

Empathetic Response Generation for Distress Support

Anonymous ACL submission

Abstract

AI-driven chatbots are seen as an attractive solution to support people undergoing emotional distress. One of the main components of such a chatbot is the ability to empathize with the user. But a significant limitation in achieving this goal is the lack of a large dialogue dataset containing empathetic support for those undergoing distress. In this work, we curate a large-scale dialogue dataset that contains $\approx 1.3M$ peer support dialogues spanning across more than 4K distress-related topics. We analyze the empathetic characteristics of this dataset using statistical and visual means. To demonstrate the utility of this dataset, we train four baseline neural dialogue models that can respond empathetically to distress prompts. Two of the baselines adapt existing architecture and the other two incorporate a framework identifying levels of cognitive and emotional empathy in responses. Automatic and human evaluation of these models validate the utility of the dataset in generating empathetic responses for distress support and show that identifying levels of empathy in peer-support responses facilitates generating responses that are lengthier, richer in empathy, and closer to the ground truth.

1 Introduction

The World Health Organization estimates that psychological distress affects 29% of people in their lifetime (Steel et al., 2014). Despite the availability of mental health services, people hesitate to reach them because of the public stigma associated with mental health. There is also a severe shortage of mental health workers (Vaidyam et al., 2019). Thus, recent work investigates how technology can be utilized to meet the needs of people suffering from distress. One such solution is the development of conversational agents or chatbots to deliver distress support (Fitzpatrick et al., 2017; Inkster et al., 2018; Mousavi et al., 2021).

Deep neural networks work very effectively in the development of task-oriented and open-

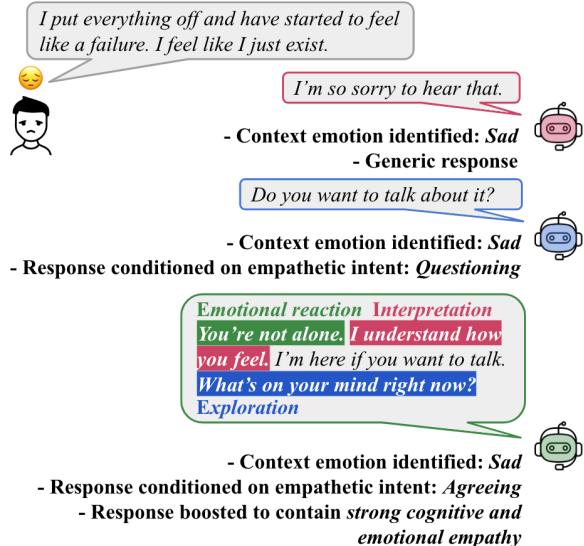


Figure 1: Distress support responses generated by our different chatbot models trained on peer support dialogues. The last response boosted with cognitive and emotional empathy communication mechanisms stands out from the rest as its lengthier and richer in empathy.

domain conversational agents (Sutskever et al., 2014; Vinyals and Le, 2015; Wen et al., 2015). Most of such dialogue agents can generate syntactically correct and contextually relevant responses. But a major challenge faced by these systems is identifying human emotion and responding in an empathetic manner (Rashkin et al., 2018; Welivita et al., 2021). This is very important when developing chatbots to support distress as one of the major components that contributes to the success of such interaction is the ability to empathize (Bohart et al., 2002; Thwaites and Bennett-Levy, 2007). Recently, researchers have curated emotion-labeled and empathetic datasets such as EmotionLines (Hsu et al., 2018), EmoContext (Chatterjee et al., 2019), EmpatheticDialogues (Rashkin et al., 2018), and ESConv (Liu et al., 2021) to enable training dialogue systems that can generate emotion-aware and empathetic responses. However, the above datasets include only a limited amount of dialogues deal-

ing with distress. The dialogues in the first three datasets are more open-domain and span across topics less related to distress. The ESConv dataset that is more focussed on distress contains only 1.3K dialogues covering only 13 distress-related topics. Recent research has curated and conducted analysis on real counseling conversations (Althoff et al., 2016; Zhang and Danescu-Niculescu-Mizil, 2020). But these datasets are not publicly accessible due to privacy and ethical reasons.

To address the above limitations, we curated a large-scale dialogue dataset, named RED (Reddit Emotional Distress), containing \approx 1.3M dialogues spanning across more than 4K distress-related topics. The dialogues are scraped from the popular peer support forum, Reddit. Peers are seen to actively engage in such forums to support others undergoing distress and thus they contain distress-related dialogues in abundance spanning a wide range of topics. Prior work has also found that responses from peers contain higher empathetic concern for posts seeking help as many peers share similar distressful experiences (Hodges et al., 2010). But as these conversations are available as long threads, the turn-taking structure has to be explicitly extracted and the conversations have to undergo a rigorous pre-processing pipeline including the removal of profanity before they are used to train chatbots. Even then, the dataset can still possess less ideal responses to distress since peers are not trained in delivering distress support as professionals. We take steps to address this by making use of existing empathetic frameworks based on psychology that can be used to identify highly empathetic responses in such dialogues and enabling chatbot models to favor such responses over others.

Empathy is a complex multi-dimensional construct with two broad aspects related to emotion and cognition. The emotion aspect refers to the ability to share the feelings of another person and the cognition aspect refers to the ability to understand and acknowledge how a person feels. In mental health therapy, both emotional and cognitive empathy are equally important (Selman, 1981). Thus, for the development of distress support chatbots, it is vital to understand these types of empathy and the techniques by which these different types of empathy can be elicited. We apply such empathy recognition frameworks on RED to develop several distress support chatbots models. Figure 1 shows an example. In the first instance, identification of

the context emotion enables the chatbot to produce a suitable generic response. In the second instance, the chatbot’s response is conditioned on a specific empathetic response intent that helps to generate a diversified response. In the third instance, training the model to favour more cognitive and emotional empathy helps in generating lengthier responses containing specific cognitive and emotional empathy communication strategies.

Our contributions are three folds. 1) We curate a large-scale dialogue dataset containing \approx 1.3M distress support dialogues spanning across more than 4K distress topics, from a set of carefully selected subreddits. 2) We describe the empathetic dialogue characteristics between the speakers and the listeners in this dataset using statistical and visual means. 3) Using this dataset as a benchmark, we develop four baseline chatbot models. The first two baseline models adapt existing empathetic response generation architectures. On top of them, we develop two new baselines by incorporating a framework that can identify levels of emotional and cognitive empathy in responses contained in RED. Automatic and human evaluation of the models’ responses validate the utility of the RED dataset in facilitating empathetic response generation and show that identifying different levels of emotional and cognitive empathy enables generating responses that are lengthier, richer in empathy, and closer to the ground-truth.

2 Related Work

Many dialogue datasets such as IEMOCAP (Busso et al., 2008), SEMAINE (McKeown et al., 2011), and MELD (Poria et al., 2019) are developed to make chatbots understand users’ emotions and respond appropriately. These datasets contain visual, acoustic, and textual signals. More recent work such as EmotionLines (Hsu et al., 2018), Open-Subtitles (Lison et al., 2019), and EDOS (Welivita et al., 2021) are conversation datasets containing TV and movie transcripts translated from voice to text. Though these works intend to build dialogue datasets by improving the sentence quality, they are still unable to fully model interactions occurring only via text. And most of the dialogues contained in these datasets represent generic day-to-day situations and not psychological distress in particular.

Rashkin et al. (2018) developed the EmpatheticDialogues dataset, inclusive of 25K dialogues grounded on 32 positive and negative emotions.

Liu et al. (2021) developed the ESConv dataset, containing \approx 1.3K dialogues discussing emotional distress and whose responses are grounded on the Helping Skills Theory (Hill, 2009). But the crowd-sourced artificial setting used to curate them makes the dialogue prompts less authentic and the responses less genuine. Because of the cost of crowdsourcing, it also limits the size of these datasets as well as their topic coverage. Thus, a large-scale topically diverse dataset focused on textual conversations between speakers who are emotionally distressed and listeners who actively offer emotional support is lacking in the literature. This type of conversation could be available as recorded therapy sessions between psychologically distressed patients and therapists. However, such counseling datasets used to conduct recent research (Althoff et al., 2016; Zhang and Danescu-Niculescu-Mizil, 2020) are not directly accessible to the public due to ethical reasons. To address these limitations, we curate a large dataset containing peer support dialogues related to a variety of distress-related topics and validate that combined with existing empathy-identifying frameworks, it can potentially be used to develop chatbots that can offer empathetic support to distressful user prompts.

3 Reddit Emotional Distress Dataset

3.1 Data Curation and Preprocessing

Online peer support forums encourage open discussion of often stigmatized psychological concerns and personal distress (De Choudhury and De, 2014; Sharma et al., 2017). They provide alternative means for connection and support when other means of care are less accessible. The anonymity in such platforms facilitates self-disclosure and such discussions help people to feel more supported and less stressed in times of crisis (De Choudhury and De, 2014; Smith-Merry et al., 2019). Reddit is one such platform, which ranks among the most visited websites in the world (Sharma et al., 2017). Reddit users can create community forums called “subreddits” to discuss and support each other on a breadth of topics. Reddit policies also allow researchers to scrape its data and use them for research. Since many people interact in Reddit in a day-to-day basis, the distress-related topics it covers are abundant and have a wide variety. Because of these reasons we chose Reddit to curate conversations that provide support for people in distress.

For this purpose, we choose 8 subreddits: *de-*

pression; depressed; Off My Chest; SuicideWatch; Depression Help; sad; Anxiety Help; and Mental Health Support, where such conversations were abundantly present. We used the Pushshift API (Baumgartner et al., 2020) to scrape English textual conversations from the above subreddits. We extracted one dyadic dialogue per conversation thread, thereby diversifying the conversation topics in the dataset. To preserve anonymity, we replaced the usernames with *speaker, listener_1, ... listener_n*. Next, we removed HTML tags and URLs from the data, and replaced numerals with a special tag <NUM>. But punctuation marks, emoticons, and emojis were preserved as they can be useful indicators to identify users’ emotions. We applied *profanity-check* (Zhou et al., 2020), a fast and robust library to detect profane language in listeners’ turns and removed such sentences containing profanity. But we retained profane speaker turns as they can inform the speakers’ state of mind.

3.2 Descriptive Statistics

The resultant RED dataset contains \approx 1.3 million dyadic conversations. Table 1 displays the summary of descriptive statistics of conversations present in the dataset as well as in individual subreddits. We used Agglomerative clustering (Murtagh and Legendre, 2014) to cluster distress prompts and recognize clearly identifiable topic clusters. At an optimal clustering threshold of 0.85, the prompts were separated into 4,363 topic clusters. By applying TF-IDF based topic modeling on these clusters, we uncovered some clearly distinguishable distress-related topics. Some of the most common topics identified were *Suicidal ideation, Anxiety attacks, Weight gain, Loneliness, Failing college, and Covid19*. The topics and their associated keywords are included in the appendices.

3.3 Emotion and Intent Analysis

To analyse the emotions and intents present in the RED dataset, we used a BERT transformer-based classifier proposed by Welivita and Pu (2020) and classified the utterances in RED into one of 32 fine-grained emotions and 8 empathetic response intents. This classifier was trained on the EmpatheticDialogues dataset and has a classification accuracy of 65.88% on the EmpatheticDialogues test set, which is comparable with the state-of-the-art emotion classifiers. Manual validation of the labels proposed by the classifier on a random subset of 100 utterances from the RED datasets yielded an ac-

Subreddit	# Dialogues	# Turns	# Tokens	Avg. # turns per dialog	Avg. # tokens per dialogue	Avg. # tokens per turn
r/depression	510,035	1,396,044	106,967,833	2.74	209.73	76.62
r/depressed	10,892	23,804	1,940,000	2.19	178.11	81.50
r/offmychest	437,737	1,064,467	109,459,738	2.43	250.06	102.83
r/sad	18,827	42,293	3,088,562	2.25	164.05	73.03
r/SuicideWatch	262,469	791,737	59,267,000	3.02	225.81	74.86
r/depression_help	23,678	51,849	5,412,390	2.19	228.58	104.39
r/Anxietyhelp	8,297	18,351	1,428,287	2.21	172.14	77.83
r/MentalHealth Support	3,551	7,931	772,952	2.23	217.67	97.46
All	1,275,486	3,396,476	88,336,762	2.66	226.06	84.89

Table 1: Descriptive statistics of the conversations in the RED dataset.

curacy of 64%, which allows us to have reasonable judgments about the RED dataset using the predicted labels. In Figure 2, we visualize the emotion and intent distributions in speaker and listener turns in the RED dataset. It could be seen that the speakers’ emotions are mostly centered around negative emotions. The most frequent speakers’ emotions that can be observed are *ashamed* (9.98%), *lonely* (8.41%), *sad* (7.52%), and *apprehensive* (5.32%).

A significant proportion of the listener turns contain empathetic response intents. The listeners’ intents are mostly centered around *questioning* (10.26%), *agreeing* (7.98%), *suggesting* (5.49%), and *sympathizing* (4.56%). Though empathetic response intents take prominence in the listener turns, they also contain emotional statements that mostly reflect the *sad* emotion (4.98%). This can possibly be explained by the study of affective asymmetry by Vaish et al. (2008) that states negative emotional experiences have more power in triggering negative emotions in the listener as humans are more sensitive to negative emotions.

Figure 3 shows the conversational dynamics in terms of emotion-intent flow patterns that could be observed in the first four dialogue turns. The first and the third turns represent the speaker turns, while the second and the fourth turns represent the listener turns. According to statistics, 93.71% dialogues in the dataset start with a negative emotion. Then in the next turn, the listeners tend to show empathy by means of intents such as *questioning* (35%), *agreeing* (12.43%), *suggesting* (8.11%), and *sympathizing* (7.23%). As the dialogues proceed, we can observe a 278.59% increase of positive emotions in the third turn compared to the first. The speakers mostly express emotions such as *grateful* (7.50%), *trusting* (7.26%), and *hopeful* (6.56%) as a result of the support offered by the listeners. Such conversational dynamics further

validate the use of RED in applications concerning empathetic chatbots that can lift up the emotions of people suffering from distress.

4 Conversational Baselines

Using the RED dataset as a benchmark, we trained four baseline dialogue models. The first two baselines adapted the architecture of EmoPrepend (Rashkin et al., 2018) and MEED (Xie and Pu, 2021), which are state-of-the-art empathetic chatbot models. We also examined different ways existing models can be combined to produce more empathetic responses for distress prompts. For this purpose, we developed another two experimental baselines, EPIMEED and EPIMEED+, by combining MEED with EPITOME (Sharma et al., 2020), which is a theoretically-grounded framework that can identify levels of cognitive and emotional empathy in text-based conversations and extract rationales underlying its predictions. All the models were trained on 80% of RED conversations, leaving 10% of the conversations each for validation and testing. Figure 4 show the architecture of the different models we used for evaluation.

EmoPrepend: This model proposed by Rashkin et al. (2018) is a transformer based encoder-decoder model. During training and inference, the top-k predicted emotion labels from a supervised classifier for the corresponding dialogue context is prepended to the beginning of the token sequence as encoder input. We initialized the encoder of this model with weights from the pre-trained language model RoBERTa (Liu et al., 2019) and trained it on RED, prepending the top-1 emotion or intent predicted by the BERT transformer-based classifier proposed by Welivita and Pu (2020).

MEED: This model proposed by Xie and Pu (2021) consists of two modules: 1) a response emotion/intent prediction module; and 2) a response

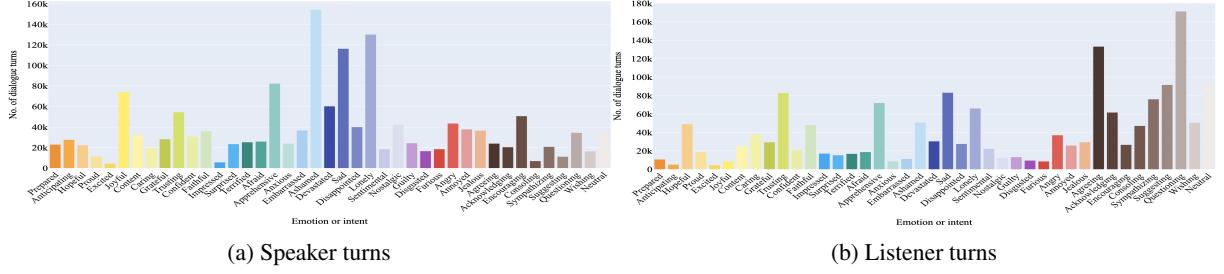


Figure 2: Emotion and intent distributions in speaker and listener turns in the RED dataset. The last 9 bars depict empathetic intents and the rest depict emotional statements.

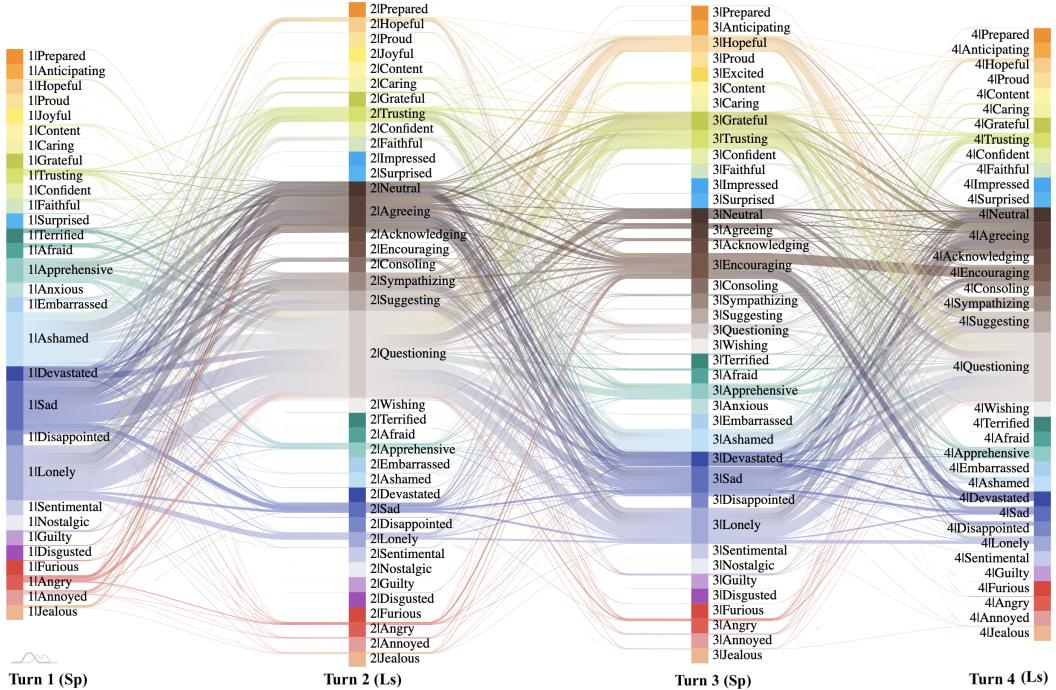


Figure 3: Frequent emotion-intent flow patterns in the RED dataset. For simplicity, only the first four dialogue turns are visualized.

generation module. The response generation module is an encoder-decoder model that uses the transformer architecture, in which the encoder is initialized with weights from RoBERTa. The response emotion/intent prediction module takes the dialogue context as input and predicts what the emotion or intent of the response should be. This prediction is used to condition the response generated by the decoder in the first module.

EPIMEED: In therapy, interacting empathetically with clients is fundamental to success (Bohart et al., 2002; Elliott et al., 2018). Even though empathy can be interpreted as reacting with emotions of warmth and compassion (Buechel et al., 2018), a separate but key aspect of empathy is also to communicate a cognitive understanding of others, referred to as cognitive empathy. For mental health support, both emotional and cogni-

tive empathy are equally important (Selman, 1981). Thus, it is important to identify such emotional and cognitive empathetic responses amongst other responses that appear in RED and train models in such a way that they favor such responses that reflect cognitive and emotional empathy over others. To support this, we experimented with a new model EPIMEED, by combining MEED with an existing text-based cognitive and emotional empathy identifying framework named EPITOME (Sharma et al., 2020). EPITOME recognizes three empathetic communication mechanisms 1) Emotional reactions (emotional empathy); 2) Interpretations (cognitive empathy); and 3) Explorations (cognitive empathy). For each of these mechanisms, it predicts a numerical value, 0, 1, or 2 — 0: peers not expressing them at all (no communication); 1: peers expressing them to some weak degree (weak

341
342
343
344
345
346
347
348
349

350
351
352
353
354
355
356
357
358

359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376

Speaker (U_1): I think about suicide often. But I don't want to kill myself. Is this something I should be concerned about?
Listener (U_2): That's a hard question. Plan to live. Plan to survive. Plan to be strong. Best wishes, friend.
Speaker (U_3): Thank you, that was helpful to me.
Listener (ground-truth): That's kind of you to say. I'm glad to hear it. You don't know it, but the world needs you.

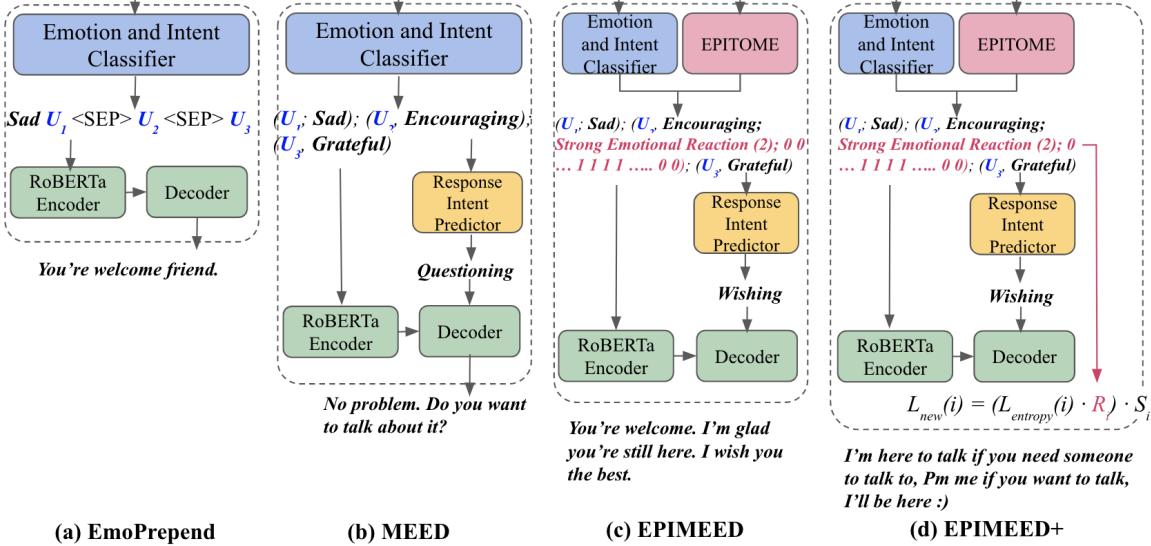


Figure 4: The four models EmoPrepend, MEED, EPIMEED, and EPIMEED+ used for evaluation.

Empathy type	Communication mechanism	Examples
Emotional	Emotional reactions	- Everything'll be fine. (weak) - I really hope things would improve. (strong)
Cognitive	Interpretations	- I realize how you feel. (weak) - If that happened to me, I would feel really isolated. (strong)
Cognitive	Explorations	- What happened? (weak) - I wonder if this makes you feel isolated. (strong)

Table 2: Examples of emotional and cognitive empathy communication mechanisms identified by EPITOME.

communication); 2: peers expressing them strongly (strong communication). Table 2 shows some examples of these communication levels identified in peer support communications.

We use this framework to assign a numerical value to each token contained in the listener responses of the RED dataset. This numerical value is the total of the values predicted by the EPITOME framework for emotional reactions, interpretations, and explorations. This is termed the rationale mask. Next, we feed this information as an additional embedding (in addition to the token embeddings, segment embeddings, position embeddings and emotion embeddings) to the encoder of the response emotion/intent prediction module and response generation module in MEED. We call this additional embedding the *communication embedding*. The

accuracy, precision, and recall of the response emotion/intent predictor of MEED were increased by 22.88%, 62.65%, and 22.89%, respectively after incorporating the communication embedding.

EPIMEED+: To enable the model to favour responses containing stronger emotional reactions, interpretations, and explorations while decoding, we further tweaked the loss function associated with MEED such that it incorporates levels of emotional and cognitive empathy predicted by EPITOME. We modified the loss function to be the dot product between the cross entropy loss and the rationale mask predicted by EPITOME. Compared to the original loss $L_{old}(i)$, the new loss $L_{new}(i)$ given an input sequence i can be written as:

$$L_{old}(i) = L_{entropy}(i) \cdot S_i$$

$$L_{new}(i) = (L_{entropy}(i) \cdot R_i) \cdot S_i$$

where $L_{entropy}(i)$, R_i , and S_i represent the cross entropy between the predicted and the ground-truth responses, the rationale mask, and the segment mask of the input i , respectively. By doing so, it facilitates the model to have a higher tendency to generate tokens annotated with stronger levels of emotional and cognitive empathy.

5 Automatic Evaluation

Automatic evaluation of the models was conducted using a variety of automatic metrics used in evalua-

Model	Diversity metrics		Word-overlap metrics				Embedding-based metrics		Avg. length (# tokens)
	D1	D2	B1	B2	ROUGE-L	METEOR	Skip Thought	Embedding Average	
EmoPrepend	0.0317	0.1178	0.0513	0.0157	0.0662	0.0434	0.4842	0.7346	16.55
MEED	0.0618	0.2889	0.0283	0.0123	0.0690	0.0331	0.4874	0.7408	9.68
EPIMEED	0.0487	0.1912	0.0271	0.011	0.0746	0.0365	0.4911	0.7285	10.30
EPIMEED+	0.0039	0.0181	0.0543	0.0191	0.0559	0.0637	0.4268	0.7650	40.82

Table 3: Automatic evaluation metrics computed on the RED test dataset. D1 and D2 stands for Distinct-1 and Distinct-2 metrics and B1 and B2 stands for BLEU-1 and BLEU-2 metrics.

Model	Emotional Reactions	Interpre-tations	Explora-tions	Total
EmoPrepend	1.148	0.216	0.364	1.728
MEED	0.464	0.412	0.520	1.396
EPIMEED	0.480	0.168	0.732	1.380
EPIMEED+	1.036	0.728	0.948	2.712

Table 4: Average levels of empathy computed for the responses generated by each model, as computed by the EPITOME framework. The maximum value for each of *emotional reactions*, *interpretations*, and *explorations* is 2, making the maximum value for *total* 6.

ing chatbots. They are grouped into diversity-based, word-overlap-based, and embedding-based metrics (details in appendices). Table 3 shows results on the RED test dataset. Accordingly, MEED ranks the top in terms of distinct-unigram and distinct-bigram scores that measures the diversity of the responses. EPIMEED+ ranks the top in majority of word-overlap based metrics and also in embedding average cosine similarity, indicating that responses generated by EPIMEED+ are most likely to contain words from the ground-truth. We also computed the average no. of tokens contained in the responses and EPIMEED+ ranked at the top generating lengthier responses closer to the average length of the ground-truth.

The levels of emotional reactions, interpretations, and explorations computed by EPITOME in the responses generated by the four models are denoted in Table 4. Accordingly, EPIMEED+ generates responses that contain stronger levels of cognitive empathy (as means of interpretations and explorations) than the rest.

6 Human Evaluation

A human evaluation experiment was designed to evaluate the empathetic appropriateness of the responses generated by the four models, by recruiting workers from Amazon Mechanical Turk. We randomly selected 200 dialogue prompts from the RED test dataset and the responses generated by

Model	Good	Okay	Bad	2/3 agree
EmoPrepend	51.82%	21.95%	26.22%	82.00%
MEED	83.42%	16.58%	0.00%	99.50%
EPIMEED	74.07%	25.93%	0.00%	94.50%
EPIMEED+	42.14%	25.16%	30.70%	79.50%

Table 5: Human evaluation results.

the four models for these prompts to be evaluated by the crowdworkers. The workers were asked to drag and drop the responses generated by the models into areas *Good*, *Okay*, and *Bad*, depending on how empathetically appropriate those responses were to the given prompt. This new way of rating makes it easy to compare many models at once instead of traditional A/B testing, which only allows the comparison of a pair of models at a time. Three workers rated the same response and the final results were computed based on the majority vote.

The human evaluation scores for each of the models is denoted in Table 5. Accordingly, it could be observed that $\approx 83\%$ of the responses generated by MEED trained on the RED dataset and $\approx 74\%$ of the responses generated by EPIMEED are rated *Good* with above 90% majority agreement between the workers. None of the responses generated by the above models were rated *Bad*. This further validates the potential of the RED dataset in generating more emotionally supportive and empathetic responses for distress-related prompts. Compared to the above models, there is more disagreement between the workers on the ratings of the responses produced by EPIMEED+ (79.5% compared to 99.50% majority agreement in MEED), leading to EPIMEED+ scoring low on the percentage of *Good* ratings. We elaborate on this under the case study.

6.1 Case Study

Table 6 shows some example responses produced by the four models for two different distress prompts. It could be observed that in both cases,

421
422
423
424
425
426
427
428
429
430
431
432
433
434
435

450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482

Prompt 1:	<i>My cat died today, and I can't get it out of my mind. I keep think of him running on a sunny hill but that only makes me cry more.</i>
EmoPrepend	<i>I'm so sorry you're going through this.</i>
MEED	<i>I'm sorry for your loss. (Sympathizing)</i>
EPIMEED	<i>I'm so sorry for your loss. (Sympathizing)</i>
EPIMEED+	<i>I'm so sorry for what you're going through. I hope you'll be able to reach out to someone who's going through something similar. Have you looked into getting help? (Sympathizing)</i>
Prompt 2:	<i>School is almost over. Things still haven't fixed with my best friend. I want to talk to her about it but I fear it'll just make my situation worse than it already is.</i>
EmoPrepend	<i>I'm so sorry you're going through this. I hope things get better for you soon.</i>
MEED	<i>How long have you been together? (Questioning)</i>
EPIMEED	<i>Do you want to talk about it? (Questioning)</i>
EPIMEED+	<i>Wanna talk about it? I'm here if you want to talk about it. I'm here to listen if you need someone to talk to. You're not alone. (Questioning)</i>

Table 6: Examples of responses generated by the models for dialogue prompts from the RED test dataset. The sentences highlighted in red are repetitive statements.

the four models generate responses that are emotionally supportive and empathetic to the given prompt. In the first case, the response generated by EPIMEED+ stands out from the rest since it is lengthier and contains stronger emotional reactions and explorations than the responses generated by the other three models. The same observation could be made for EPIMEED+ when responding to the second dialogue prompt, however, it could be noticed that the response is composed of repetitive statements, which was observed in a number of other responses as well. We believe that this is why workers might have more disagreement when assigning a rating to the responses produced by EPIMEED+. Future work should address such repetitive statements in the responses produced by EPIMEED+, while preserving the empathy-richness brought by emotional reactions, interpretations, and explorations.

7 Discussion and Conclusion

In this paper, we addressed the problem of responding empathetically to distress by curating dialogues from the peer support forum, Reddit, and developing models by adapting and hybridizing existing empathetic response generation architectures and empathy identifying frameworks. The RED dataset

can be used as benchmark to develop similar and better performing chatbot models that can respond to distress. The results of the emotion and intent analysis as well as the automatic and human evaluation results of the experiments conducted on the four baseline chatbot models validate the utility of this dataset in generating emotionally supportive and empathetic responses for distress-related dialogue prompts.

But there are some limitations to this work. Since users responding to distress-related posts in Reddit are not professionals, caution must be taken if these conversations are directly used for training automatic systems that can offer emotional support. Removal of profanity is one step that we have taken towards making such systems reliable and fail-safe. The shift in the emotion of the speaker towards more positive emotions such as gratefulness is also another indicator that the responses do help the speaker lift his/her mood. But deeper analysis such as measuring the level of speaker satisfaction in subsequent dialogue turns and identifying the specific communication techniques that lead to positive outcomes are required when developing an emotionally supportive chatbot based on these conversations. We showed that incorporating existing empathetic frameworks such as EPITOME (Sharma et al., 2020) and conditioning the response on specific empathetic response intents such as in MEED (Xie and Pu, 2021) are good advances in addressing such limitations.

8 Ethics Statement

Analysis of posts of a website like Reddit is likely considered “fair play” as individuals are anonymous, and users can understand their responses remain archived on the site unless taken action to delete them. But a study on user perceptions on social media research ethics (Fiesler and Proferes, 2018) highlights some potential harms that can be caused due to social computing research as internet users can be unaware that the data they share publicly could be used for research. In particular, this dataset contains sensitive information. So, as per Benton et al. (2017)’s guidelines for working with social media data in health research, in this paper, we share only anonymized and paraphrased excerpts from the dataset. References to usernames and URLs are removed from dialogue content for de-identification. The dataset as well as the models are intended for research purposes only.

References

- 559 Tim Althoff, Kevin Clark, and Jure Leskovec. 2016.
560 Large-scale analysis of counseling conversations: An
561 application of natural language processing to mental
562 health. *Transactions of the Association for Compu-
563 tational Linguistics*, 4:463–476.
- 564
- 565 Jason Baumgartner, Savvas Zannettou, Brian Keegan,
566 Megan Squire, and Jeremy Blackburn. 2020. The
567 pushshift reddit dataset. *Proceedings of the Interna-
568 tional AAAI Conference on Web and Social Media*,
569 14(1):830–839.
- 570 Adrian Benton, Glen Coppersmith, and Mark Dredze.
571 2017. Ethical research protocols for social media
572 health research. In *Proceedings of the First ACL
573 Workshop on Ethics in Natural Language Processing*,
574 pages 94–102.
- 575 Arthur C Bohart, Robert Elliott, Leslie S Greenberg,
576 and Jeanne C Watson. 2002. Empathy.
- 577 Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Un-
578 gar, and Joao Sedoc. 2018. Modeling empathy and
579 distress in reaction to news stories. *arXiv preprint
580 arXiv:1808.10399*.
- 581 Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe
582 Kazemzadeh, Emily Mower, Samuel Kim, Jean-
583 nette N Chang, Sungbok Lee, and Shrikanth S
584 Narayanan. 2008. Iemocap: Interactive emotional
585 dyadic motion capture database. *Language resources
586 and evaluation*, 42(4):335.
- 587 Ankush Chatterjee, Uman Gupta, Manoj Kumar Chin-
588 nakotla, Radhakrishnan Srikanth, Michel Galley, and
589 Puneet Agrawal. 2019. Understanding emotions in
590 text using deep learning and big data. *Computers in
591 Human Behavior*, 93:309–317.
- 592 Munmun De Choudhury and Sushovan De. 2014. Men-
593 tal health discourse on reddit: Self-disclosure, so-
594 cial support, and anonymity. In *Eighth international
595 AAAI conference on weblogs and social media*.
- 596 Robert Elliott, Arthur C Bohart, Jeanne C Watson, and
597 David Murphy. 2018. Therapist empathy and client
598 outcome: An updated meta-analysis. *Psychotherapy*,
599 55(4):399.
- 600 Casey Fiesler and Nicholas Proferes. 2018. “participant”
601 perceptions of twitter research ethics. *Social Media+
602 Society*, 4(1).
- 603 Kathleen Kara Fitzpatrick, Alison Darcy, and Molly
604 Vierhile. 2017. Delivering cognitive behavior ther-
605 apy to young adults with symptoms of depression
606 and anxiety using a fully automated conversational
607 agent (woebot): a randomized controlled trial. *JMIR
608 mental health*, 4(2):e7785.
- 609 Clara E Hill. 2009. *Helping skills: Facilitating, explo-
610 ration, insight, and action*. American Psychological
611 Association.
- 612 Sara D Hodges, Kristi J Kiel, Adam DI Kramer, Darya
613 Veach, and B Renee Villanueva. 2010. Giving birth
614 to empathy: The effects of similar experience on
615 empathic accuracy, empathic concern, and perceived
616 empathy. *Personality and Social Psychology Bulletin*,
617 36(3):398–409.
- 618 Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo,
619 Ting-Hao Huang, and Lun-Wei Ku. 2018. Emotion-
620 Lines: An emotion corpus of multi-party conversa-
621 tions. In *Proceedings of the Eleventh International
622 Conference on Language Resources and Evaluation
623 (LREC 2018)*.
- 624 Becky Inkster, Shubhankar Sarda, Vinod Subramanian,
625 et al. 2018. An empathy-driven, conversational artifi-
626 cial intelligence agent (wysa) for digital mental well-
627 being: real-world data evaluation mixed-methods
628 study. *JMIR mHealth and uHealth*, 6(11):e12106.
- 629 Pierre Lison, Jörg Tiedemann, Milen Kouylekov, et al.
630 2019. Open subtitles 2018: Statistical rescoring of
631 sentence alignments in large, noisy parallel corpora.
632 In *LREC 2018, Eleventh International Conference
633 on Language Resources and Evaluation*. European
634 Language Resources Association (ELRA).
- 635 Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand
636 Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie
637 Huang. 2021. Towards emotional support dialog
638 systems. In *Proceedings of the 59th Annual Meet-
639 ing of the Association for Computational Linguistics
640 and the 11th International Joint Conference on Natu-
641 ral Language Processing (Volume 1: Long Papers)*,
642 pages 3469–3483. Association for Computational
643 Linguistics.
- 644 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man-
645 dhar Joshi, Danqi Chen, Omer Levy, Mike Lewis,
646 Luke Zettlemoyer, and Veselin Stoyanov. 2019.
647 Roberta: A robustly optimized bert pretraining ap-
648 proach. *arXiv preprint arXiv:1907.11692*.
- 649 Gary McKeown, Michel Valstar, Roddy Cowie, Maja
650 Pantic, and Marc Schroder. 2011. The semaine
651 database: Annotated multimodal records of emotion-
652 ally colored conversations between a person and a
653 limited agent. *IEEE transactions on affective com-
654 puting*, 3(1):5–17.
- 655 Seyed Mahed Mousavi, Alessandra Cervone, Morena
656 Danieli, and Giuseppe Riccardi. 2021. Would you
657 like to tell me more? generating a corpus of psy-
658 chotherapy dialogues. In *Proceedings of the Second
659 Workshop on Natural Language Processing for Medi-
660 cal Conversations*, pages 1–9.
- 661 Fionn Murtagh and Pierre Legendre. 2014. Ward’s hi-
662 erarchical agglomerative clustering method: which
663 algorithms implement ward’s criterion? *Journal of
664 classification*, 31(3):274–295.
- 665 Soujanya Poria, Devamanyu Hazarika, Navonil Ma-
666 jumder, Gautam Naik, Erik Cambria, and Rada Mi-
667 halcea. 2019. MELD: A multimodal multi-party

- 668 dataset for emotion recognition in conversations. In 721
669 *Proceedings of the 57th Annual Meeting of the Association 722*
670 for Computational Linguistics, pages 527–536. 723
671
672 Hannah Rashkin, Eric Michael Smith, Margaret Li, and 724
673 Y-Lan Boureau. 2018. Towards empathetic open-
674 domain conversation models: A new benchmark and
dataset. *arXiv preprint arXiv:1811.00207*.
- 675 Nils Reimers and Iryna Gurevych. 2019. Sentence- 727
676 BERT: Sentence embeddings using Siamese BERT- 728
677 networks. In *Proceedings of the 2019 Conference 729*
678 on Empirical Methods in Natural Language Process- 730
679 ing and the 9th International Joint Conference on 731
680 Natural Language Processing (EMNLP-IJCNLP)
- 681 Robert L Selman. 1981. The development of inter- 732
682 personal competence: The role of understanding in 733
683 conduct. *Developmental review*, 1(4):401–422.
- 684 Ashish Sharma, Adam Miner, David Atkins, and Tim Al- 732
685 thoff. 2020. A computational approach to understand- 733
686 ing empathy expressed in text-based mental health 734
687 support. In *Proceedings of the 2020 Conference on 735*
688 Empirical Methods in Natural Language Processing 736
689 (EMNLP), pages 5263–5276.
- 690 Ratika Sharma, Britta Wigginton, Carla Meurk, Pauline 737
691 Ford, and Coral E Gartner. 2017. Motivations and 738
692 limitations associated with vaping among people with 739
693 mental illness: A qualitative analysis of reddit dis- 740
694 cussions. *International journal of environmental re- 741*
695 search and public health, 14(1):7.
- 696 Jennifer Smith-Merry, Gerard Goggin, Andrew Camp- 737
697 bell, Kirsty McKenzie, Brad Ridout, Cherry Baylo- 738
698 sis, et al. 2019. Social connection and online en- 739
699 gagement: insights from interviews with users of a 740
700 mental health online forum. *JMIR mental health*, 741
701 6(3):e11084.
- 702 Zachary Steel, Claire Marnane, Changiz Iranpour, Tien 742
703 Chey, John W Jackson, Vikram Patel, and Derrick 743
704 Silove. 2014. The global prevalence of common 744
705 mental disorders: a systematic review and meta-analysis 745
706 1980–2013. *International journal of epidemiology*, 746
707 43(2):476–493.
- 708 Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. 747
709 Sequence to sequence learning with neural networks. 748
710 *arXiv preprint arXiv:1409.3215*.
- 711 Richard Thwaites and James Bennett-Levy. 2007. Con- 749
712 ceptualizing empathy in cognitive behaviour therapy: 750
713 Making the implicit explicit. *Behavioural and Cog- 751*
714 nitive Psychotherapy, 35(5):591–612.
- 715 Aditya Nrusimha Vaidyam, Hannah Wisniewski, 749
716 John David Halamka, Matcheri S Kashavan, and 750
717 John Blake Torous. 2019. Chatbots and conversa- 751
718 tional agents in mental health: a review of the psychi- 752
719 atric landscape. *The Canadian Journal of Psychiatry*, 753
720 64(7):456–464.
- Amrisha Vaish, Tobias Grossmann, and Amanda Wood- 721
ward. 2008. Not all emotions are created equal: the 722
negativity bias in social-emotional development. *Psy- 723
chological bulletin*, 134(3):383.
- Oriol Vinyals and Quoc Le. 2015. A neural conversa- 725
tional model. *arXiv preprint arXiv:1506.05869*.
- Anuradha Welivita and Pearl Pu. 2020. A taxonomy of 727
empathetic response intents in human social conversa- 728
tions. In *Proceedings of the 28th International Con- 729*
ference on Computational Linguistics, pages 4886– 730
4899.
- Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A 732
large-scale dataset for empathetic response genera- 733
tion. In *Proceedings of the 2021 Conference on 734*
Empirical Methods in Natural Language Processing, 735
pages 1251–1264.
- Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei- 737
Hao Su, David Vandyke, and Steve Young. 2015. 738
Semantically conditioned LSTM-based natural lan- 739
guage generation for spoken dialogue systems. In 740
Proceedings of the 2015 Conference on Empirical 741
Methods in Natural Language Processing, pages 742
1711–1721.
- Yubo Xie and Pearl Pu. 2021. Empathetic dialog gen- 744
eration with fine-grained intents. In *Proceedings of 745*
the 25th Conference on Computational Natural Lan- 746
guage Learning, pages 133–147.
- Justine Zhang and Cristian Danescu-Niculescu-Mizil. 748
2020. Balancing objectives in counseling conversa- 749
tions: Advancing forwards or looking backwards. In 750
ACL.
- Victor Zhou, Domitrios Mistriotis, and Vadim 752
Shestopalov. 2020. [profanity-check](#).

A Topic Coverage

We used automatic clustering to identify clearly distinguishable topics present in the Reddit distress dialogues. For this purpose, we used “Agglomerative Clustering” tuned for large datasets ([Murtagh and Legendre, 2014](#)). It recursively merges pairs of clusters that minimally increase a given linkage distance. The linkage distance was computed using the cosine similarity between pairs of embeddings generated by Sentence-BERT ([Reimers and Gurevych, 2019](#)) since the resulting embeddings have shown to be of high quality and working substantially well for document-level embeddings.

We experimented with 8 similarity thresholds from 0.6 to 0.95 with 0.05 increments to cluster distress prompts. At an optimal threshold of 0.85 identified by manual inspection of a randomly selected subset of 10 clusters resulted in 4.93% of the distress prompts (47, 109 prompts in total) getting clustered into 4, 363 clearly identifiable clusters. After applying TF-IDF-based topic modeling to these clusters, clearly distinguishable topics were uncovered. Table 7 shows some distress-related topics and their corresponding keywords.

Distress topic	Keywords
Suicidal	<i>commit, killing, death, painless, option</i>
Anxiety attacks	<i>anxiety, anxious, attacks, social, attack</i>
Weight gain	<i>eating, weight, eat, lose, fat</i>
Loneliness	<i>lonely, surround, connect, isolated, social</i>
Failing college	<i>study, college, class, semester, failing</i>
Alcoholic	<i>drinking, drink, alcohol, drunk, sober</i>
US election	<i>trump, president, donald, election, war</i>
Covid19	<i>covid, 19, pandemic, shambolic, brought</i>

Table 7: Some distress-related topics identified in the RED dataset along with corresponding keywords.

B Human Evaluation Experiment

In the human evaluation experiment, randomly selected 200 dialogues were bundled into 20 HITs (Human Intelligent Tasks) with each HIT containing 10 such dialogues. Three workers were assigned per HIT. To evaluate the workers’ attentiveness to the task, we randomly inserted 3 checkpoints among the 10 dialogues by including the ground-truth response to be rated among the other chatbot-generated responses. Ideally, the ground-truth response should be rated either as *Good* or *Okay* by the workers. If a worker was able to pass at least 2 out of the 3 checkpoints, he was offered

Figure 5: Dialogue response rating interface. A drag and drop mechanism was implemented for ease of rating dialogue responses by the crowdworkers.

a bonus of \$0.1 in addition to the compensation of \$0.4 offered for completing an assignment. The bonus compensation was used as a mean to motivate workers to pay attention to the task and deliver quality work. In addition, workers who accepted more than 10 HITs were blocked automatically to avoid one worker monopolising the task. We also monitored the total time a worker takes to complete an assignment and rejected all the assignments that were completed in less than 2 minutes and added them back to the pool of available assignments. This enabled us to avoid ratings generated by inattentive workers and spammers. Figure 5 shows the task interface. An overall inter-annotator agreement score (Fleiss’ kappa) of 0.2294 was observed, indicating fair agreement between the workers.

C Example Dialogue Responses

Table 8 shows some additional example dialogue responses generated by the four models EmoPrepend, MEED, EPIMEED, and EPIMEED+ for dialogue prompts from the RED test dataset.

Speaker: Back then I used to be all cheery, and went through my day. Now I just sit at my chair, looking down on myself, and sleeping days away. I reflect on myself no matter what, especially how I am around other people. I just lost all drive to do anything. I just want to lay down, and wish death.

EmoPrepend	I'm so sorry you're going through this. You talked to anyone in your life about any of the stuff you're going through?
MEED	What kind of stuff is making you feel this way? (Questioning)
EPIMEED	Do you want to talk about it? (Questioning)
EPIMEED+	I've been in a similar situation. You're not alone. Did you tell anyone what you're going through or looked into getting help? (Questioning)

Speaker: I want out. I want to shoot someone, I want to be shot. I don't know what the fuck I want, I know exactly what I want. I want someone to shoot me, I don't want a new prescription doc, I want to be f***ing shot.

EmoPrepend	What country do you live in?
MEED	Why do you want to do this? (Questioning)
EPIMEED	Why would you want to do this? (Questioning)
EPIMEED+	What's going on in your life? I'm here to talk if you want to talk. PM me if you need someone to talk to. (Questioning)

Speaker: I don't have one friend, when my family come into this house all they do is look threw me. No one talks to me unless it's something perverted online. I feel terrible.

EmoPrepend	I'm sorry you're going through this.
MEED	How old are you? (Questioning)
EPIMEED	I'm here if you want to talk. What's going on? (Questioning)
EPIMEED+	What's making you feel this way? I'm here if you need someone to talk to. (Questioning)

Speaker: I'm dealing with severe anxiety. I feel like I'm failing my fiance. Sorry for rambling but I just feel like giving up. I just feel like I can't win.

Listener: I'm going through a similar situation. I also feel like I'm failing at life.

Speaker: Well, I guess we can at least go down swinging. At least I'm not alone.

EmoPrepend	I'm sorry you're going through this.
MEED	You're not alone. (Agreeing)
EPIMEED	That's a good point. (Acknowledging)
EPIMEED+	You're not alone. I'm in the same boat. I can relate to this. I wish you'd be happy to talk to me if you want to talk about it. (Acknowledging)

Speaker: I just want to be dead, I don't want to live in this shitty world, I want to be released. I'm planning to kill myself later in life. Nobody will notice about my success.

EmoPrepend	I'm so sorry you're going through this. I hope things get better for you soon.
MEED	Why do you want to die? (Questioning)
EPIMEED	Please don't. Why do you feel this way? (Questioning)
EPIMEED+	Why do you want to kill yourself? Why do you feel like you want to do this? Have you told anyone what you're going through or looked into getting help? If you want to talk I'm here. (Questioning)

Speaker: They say one person eventually gets attached, Funny how it's always me..

Listener: How did it become that?

Speaker: Cos I'm a very emotional person who gets attached very easily. I eventually become some side pet in any relationship.

EmoPrepend	I'm so sorry you're going through this.
MEED	Are you in an open relationship? (Questioning)
EPIMEED	That's why I don't want any relationship. (Neutral)
EPIMEED+	I'm sure you're not the only one. I'm here to talk if you need to talk. (Neutral)

Table 8: Examples of responses generated by the models for dialogue prompts from the RED test dataset.