Empathy Identification Systems are not Accurately Accounting for Context

Andrew Lee[†] **Jonathan Kummerfeld**[†‡] **Larry An**[†] **Rada Mihalcea**[†]
[†] University of Michigan [‡] University of Sydney
ajyl@umich.edu

Abstract

Understanding empathy in text dialogue data is a difficult, yet critical, skill for effective human-machine interaction. In this work, we ask whether systems are making meaningful progress on this challenge. We consider a simple model that checks if an input utterance is similar to a small set of empathetic examples. Crucially, the model does not look at what the utterance is a response to, i.e., the dialogue context. This model performs comparably to prior work on standard benchmarks and even outperforms state-of-the-art models for empathetic rationale extraction by 16.7 points on T-F1 and 4.3 on IOU-F1. This indicates that current systems rely on the *surface form* of the response, rather than whether it is suitable in context. To confirm this, we create examples with dialogue contexts that change the interpretation of the response and show that current systems continue to label utterances as empathetic. We discuss the implications of our findings, including improvements for empathetic benchmarks and how our model can be an informative baseline.

1 Introduction

Empathy is a fundamental phenomenon that allows us to better communicate and relate with others. Studies show that empathy is significantly correlated with counseling treatment outcomes (Moyers and Miller, 2013; Elliott et al., 2018). Computer systems could be improved by adding the ability to understand empathy.

EPITOME (Sharma et al., 2020b) took a step towards understanding empathy in language, introducing tasks such as *empathy identification* and *empathetic rationale extraction*. Models built using EPITOME have been used to build or evaluate empathetic dialogue systems (Sharma et al., 2021; Zheng et al., 2021; Majumder et al., 2022; Kim et al., 2021) or study social effects of empathy (Chen and Xu, 2021).

In this work, we explore whether current models are effectively considering dialogue context (shortened to *context* for the rest of this paper). We show that a simple model that does not consider context can achieve strong results, and that a model from prior work does not change its predictions when we make substantial changes to the context. Together, these results indicate that models are more limited than previously thought.

We introduce an adapted version of *micromodels* (Lee et al., 2021), a simple and explainable approach that combines a set of models, with each model identifying a specific linguistic phenomenon. This approach performs much better than the EPITOME's model on five metrics, comparably on five metrics, and much worse on two. Critically, we achieve this *without any use of context*.

We inspect our model's behavior and find that it can achieve accuracy scores that are as good as or better than the EPITOME model's for empathetic rationale extraction with as few as three seed/training utterances for our model.

We also conduct an experiment to probe EPITOME's behavior. We take utterances from empathetic responses and randomly insert them as part of the response in another context. Despite these insertions mainly being nonsensical and non-empathetic, prior models nearly *always* predict these responses as empathetic, demonstrating that the models rely on the *surface form* of the response rather than contextual understanding.

The authors of EPITOME noted that empathy is *contextual*; a "reaction to an emotional stimulation" or a "deliberate process of understanding and interpreting the experiences" of others.[1] However, current systems do not effectively account for context: they may identify empathetic *style*, but they do not consider whether a response is indeed a *reaction* to feelings and experiences. Future work

[1] Section 2.1 of Sharma et al. (2020b)

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1686–1695

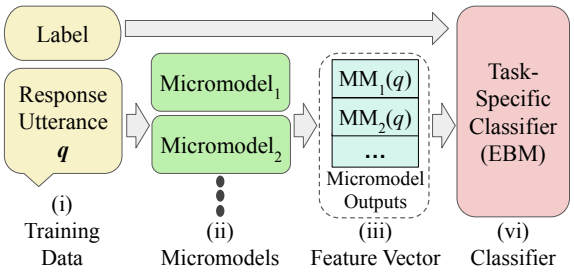

Figure 1: The micromodel framework training process. Notably, our approach does not use any *dialogue context*, yet still performs strongly on current empathy detection benchmarks, raising the question of whether current systems and benchmarks are accounting for context.

should conduct probing experiments like the one we use here, and could consider micromodels as a baseline. This will help assess the contextual understanding of empathy for future models.

2 Benchmarks and Tasks: EPITOME

EPITOME (Sharma et al., 2020b) is a framework for computationally assessing empathy in text-based dialogue. We denote their dataset as $\text{EPITOME}_{\text{Data}}$ and the model as $\text{EPITOME}_{\text{Model}}$.

$\text{EPITOME}_{\text{Data}}$ consists of pairs of dialogues from Talklife and Reddit, and two tasks: empathy prediction (**EmpPred**) and empathetic rationale extraction (**EmpRE**). Given a seeker's post $S_i = s_{i1}, ..., s_{im}$ and a response $R_i = r_{i1}, ..., r_{in}$, each response R_i is annotated with an empathy level (None, Weak, or Strong) in the context of S_i across three communication mechanisms: Emotional Reactions, Interpretations, and Explorations.[2]

Empathetic rationales are spans of text that provide evidence of empathy. They are annotated at the token-level, e.g., the response *"I feel you. Are you okay?"* is represented as $[1, 1, 1, 0, 0, 0]$, $[0, 0, 0, 1, 1, 1]$, and $[0, 0, 0, 0, 0, 0]$ for Emotional Reactions, Exploration, and Interpretations, with one digit per token.

The goal of **EmpPred** is to predict the correct level of empathy given (S_i, R_i) across each communication mechanism. The goal of **EmpRE** is to correctly extract the rationale spans.

3 Micromodels

Lee et al. (2021) introduced the micromodel framework to assess the mental health status of social media users. We give a brief overview of the framework, followed by our adaptations to tackle each task. Further details are provided in the appendix.

3.1 Micromodel Framework

Figure 1 depicts the micromodel framework. The framework consists of a set of micromodels, in which each micromodel MM is a binary classifier that is initialized with a set of *seed* utterances $Z = z_1, ..., z_n$. Given an input query q, micromodel MM gives a binary prediction if q is semantically similar to any of the seed utterances in Z:

$$MM(q) = \exists_{z \in Z} CosSim(BERT(q), BERT(z)) > \theta \quad (1)$$

The outputs of the micromodels are used as features to train a task-specific classifier. Lee et al. (2021) uses explainable boosting machines (EBM) (Caruana et al., 2015), which are generalized additive models (Lou et al., 2012) that make predictions based on adding a set of feature functions learned on each input feature, where each feature function is trained using bagging and gradient boosting.

3.2 Micromodels for EmpPred

For EPITOME's tasks, we build three micromodels, one for each communication mechanism c. For each micromodel MM_c, the seed utterances Z_c are initialized using the annotated rationales of each communication mechanism in the training split of $\text{EPITOME}_{\text{Data}}$. Once initialized, rather than using the binary outputs from each micromodel, we use the maximum similarity score between each sentence from the response post $r_{ij} \in R_i$ and each of the seed utterances $z \in Z_c$.

$$MM_c(R_i) = \max_{\substack{r_{ij} \in R_i \\ z \in Z_c}} (Sim(BERT(r_{ij}), BERT(z))) \quad (2)$$

We use the resulting similarity scores as features to train an EBM model[3] to predict the empathy level. We use S-BERT (Reimers and Gurevych, 2019) models[4] to compute similarity scores.

3.3 Micromodels for EmpRE

Figure 2 depicts how we apply micromodels to extract empathetic rationales. Given a response post R_i, we first split it into sentences $r_{i1}, ..., r_{in}$. Each micromodel MM_c runs on each sentence r_{ij}, returning 1 if sentence r_{ij} is semantically similar to any of the seed utterances Z_c and 0 otherwise. This results in a binary vector v_c of length n. Each

[2]The definition of each communication mechanism can be found in the appendix.

[3]https://github.com/interpretml/interpret

[4]"paraphrase-xlm-r-multilingual-v1"

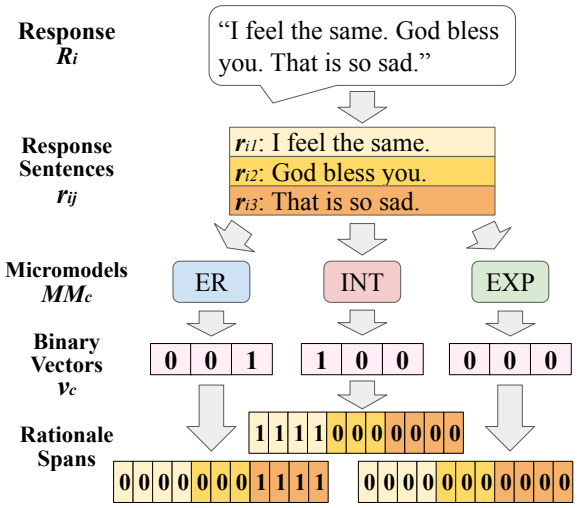

Figure 2: Extracting empathy rationales using micromodels. Each micromodel MM_c determines if each response sentence r_{ij} is empathetic. Each token w_k^{ij} of sentence r_{ij} is then assigned the binary value of r_{ij}.

sentence r_{ij} is then tokenized[5] into a set of tokens $w_1^{ij}, ..., w_l^{ij}$, where $l = len(r_{ij})$ and each token w_k^{ij} is assigned the binary value of r_{ij}. This results in a sequence of 0's and 1's where spans of 1's represent rationales.

4 Experiments and Results

4.1 Experimental Setup

For our experiments, we use random splits of 75:5:20 for our train, validation, and test sets.[6] We report average scores from 10 runs. For θ (Eq. 1), we use a threshold value of 0.7, based on experiments from the validation set. Following EPITOME's authors, we report token-level F1 (T-F1) and Intersection Over Union F1 (IOU-F1) scores.[7]

4.2 Baselines: EPITOME$_{\text{Model}}$

We compare our approach to several baselines, including EPITOME$_{\text{Model}}$. EPITOME$_{\text{Model}}$ is a multi-task bi-encoder model initialized with the weights of RoBERTa. Each encoder encodes the seeker's post S_i and the response post R_i. An attention layer attends over both encodings, which is then jointly trained on the two EPITOME tasks. Further details are provided in the appendix.

[5]We use NLTK for tokenization

[6]This is the same ratio used in the original EPITOME paper. There were no official splits.

[7]We use the same IOU match threshold as EPITOME (0.5).

4.3 Baselines: Other

We also include baseline results as reported by the authors of EPITOME. These baseline models include popular models used in similar tasks, each of which have been fine-tuned or trained on the EPITOME tasks:

- Logistic regression over tf-idf vectors
- Recurrent neural network
- Hierarchical recurrent encoder-decoder (HRED, Sordoni et al. (2015))
- BERT (Devlin et al., 2019)
- GPT-2 (Radford et al.)
- DialoGPT (GPT-2 adapted for dialogue, Zhang et al. (2020))
- RoBERTa (Liu et al., 2019)

4.4 Results

EmpPred$_{\text{EPIT}}$. The first six columns of Table 1 show the accuracy and F-1 scores of empathy prediction. While our F-1 scores are lower than EPITOME$_{\text{Model}}$, they often outperform other fine-tuned language models.

EmpRE. The last six columns of Table 1 show the T-F1 and IOU-F1 scores for empathetic rationale extraction. Other than for Interpretations, we demonstrate significant improvements of up to 16.7 points for T-F1 and 4.3 points for IOU-F1, resulting in the highest scores to our knowledge.

4.5 Follow-Up Analyses

Probing our model: As a post-analysis, we examine which seed utterances $z \in Z_c$ trigger each micromodel during testing and observe that only a small subset of seed utterances are meaningfully used. To demonstrate this, we conduct an experiment in which we iteratively reduce the number of seed utterances used by each micromodel based on how frequently they trigger a micromodel during testing. Figure 3 shows the resulting IOU-F1 scores[8] from 10 random runs, demonstrating that for some communication mechanisms, state-of-the-art results can be achieved by simply checking for semantic similarity against as few as three seed utterances, which are shown in Table 2. Note, this analysis requires modifying the training based on the test performance, so the results are not necessarily representative beyond that dataset.

Probing EPITOME$_{\text{Model}}$: We run an experiment to study whether these utterances are also driving EPITOME$_{\text{Model}}$'s behavior. We gather 1,000

[8]T-F1 showed nearly identical patterns

Model	Empathy Prediction						Empathetic Rationale Extraction					
	Emotional Reactions		Interpretations		Explorations		Emotional Reactions		Interpretations		Explorations	
	Acc.	F-1	Acc.	F-1	Acc.	F-1	T-F1	IOU-F1	T-F1	IOU-F1	T-F1	IOU-F1
Majority	66.38	26.60	54.58	23.54	83.90	30.41	66.98	66.98	54.94	54.94	84.53	84.53
Log. Reg.	41.69	42.69	70.58	49.77	67.08	46.63	43.26	61.27	49.85	31.31	48.21	70.36
RNN	71.63	42.85	76.21	51.76	85.58	30.74	45.54	43.94	48.22	51.35	65.11	78.27
HRED	71.11	44.10	79.65	54.16	85.58	30.74	46.34	45.65	48.88	52.12	66.66	80.33
BERT	72.13	50.41	82.16	61.20	89.35	56.54	51.06	54.81	48.38	50.75	67.91	71.00
GPT-2	76.69	71.65	82.32	62.27	88.25	58.28	51.44	57.10	54.53	52.38	73.39	82.89
DialoGPT	66.07	51.16	81.85	**68.95**	89.65	70.65	51.83	49.37	54.43	55.85	73.43	85.20
RoBERTa	76.99	70.35	82.16	61.38	90.58	63.41	51.89	58.31	55.62	54.60	69.76	83.33
EPITOME$_{\text{Model}}$	79.43	**74.46**	84.04	62.60	92.61	**72.58**	53.57	64.83	**57.40**	**55.90**	71.56	84.48
Micromodels	**88.26**	59.52	**92.71**	62.73	**95.27**	61.47	**70.30**	**69.13**	54.94	54.08	**86.92**	**86.64**

Table 1: Performance on empathy prediction and empathetic rationale extraction.

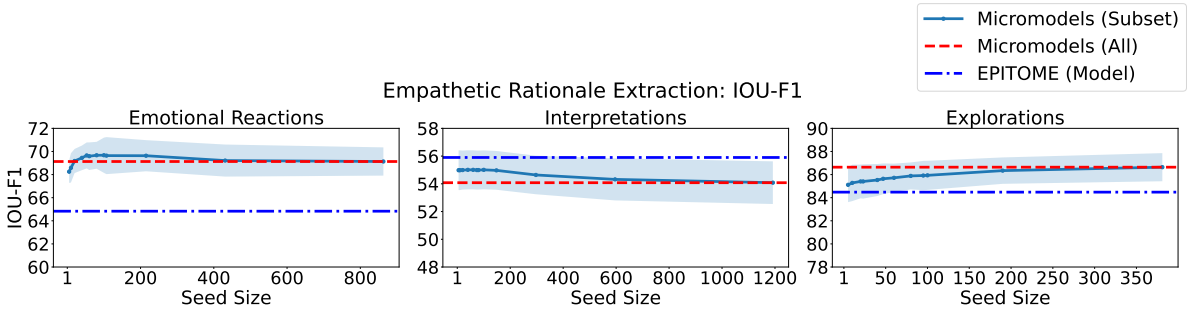

Figure 3: F1 scores with varying seed sizes per micromodel. Simply checking for semantic similarity with as few as three utterances in Table 2 demonstrates either better or competitive performance scores compared to previous state-of-the-art models. The shaded light blue regions indicate the standard deviation across our 10 runs.

Communication Mechanism	Seed Utterance
Emotional Reactions	"I know how you feel." "I'm sorry." "I feel you."
Interpretations	"I feel the same way." "I know how you feel." "I understand how you feel."
Explorations	"Why?" "What happened?" "Why do you feel like that?"

Table 2: Utterances from the smallest subset of seed data used in Figure 3. Simply checking for semantic similarity between response utterances and these seed utterances may outperforms prior state-of-the-art models in empathetic rationale extraction.

Is Empathetic?	PersonaChat		Ubuntu	
	No	Yes	No	Yes
Emotional Reactions	3	997	8	992
Interpretations	305	695	556	444
Interpretations†	10	990	11	989
Explorations	31	969	6	994

Table 3: Number of times non-empathetic dialogues are predicted as empathetic by EPITOME$_{\text{Model}}$. †indicates when we always insert "I feel the same way" (The most commonly seen seed utterance for Interpretations – see Table 2).

random conversations from PersonaChat (Zhang et al., 2018), an open-domain dialogue dataset, and Ubuntu Dialogue (Lowe et al., 2015), a dialogue dataset around technical support for Ubuntu-related problems. For each sample, we insert one of the three utterances from Table 2 as part of the response, resulting in a nonsensical and non-empathetic response. Table 3 shows the results, with EPITOME$_{\text{Model}}$ almost always predicting these artificial dialogues as empathetic. This indicates that it relies on the *surface form* of the response for its prediction, regardless of context.

5 Related Work

Limited access to treatment for mental health, along with a rise in demand for scalable yet high-quality interventions (Miner et al., 2019), has led to an abundance of conversational systems that provide mental health support (Shen et al., 2020; Welch et al., 2020; Han et al., 2013).[9][10] A critical capability for these systems is to understand and interact with empathy, as studies show that

[9]https://woebothealth.com/
[10]https://www.wysa.io/

empathy is significantly correlated with counseling treatment outcomes (Moyers and Miller, 2013; Elliott et al., 2018).

Broadly, recognizing empathy within dialogue has been studied under the following contexts: online platforms (Sharma et al., 2020a, 2021; Khanpour et al., 2017), formal counseling settings (Gibson et al., 2015; Zhang and Danescu-Niculescu-Mizil, 2020; Pérez-Rosas et al., 2017), or social media interactions (Hosseini and Caragea, 2021; Lahnala et al., 2021; Wang and Jurgens, 2018; Zhou and Jurgens, 2020).

For instance, Lahnala et al. (2021) examined the interactions between practitioners and non-practitioners that provide support on Reddit. Wang and Jurgens (2018) and Zhou and Jurgens (2020) analyzed the language of condolence and empathy in various social platforms. Other settings for assessing empathy include reacations to news stories (Buechel et al., 2018).

Another direction in the study of empathy within dialogue includes empathetic response generation (Rashkin et al., 2019; Liu et al., 2021; Zhong et al., 2020; Zheng et al., 2021). In order to assess the empathy level of their systems, researchers often use models such as EPITOME. (Sharma et al., 2021; Zheng et al., 2021; Majumder et al., 2022; Kim et al., 2021). Our work demonstrates the pitfalls that researchers should be aware of when taking such approach.

Other efforts include a taxonomy of empathetic responses (Welivita and Pu, 2020), fine-tuned language models for empathy (Guda et al., 2021), as well as empathy-lexicons (Sedoc et al., 2020).

6 Conclusion

In this paper, we assessed whether empathetic systems are correctly taking dialogue context into account. We demonstrated that a simple model with no contextual understanding can achieve results comparable to the EPITOME model and better than all baselines. We find that these results can be achieved by simply checking for semantic similarity to just three utterances. We also found that $\text{EPITOME}_{\text{Model}}$ nearly always classifies a response as empathetic regardless of context, as long as it contains one of these three utterances. We conclude that current empathy recognition models do not effectively take contextual information into account.

Future work on benchmarks should consider including examples that require contextual understanding to answer (S: *"I got promoted!"* R: *"That's terrible, I'm sorry."*). Work on models should consider comparing with Micromodels, a simple and practical baseline that serves as a reference point for performance without contextual understanding. These changes will better inform progress on models that meaningfully capture empathy.

The code for our experiments is publicly available at `https://github.com/MichiganNLP/micromodels`.

7 Limitations

The main focus of our paper is on the limitations of empathy recognition models, both our micromodel approach as well as the prior state-of-the-art model $\text{EPITOME}_{\text{Model}}$. Namely, our micromodel approach is based on semantic similarity matching, and lacks any representation learning and contextual knowledge. On the other hand, our experiments demonstrate that $\text{EPITOME}_{\text{Model}}$ also does not account for context. Despite given non-empathetic contexts, it continues to predict a response as empathetic as long as the *style* of empathy is present. Other limitations of our work include the scope of our study, as we only examine a single benchmark. This is due to the lack of available resources regarding the task of empathy recognition in dialogue. Datasets like EmpatheticDialogue (Rashkin et al., 2019) consists of empathetic conversations, but do not measure the empathy level of utterances. Other empathy prediction tasks (Hosseini and Caragea, 2021; Buechel et al., 2018) do not pertain to dialogue. With more benchmarks regarding empathy recognition in dialogue available, a more thorough study should be conducted.

8 Ethical Considerations

It is important to distinguish our improved accuracy scores from the ability to computationally understand empathy. Because of the lack of representational learning or contextual knowledge, our approach would undoubtedly fail in distinguishing empathetic utterances from false-positive cases, such as sarcastic or even offensive statements (R: *"What's the matter?"* versus *"What's the matter with you?"*). Given the sensitive nature of the mental health domain, mishandling these situations can exacerbate one's situation. Another risk of relying too heavily on such accuracy numbers include overlooking the degree to which a mishandling of

a situation can affect an afflicted user. Measuring this additional personal and humanistic dimension in benchmarks for computational systems is undeniably a difficult problem, but likely a necessary step to bridge the gap for effective systems for mental health support. Lastly, while we are able to do a thorough analysis of our findings with the explanations provided by micromodels, there are still portions of their decision making process that remain opaque. Concretely, the computation of semantic similarity by large pre-trained language models like BERT is a key step in our procedure. Using a simpler, more transparent representation for micromodels may mitigate this problem. We believe there is an interesting trade off between accuracy and explainability in designing each micromodel.

Acknowledgments

This material is based in part upon work supported by the John Templeton Foundation (#62256) and by a Berman Research award from the University of Michigan. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the John Templeton Foundation or the University of Michigan.

References

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1721–1730, New York, NY, USA.

Yixin Chen and Yang Xu. 2021. Exploring the effect of social support and empathy on user engagement in online mental health communities. *International Journal of Environmental Research and Public Health*, 18(13):6855.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399.

James Gibson, Nikolaos Malandrakis, Francisco Romero, David C Atkins, and Shrikanth S Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Sixteenth annual conference of the international speech communication association*.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. EmpathBERT: A BERT-based framework for demographic-aware empathy prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3072–3079, Online. Association for Computational Linguistics.

Sangdo Han, Kyusong Lee, Donghyeon Lee, and Gary Geunbae Lee. 2013. Counseling dialog system with 5W1H extraction. In *Proceedings of the SIGDIAL 2013 Conference*, pages 349–353, Metz, France. Association for Computational Linguistics.

Mahshid Hosseini and Cornelia Caragea. 2021. Distilling knowledge for empathy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3713–3724, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 246–251, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Allison Lahnala, Yuntian Zhao, Charles Welch, Jonathan K. Kummerfeld, Lawrence C An, Kenneth Resnicow, Rada Mihalcea, and Verónica Pérez-Rosas. 2021. Exploring self-identified counseling expertise in online support forums. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4467–4480, Online. Association for Computational Linguistics.

Andrew Lee, Jonathan K. Kummerfeld, Larry An, and Rada Mihalcea. 2021. Micromodels for efficient, explainable, and reusable systems: A case study on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4257–4272, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2022. Exemplars-guided empathetic response generation controlled by the elements of human communication. *IEEE Access*, 10:77176–77190.

Adam S Miner, Nigam Shah, Kim D Bullock, Bruce A Arnow, Jeremy Bailenson, and Jeff Hancock. 2019. Key considerations for incorporating conversational ai in psychotherapy. *Frontiers in psychiatry*, 10:746.

Theresa B Moyers and William R Miller. 2013. Is low therapist empathy toxic? *Psychology of Addictive Behaviors*, 27(3):878.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2020. Learning word ratings for empathy and distress from document-level user responses. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1664–1673, Marseille, France. European Language Resources Association.

Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020a. Engagement patterns of peer-to-peer interactions on mental health platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 614–625.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020b. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20, 1st virtual meeting. Association for Computational Linguistics.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *proceedings of the 24th ACM international on conference on information and knowledge management*, pages 553–562.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45, Brussels, Belgium. Association for Computational Linguistics.

Charles Welch, Allison Lahnala, Veronica Perez-Rosas, Siqi Shen, Sarah Seraj, Larry An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2020. Expressive interviewing: A conversational system for coping with COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP*

2020, Online. Association for Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5276–5289, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. CoMAE: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824, Online. Association for Computational Linguistics.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566.

Naitian Zhou and David Jurgens. 2020. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 609–626, Online. Association for Computational Linguistics.

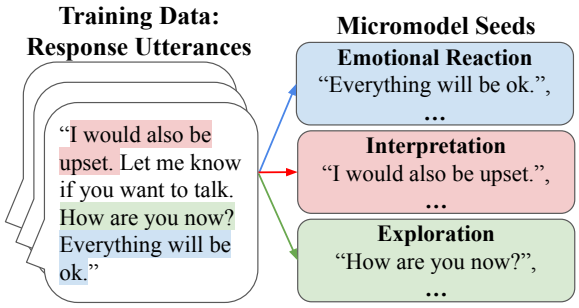

Figure 4: The annotated rationales in the training split of $\text{EPITOME}_{\text{Data}}$ are used as the seed data for each micromodel.

A Empathetic Communication Mechanisms

The EPITOME (Sharma et al., 2020b) framework introduces three communication mechanisms to capture the multi-dimensionality of empathy in text-based dialogue – *Emotional Reactions*, *Interpretations*, and *Explorations*. The definitions and examples of each communication mechanism according to the original EPITOME paper can be found below.

Emotional Reactions Expressions of emotions such as warmth, compassion as a response to the seeker's post. These expressions can explicitly label an emotion (e.g., *"I feel really sad for you."*) or may allude to an emotion (e.g., *"Everything will be fine."*).

Interpretations A reactive statement of one's own understanding of feelings and experiences inferred from the seeker's post. Such statement may simply state their understanding (e.g., *"I understand how you feel."*) or specify their inferred feelings or similar experiences (e.g., *"This must be terrifying.", "I also have anxiety attacks at times which makes me really terrified."*).

Explorations Seeking further information to improve one's understanding of the seeker and their feelings and experiences. These can include generic follow-ups (e.g., *"what happened?"*) or specific inquiries (e.g., *"Are you feeling alone right now?"*).

B Detailed Explanation of Micromodels

Micromodels (Lee et al., 2021) were originally inspired by recent work in microservice architectures, in which complex web applications are built by orchestrating a collection of loosely coupled *microservices*. Each of these microservices has a fine-grained focus of responsibility. In a similar manner, the micromodel framework consists of multiple *micromodels*, in which each micromodel is responsible for representing or identifying a specific linguistic phenomena. In this work we build a micromodel for each communication mechanism.

Here we describe the training procedure using micromodels.

The first step in the framework is to initialize each micromodel. The original authors scrape Reddit and use BERT to look for utterances that are representative of each micromodel. In this work, we simply use the annotated rationales that are available in the training split of EPITOME (see Figure 4). Each micromodel only needs to be initialized once.

Next, given supervised training data in the form of (x, y), each micromodel MM runs on the input query x. While any algorithm of choice can be used for micromodels, the original authors use binary classifiers based on semantic similarity. Concretely, each micromodel that is initialized with a set of seed utterances $Z = z_1, ..., z_n$ makes a binary decision, returning 1 if its input query q is semantically similar to any of the seed utterances $z \in Z$:

$$MM(x) = \exists_{z \in Z} CosSim(BERT(x), BERT(z)) > \theta \quad (3)$$

In this work we use our validation set to determine the θ value.

Once every micromodel runs on x, we are left with a binary vector $\mathbf{v}$ of size n where n is the number of micromodels that were used. The binary vector $\mathbf{v}$ serves as a feature vector for a task-specific classifier to train off of. One can think of the inference from the micromodels to be a featurization step to convert the input data (x, y) into a feature vector $(\mathbf{v}, y)$, while the task-specific classifier is an independent classification model that actually learns a task from such featurized values.

Given $(\mathbf{v}, y)$, a task-specific classifier can be trained. Similar to micromodels, the framework allows for any algorithm of choice to be used for task-specific classification, from simple regression models to complex neural networks. The original authors use explainable boosting machines (EBM) (Caruana et al., 2015) because of the explanations it provides. More specifically, the use of EBMs allows one to understand the impact that each micromodel had on the task-specific classifier's decision making process. For more details on EBMs, we point our readers to both the original paper

as well as its widely used open-source repository (https://github.com/interpretml/interpret).

C Detailed Explanation of EPITOME$_{\text{Model}}$

Epitome$_{\text{Model}}$ is a multi-task bi-encoder model in which the two encoders are initialized with the weights of RoBERTa and pre-trained with in-domain data that was available to the authors of EPITOME. The two encoders then each encode the seeker's post S_i and the response post R_i.

$$e_i^{(S)} = \text{S-Encoder}(S_i); e_i^{(R)} = \text{R-Encoder}(R_i) \quad (4)$$

Borrowing terminology from transformers, the response post encoding is used as a query and the seeker post is used as keys and values.

$$a_i(e_i^{(R)}, e_i^{(S)}) = \text{softmax}(e_i^{(R)} e_i^{(S)} / \sqrt{d}) e_i^{(S)} \quad (5)$$

The encoded response $e_i^{(R)}$ is summed with the output of the attention layer $a_i(e_i^{(R)}$ to obtain a residual mapping, resulting in a seeker-context aware representation of the response post, which is used to jointly train on the two tasks.