3 The Properties of the Dataset

Table 3 provides the basic statistics of RESEMO. We also perform a statistical analysis of our dataset. In Figure 2, we present the occurrences of 16 different emotions. We observe that the distribution of positive, negative, and neutral emotions is 40.02%, 29.51%, and 30.47% respectively. However, we notice that the distribution of fine-grained emotions is quite uneven. This is to be expected, as users on social platforms primarily rely on text to express their emotions. Emotions such as joy, trust, and disgust are generally easier to convey through text compared to other subtle emotions.

Table 4 presents the distribution of responsive relationships within our dataset. We observe that for first-level comments, 57.26% of them are in response to the original posts, while the remaining 42.74% are in response to other first-level comments. In the case of second-level comments, 19.41% of them respond to the posts, 49.49% respond to the parent first-level comments and the remaining 31.10% respond to other first-level comments. These results demonstrate the diverse nature of responsive relationships in Weibo, further em-

Annotated Objects	Responsive Relationship	Amount	Proportion
First-level Comments	Posts	**50,611**	**57.26%**
	Other first-level comments	37,773	42.74%
Second-level Comments	Posts	5,800	19.41%
	Affiliated first-level comments	**14,792**	**49.49%**
	Second-level comments under affiliated comments	9,294	31.10%

Table 4: The distribution of responsive relationships

phasizing the importance of studying responsive relationships in the context of responsive emotion.

We further compare RESEMO with existing datasets for emotion analysis, where Table 5 presents the comparison results, and RESEMO shows several advantages as follows.

- RESEMO presents a collection of fine-grained emotion categories. In Table 5, we have compiled a set of 13 existing datasets focused on emotions. Our dataset stands out by offering 16 distinct emotion categories, demonstrating a higher level of granularity compared to the majority of these datasets. It is worth noting that only two datasets, namely GoEmotions and EmpatheticDialogues, feature a greater number of categories than ours. As a result, our dataset presents a competitive range of emotion categories, enabling a nuanced representation of emotions and contributing to the advancement of emotion analysis and understanding.

- RESEMO comprises a rich collection of responsive relationships, documenting user interactions and responses, a feature that distinguishes it from previous datasets. These responsive relationships are crucial, as leveraging this information can significantly enhance the accuracy of emotion analysis tasks. However, they have been overlooked in previous datasets. User interactions on social media inherently involve responsive behavior, where they naturally express opinions and emotions in a responsive manner. The responsive relationships fundamentally capture such characterization and will be beneficial to emotion analysis tasks on social media.

3 Experiments

The RESEMO dataset serves as a foundational resource for analyzing emotions in social media. To assess the effectiveness of this dataset, We propose LLM-based baseline models for tasks, such as responsive emotion category detection, responsive emotion cause extraction, and responsive relationship extraction.

3.1 Models

In recent years, LLMs have led a significant revolution in the field of Natural Language Processing (NLP) (Devlin et al., 2018; Diao et al., 2020; Song et al., 2021; Gan et al., 2023; Peng et al., 2023; Kosinski, 2023; Tian et al., 2023b). With extensive pre-training on massive datasets and numerous parameters, LLMs demonstrate exceptional performance in tasks involving language understanding and text generation (Dong et al., 2022). For our experiments, we use two well-known LLMs, i.e., ChatGPT (OpenAI, 2022) and Chinese-Alpaca-2 (Cui et al., 2023).

For the experiment with ChatGPT, we utilize the API provided by OpenAI and conduct experiments with few-shot learning. For the experiment with the Chinese-Alpaca-2, since it is publicly available, allow us to run it locally. We adopt the Chinese-Alpaca-2-13b version[3], employing the default hyper-parameter settings and conduct experiments involving 0-shot learning and supervised instruction fine-tuning. For the fine-tuning experiment, we train the model for three epochs on two NVIDIA A800 GPUs, with introduction prompts illustrated in Appendix A.

3.2 Settings

The data is divided into training, development, and test sets using a ratio of 7:1:2. Experiments are conducted for three tasks: responsive relationship extraction (RRE), responsive emotion cause extraction (RECE), and responsive emotion category detection (RED). In our evaluation, we employ

[3]The model can be obtained from https://huggingface.co/hfl/chinese-alpaca-2-13b according to its intended use.

Dataset	Lang	Genre	Domain	Sent. #	Emotion Type #	RR	EC
Dailydialog (Li et al., 2017)	en	Dialogue	Daily Conversation	102K	7	×	×
Meld (Poria et al., 2018)	en	Dialogue	TV Program	13K	7	×	×
IEMOCAP (Busso et al., 2008)	en	Dialogue	Dramatic	7K	10	×	×
CPED (Chen et al., 2022)	ch	Dialogue	TV Program	133K	13	×	×
RECCON (Poria et al., 2021)	en	Dialogue	Daily Conversation	12K	9	×	✓
GoEmotions (Demszky et al., 2020)	en	Comments	Social Media	58K	28	×	×
EMOTyDA (Saha et al., 2020)	en	Dialogue	Daily Conversation	19K	7	×	×
MEmoR (Shen et al., 2020)	en	Dialogue	TV Program	22K	14	×	×
EmpatheticDialogues (Shen et al., 2020)	en	Dialogue	Daily Conversation	100K	32	×	×
ESTC (Zhou et al., 2018)	ch	Dialogue	Daily Conversation	4.5M	6	×	×
PELD (Zhiyuan et al., 2021)	en	Dialogue	TV Program	10K	7	×	×
ECF (Wang et al., 2021)	en	Dialogue	TV Program	13509	6	×	✓
GoodNewsEveryone (Oberländer et al., 2020)	en	News Headlines	News	5000	15	×	✓
Ours	ch	Posts and Comments	Social Media	92K	16	✓	✓

Table 5: The comparison between our dataset and existing datasets for emotion analysis. "Lang" denotes the language of the dataset with "ch" denoting Chinese and "en" denoting English. "Sent. #" means the number of sentences in the datasets. "RR" and "EC" mark whether the dataset has annotations for responsive relationship and emotion cause, respectively.

different metrics for each task. For responsive relationship extraction, we utilize Precision, Recall, and F1 scores. We use ROUGE-1, ROUGE-2, and ROUGE-L to evaluate the quality of the RECE task[4]. For responsive emotion detection, we use accuracy as the evaluation metric.

3.3 Experiment Results

In this section, we first provide the performance of different models on three tasks and then show the experimental results of the ablation study.

3.3.1 Overall Performance

We first conduct the experiments treating the above three tasks independently. For example, in the task of RRE, we construct the prompt as shown in Table 7 of Appendix A. In this prompt, a blog post with its comment list is given, and it asks LLMs to output the comment or the blog post IDs that have responsive relationships with the target comment. Similarly, for RECE, the prompt asks LLMs to extract the text spans standing for the emotion cause of the target comment from the given blog post and comment list. For RED, we transform the classification task into a multi-choice question-answering task. The prompt includes a question regarding the responsive emotion category of the target comment. A list of category options is provided, and the output is expected to be one of the options. In order to reduce positional bias in this experiment, we randomly shuffle the option positions in the candidate list in our prompts.

[4]We use the rouge python package to accomplish this evaluation

The experimental results are presented in Table 6. It showed that both the Chinese-Alpaca-2-13B without fine-tuning and ChatGPT models performed poorly across the three tasks. For instance, in the case of zero-shot, ChatGPT outperformed Chinese-Alpaca-2-13B without fine-tuning but still achieved only 0.19 accuracy in emotion detection. We attempt to fine-tune the Chinese-Alpaca-2-13B model with data and found that fine-tuning significantly improved its performances across different tasks, surpassing the performances of the 8-shot ChatGPT model. The reason behind this is that large language models are trained in an autoregressive manner, which makes them excel in generative tasks. However, their performance in discriminative tasks may be inferior to specifically designed models that are fine-tuned using data. This demonstrates that training with data can significantly enhance the performance of large models as well, thereby demonstrating the value of our dataset.

3.3.2 Ablation Study

We then conduct the ablation study and compare the results with and without responsive relationships. The prompts are provided in Table 8 of Appendix A, where the information on responsive relationships is given to assist the tasks of RED and RECE. The results are shown in Table 6 where "+ RR" denotes the results exploiting responsive relationship information. We can find that for the few shots setting of ChatGPT, with the information of responsive relationships, the performances of both RED and RECE are improved by 29.55% and 6.47%, respectively. This phenomenon can also be observed in the results of Chinese-Alpaca with

Model		RED	RRE			RECE		
		ACC	P	R	F1	ROUGE-1	ROUGE-2	ROUGE-L
Alpaca	CA-2(0-shot)	0.1219	0.0700	0.2324	0.1075	0.2753	0.1202	0.2503
	CA-2 (FT)	0.5580	**0.8253**	**0.6365**	**0.7028**	0.5177	0.4697	0.5129
	CA-2(FT + RR)	**0.5740**	*	*	*	**0.6459**	**0.6043**	**0.6416**
ChatGPT	ChatGPT(0-shot)	0.1900	0.2244	0.0702	0.1069	0.3009	0.2024	0.2779
	ChatGPT(8-shots)	0.2007	0.3694	0.6153	0.4616	0.4148	0.3470	0.4063
	ChatGPT(8-shots + RR)	0.2600	*	*	*	0.4384	0.3755	0.4298

Table 6: Experimental results of different models. "RED", "RRE" and "RECE" stand for responsive emotion detection, responsive relation extraction, and responsive emotion cause extraction. "FT" denotes the model is fine-tuned on RESEMO. "CA-2" denotes Chinese-Alpaca-2-13B. "P" denotes Precision and "R" denotes Recall. "*" denotes that no experimentation is required for this component. "+ RR" denotes the results exploiting responsive relationship information.

fine-tuning, where the performances of both RED and RECE are improved by 2.88% and 26.09%, respectively. These results show that exploiting the information of responsive relationships can be beneficial for both emotion detection and emotion cause extraction tasks.

To sum up, the above experimental results validate our proposed RESEMO dataset. Firstly, in the few shots and fine-tuning experiments, the training data from our dataset can prominently improve the three tasks, which demonstrates the consistency and effectiveness of our dataset. Secondly, the ablation study shows that the unique characteristic of responsive relationships in our dataset is beneficial to both RED and RECE tasks, showing its advantages in responsive emotion analysis.

4 Related Work

As shown in Table 5, There are existing datasets for emotion analysis. The DailyDialog dataset (Li et al., 2017) is collected from various websites that provide services for English language learners to practice English conversations. The dialogue segments in this dataset are presented in English and annotated with 7 emotion categories. However, response relationships and emotion causes are not annotated.

The MELD dataset (Poria et al., 2018) includes 13,000 utterances from 1,433 dialogue segments of the TV series "Friends". In addition to dialogue text, the dataset also annotates various modal information, including speech and facial expressions.

The IEMOCAP dataset (Busso et al., 2008) provides detailed motion capture information about the head, face, hand movements, etc., to showcase emotional expressions and postures during interpersonal interactions.

The CPED dataset (Chen et al., 2022) consists of over 12,000 dialogues from 392 speakers in 40 TV programs. It is a large-scale Chinese personalized and emotion-driven dialogue dataset, aiming to perform both personality and emotion analysis.

Besides, the GoEmotions dataset (Demszky et al., 2020) includes 58,000 English Reddit comments that have been labeled with 27 different emotions or neutral. Mastodon explores the relationship between conversational behavior and emotion recognition, suggesting that this correlation can be leveraged for transfer learning between two tasks (Cerisara et al., 2018). The EMOTyDA dataset investigates the role of multimodality and emotion recognition in dialog behavior classification (Saha et al., 2020). The MEmoR dataset is proposed to perform emotion category detection in situations with missing modalities (Shen et al., 2020). The PELD dataset collects data from TV show dialogues, creating an emotional dialogue dataset with personality traits (Zhiyuan et al., 2021). The EmpatheticDialogues dataset aims to advance the development of conversational agents capable of better understanding and responding to human emotions and concerns (Rashkin et al., 2018). ECF establishes a dataset comprising 9,272 multimodal emotion-cause pairs (Wang et al., 2021).

5 Conclusion

In this paper, we propose a Chinese dataset named RESEMO for responsive emotion analysis in the social media domain. The dataset consists of 3,813 posts with 68,781 comments collected from Weibo, including three types of annotation, namely, responsive relationship, responsive emotion cause and responsive emotion categories.

We test several LLM-based baseline methods on RESEMO for responsive relationship extraction, responsive emotion detection, and responsive emo-

tion cause extraction. The experimental results validate our proposed RESEMO dataset in terms of consistency and effectiveness. Besides, the unique characteristic of responsive relationships also shows its advantages in responsive emotion analysis.

6 Limitations

This study also presents certain constraints and limitations. Firstly, our dataset only comprises the Chinese language, thereby limiting its applicability to other languages. Secondly, our data collection is collected from the Chinese social media platform Weibo, which primarily reflects the language behaviors and interaction patterns specific to users in mainland China. Consequently, there may be risks associated with generalizing the findings to other Chinese language data. Finally, there is multimodal content on current social media platforms, which has the potential to contribute to emotion analysis. To address these aforementioned issues, we intend to expand our dataset in three main ways. Firstly, we will collect data in the English language to create a bilingual dataset. Secondly, we will acquire data from other popular social media platforms to enhance the diversity of our dataset. Finally, in the next version of dataset research, we will cover multimodel content to further enhance our dataset.

References

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577.

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.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. *arXiv preprint arXiv:1807.05013*.

Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint Aspect Extraction and Sentiment Analysis with Directional Graph Convolutional Networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 272–279.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

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*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, Imre Solti, et al. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144. American Medical Informatics Association.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of Fine-grained Emotions. *arXiv preprint arXiv:2005.00547*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020. ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4729–4740.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Bharat Gaind, Varun Syal, and Sneha Padgalwar. 2019. Emotion detection and analysis on social media. *arXiv preprint arXiv:1901.08458*.

Ruyi Gan, Ziwei Wu, Renliang Sun, Junyu Lu, Xiaojun Wu, Dixiang Zhang, Kunhao Pan, Ping Yang, Qi Yang, Jiaxing Zhang, and Yan Song. 2023. Ziya2: Data-centric Learning is All LLMs Need. *arXiv preprint arXiv:2311.03301*.

Jan Hruska and Petra Maresova. 2020. Use of social media platforms among adults in the united states—behavior on social media. *Societies*, 10(1):27.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A Manually Labelled Multi-turn Dialogue Dataset. *arXiv preprint arXiv:1710.03957*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Yuanhe Tian, Tsung-Hui Chang, Song Wu, Xiang Wan, and Yan Song. 2021. Exploring word segmentation and medical concept recognition for chinese medical texts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 213–220.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named Entity Recognition for Social Media Texts with Semantic Augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1383–1391, Online.

Laura Ana Maria Oberländer, Evgeny Kim, and Roman Klinger. 2020. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566.

OpenAI. 2022. Introducing chatgpt. OpenAI Blog.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*.

Robert Plutchik and Henry Kellerman. 1980. *Emotion, theory, research, and experience: theory, research and experience*. Academic press.

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*.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing Emotion Cause in Conversations. *Cognitive Computation*, 13:1317–1332.

Han Qin, Yuanhe Tian, Fei Xia, and Yan Song. 2022. Complementary Learning of Aspect Terms for Aspect-based Sentiment Analysis. In *Proceedings of the 13th Language Resources and Evaluation Conference*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multimodal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 493–502.

Yan Song, Tong Zhang, Yonggang Wang, and Kai-Fu Lee. 2021. ZEN 2.0: Continue Training and Adaption for N-gram Enhanced Text Encoders. *arXiv preprint arXiv:2105.01279*.

Liyaning Tang, Yiming Zhang, Fei Dai, Yoojung Yoon, and Yangqiu Song. 2018. What construction topics do they discuss in social media? a case study of weibo in china. In *Construction Research Congress 2018*, pages 612–621.

Yuanhe Tian, Weidong Chen, Bo Hu, Yan Song, and Fei Xia. 2023a. End-to-end Aspect-based Sentiment Analysis with Combinatory Categorial Grammar. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13597–13609, Toronto, Canada.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2023b. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. *arXiv preprint arXiv:2311.06025*.

Yuanhe Tian, Chang Liu, Yan Song, Fei Xia, and Yongdong Zhang. 2024. Aspect-based Sentiment Analysis with Context Denoising. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Mexico City, Mexico.

Yuanhe Tian, Renze Lou, Xiangyu Pang, Lianxi Wang, Shengyi Jiang, and Yan Song. 2022a. Improving English-Arabic transliteration with phonemic memories. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3262–3272, Abu Dhabi, United Arab Emirates.

Yuanhe Tian, Han Qin, Fei Xia, and Yan Song. 2022b. Chimst: A chinese medical corpus for word segmentation and medical term recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5654–5664.

Dingmin Wang, Meng Fang, Yan Song, and Juntao Li. 2019. Bridging the gap: Improve part-of-speech tagging for chinese social media texts with foreign words. In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 12–20.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2021. Multimodal emotion-cause pair extraction in conversations. *arXiv preprint arXiv:2110.08020*.

Wen Zhiyuan, Cao Jiannong, Yang Ruosong, Liu Shuaiqi, and Shen Jiaxing. 2021. Automatically select emotion for response via personality-affected emotion transition. *arXiv preprint arXiv:2106.15846*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

A Illustration of Instruction Prompts

We conduct few-shot and instructional fine-tuning experiments using RESEMO, and the instruction prompts are illustrated in Table 7 and Table 8.

B Detailed Annotation Guidelines

In this appendix, we provide the detailed annotation guidelines. Each post with its corresponding comments forms a unit for annotation in our dataset. For each post, the comments are organized in two levels as illustrated in Figure 1, and comments under the same parent node are assigned with indices in chronological order from oldest to newest. When assigned to label a post, an annotator is required to sequentially annotate each comment following such chronological order, and the responsive relationship, emotion cause, and emotion category for each comment should be labeled. To understand the context for annotation, the annotator can only refer to the post blog and the preceding comments, which we refer to as preceding utterances, before the target comment. Detailed instructions on how to annotate the three types of annotations are then provided.

- Responsive relationship: To determine the responsive relationship for the target comment, the annotator begins by identifying the preceding utterances to which the target comment responds. The annotator then annotates those specific utterances as the responsive relationships of the target comment. Since the responsive relationship can be one to many, he can annotate multiple utterances. For example, in Figure 1, the target comment is *id:1-3* "Some people are so excellent that they only read physical books ." It responds to the comments *id:1-1* "I was just about to buy a book and see this post. I feel that excellent individuals only read physical books" and *id:1-2* "Are there people who never read e-books?" Therefore, the IDs of these two comments, namely *id:1-1* and *id:1-2*, are marked as the responsive relationship for the target comment.

- Emotion cause: The emotion causes are found in the utterances to which the target comment responds. However, since the target comment may have multiple responsive relationships, we simplify the annotation process by asking annotators to identify the utterance that is most likely to be the cause of the emotion expressed in the target comment. Once this utterance is selected, annotators are instructed to choose the text span that conveys the emotion cause using the fewest words possible. In Figure 1, for example, the target comment responds to two utterances. However, the emotion is most likely originating from the text span of comment *id:1-1* "I feel that excellent individuals only read physical books." Therefore, this text span is annotated as the emotion cause.

- Emotion category: Annotators are tasked with selecting the most appropriate emotion category from a set of 16 options. This selection should be made based on their understanding of the context provided by the preceding utterances. As shown in Figure 1, the target comment appears to be praising someone's excellence, but when considered in the context, it is obvious that it is actually sarcastic towards those who do not read e-books. Therefore, the emotion category here is annotated as *Cynicism*.

Task	Chinese	English Translation
RED	Instruction: 请你完成情绪分类任务。接下来将给你一条博文及其相关评论，请根据博文和相关评论的内容和语境给出目标评论的情绪标签，情绪标签从[选项]中选择。请只返回标签，不要包含其它解释性语言。 [博文和评论] [目标博文或目标评论] A. [类别1] B. [类别2] ... P. [类别16] [示例博文和评论1] [目标博文或目标评论] [选项] ... [示例博文和评论n] [目标博文或目标评论] [选项] Output: [选项]	Instruction: Please complete the emotion classification task. Next, you will be provided with a blog post and its related comments. Based on the content and context of the blog post and related comments, assign an emotion label to the target reply. Choose from the [options]. Please only provide the label, without including other explanatory language. [Post and comments] [Target Post or Target comment] A. [Category 1] B. [Category 2] ... B. [Category 16] [Example Blog Post and Comment 1] [Target Post or Target comment] [Option] ... [Example Blog Post and Comment n] [Target Post or Target comment] [Option] Output: [Option]
RRE	Instruction: 请你完成响应关系提取任务。响应关系是指目标评论是对之前评论或博文的响应或回应，表示对前文内容的回答、评论、赞同、补充、反驳等。我们称此处之前的某条评论或博文与目标评论有响应关系。接下来将给你一条博文及其相关评论，请根据博文和相关评论的内容和语境给出与目标评论存在响应关系的博文或评论的【id】，其中【id:0】为博文id，其余为评论id。 [博文和评论] [目标评论] [示例博文和评论1] [目标评论] [IDs] ... [示例博文和评论n] [目标评论] [IDs] Output: [IDs]	Instruction: Please complete the responsive relationship extraction task. A responsive relation refers to the target reply being a response to a previous comment or blog post, indicating an answer, reply, agreement, addition, refutation, etc., to the preceding content. We refer to a specific preceding reply or blog post here that has a responsive relation with the target reply. Next, you will be given a blog post and its related comments, please give the [id] of the blog post or comments that have a responsive relationship with the target comment based on the content and context of the blog post and the related comments, where [id:0] is the id of the blog post and the rest are the ids of the comments. [Post and comments] [Target comment] [Example Blog Post and Comment 1] [Target comment] [IDs] ... [Example Blog Post and Comment n] [Target comment] [IDs] Output:[IDs]
RECE	Instruction: 请你完成情绪原因提取任务。接下来将给你一条博文及其相关评论，请根据博文和相关评论的上下文信息和语境，给出之前的评论或博文中哪些内容片段是造成目标评论当前情绪的原因。请只返回之前博文和评论中的内容片段，不要包含其它解释性语言。 [博文和评论] [目标评论] [示例博文和评论1] [目标评论] [情绪原因] ... [示例博文和评论n] [目标评论] [情绪原因] Output: [情绪原因]	Instruction: Please complete the emotion cause extraction task. Next, you will be provided with a blog post and its related comments. Based on the contextual information and the context of the blog post and related comments, identify which content segments in the preceding comments or blog posts are the reasons for the current emotion expressed in the target reply. Please only provide the content segments from the previous blog post and comments, without including other explanatory language. [Post and comments] [Target comment] [Example Blog Post and Comment 1] [Target comment] [Emotion Cause] ... [Example Blog Post and Comment n] [Target comment] [Emotion Cause] Output: [Emotion Cause]

Table 7: The prompt template is used for responsive relationship extraction (RRE), responsive emotion detection (RED), and responsive emotion cause extraction (RECE). '[]" marks the template areas to be realized by text data. In our experiments, we use the Chinese prompt where the English translation is given for reference.

Task	Chinese	English Translation
RED + RR	Instruction: 请你完成情绪分类任务。接下来将给你一条博文及其相关评论，请根据博文和相关评论的上下文信息和语境给出目标评论的情绪标签，为了帮助你更好地完成任务，我们还提供了与目标评论具有响应关系的博文或评论的id，其中响应关系是指目标评论是对之前评论或博文的响应或回应，表示对前文内容的回答、评论、赞同、补充、反驳等，【id:0】为博文id，其余为评论id。情绪标签从[选项]中选择。请只返回标签，不要包含其它解释性语言。 [博文和评论] [与目标评论存在响应关系的id] [目标评论] A. [类别1] B. [类别2] ... P. [类别16] [示例博文和评论1] [与目标评论存在响应关系的id] [目标评论] [选项] ... [示例博文和评论n] [与目标评论存在响应关系的id] [目标评论] [选项] Output: [选项]	Instruction: Please complete the emotion classification task. Next, we will provide you with a blog post and related comments. Based on the contextual information and context of the blog post and comments, provide the emotion label for the target comment. To assist you in completing the task more effectively, we also provide the IDs of blog posts or comments that are in response to the target comment. Response refers to the target comment being a response or reply to a previous comment or blog post, indicating an answer, comment, agreement, addition, refutation, etc., to the preceding content. [id:0] is the blog post ID, and the rest are comment IDs. Choose the emotion label from the [options]. Please only return the label without including any explanatory language. [Post and comments] [The id of comment that has a responsive relationship with the target comment] [Target Post or Target comment] A. [Category 1] B. [Category 2] ... B. [Category 16] [Example Blog Post and Comment 1] [The id of comment that has a responsive relationship with the target comment] [Target comment] [Option] ... [Example Blog Post and Comment n] [The id of comment that has a responsive relationship with the target comment] [Target comment] [Option] Output: [Option]
RECE + RR	Instruction: 请你完成情绪原因提取任务。接下来将给你一条博文及其相关评论，请根据博文和相关评论的上下文信息和语境，给出之前的评论或博文中哪些内容片段是造成目标评论当前情绪的原因。为了帮助你更好地完成任务，我们还提供了与目标博文具有响应关系的博文或评论的id，其中响应关系是指目标评论是对之前评论或博文的响应或回应，表示对前文内容的回答、评论、赞同、补充、反驳等，【id:0】为博文id，其余为评论id。请只返回之前博文和评论中的内容片段，不要包含其它解释性语言。 [博文和评论] [与目标评论存在响应关系的id] [目标评论] [情绪原因] [示例博文和评论1] [与目标评论存在响应关系的id] [目标评论] [情绪原因] ... [示例博文和评论n] [与目标评论存在响应关系的id] [目标评论] [情绪原因] Output: [情绪原因]	Instruction: Please complete the emotion cause extraction task. Next, we will provide you with a blog post and related comments. Based on the contextual information and context of the blog post and comments, identify which content segments from previous comments or blog posts caused the current emotion in the target comment. To assist you in completing the task more effectively, we also provide the IDs of blog posts or comments that are in response to the target comment. Response refers to the target comment being a response or reply to a previous comment or blog post, indicating an answer, comment, agreement, addition, refutation, etc., to the preceding content. [id:0] is the blog post ID, and the rest are comment IDs. Please only return the content segments from previous blog posts and comments without including any explanatory language. [Post and comments] [The id of comment that has a response relationship with the target comment] [Target comment] [Emotion Cause] [Example Blog Post and Comment 1] [The id of comment that has a responsive relationship with the target comment] [Target comment] [Emotion Cause] ... [Example Blog Post and Comment n] [The id of comment that has a responsive relationship with the target comment] [Target comment] [Emotion Cause] Output: [Emotion Cause]

Table 8: The prompt template exploits responsive relationships for RED and RECE tasks. "RED + RR" means responsive emotion detection with responsive relationship information, and "RECE + RR" means responsive emotion cause extraction with responsive relationship information. '[]" marks the template areas to be realized by text data. In these experiments, we use the Chinese prompt where the English translation is given for reference.