Emotion Modeling for Improving Empathy in Conversations

Travis Metz
University of California, Berkeley
tmetz@berkeley.edu

Evelyn You
University of California, Berkeley
evelyn.you@berkeley.edu

August 2, 2020

Abstract

Empathy is an important element of interpersonal communications, with broad applications in multiple domains. Recent advances in language modeling have improved the ability of dialogue systems to generate appropriate responses in a variety of settings, though it has been challenging to develop digital agents that exhibit empathetic behavior, due to the lack of publicly available datasets for identifying and evaluating empathy in conversations. In this paper, we combine existing dialogue corpus with emotion annotation frameworks to gain insights into patterns of emotion interaction and empathy in conversations. Dialogue corpus selected using our findings shows improved human self-evaluation scores on empathy and higher automatic evaluation measures adjusted for training size. These findings help to lay the groundwork for identifying and building improved training datasets for empathy, toward the ultimate goal of improving empathetic behavior in human-machine interactions.

1 Introduction

Empathy describes the ability to understand and react to emotions of others. Our interpersonal communications and relationships are supported by empathy, and society relies on empathy to connect diverse groups of people.[1] Higher degrees of empathy is associated with better outcomes in multiple domains involving human-human communication including customer service, conflict resolution, mental health therapy, and relationship counseling.[11, 1, 4] As digital communication forms including text and online messaging continue to gain popularity, many fields will benefit from digital agents capable of identifying and generating empathetic behavior.

Historically, training digital agents on empathy has faced significant challenges stemming from the lack of publicly available datasets for reliably identify and evaluate degree of empathy in conversations.[11, 1, 4] While recent language models trained on vast amounts of text scraped from sources such as Wikipedia and online social platforms have shown remarkable advances in performance across a variety of tasks, language generated by these models are unlikely to exhibit empathetic behavior in dialogue settings, and may in fact be prone to reproduce insensitive and aggressive comments typical of certain online forums.[11]

Our interest in the topic is driven by the ultimate goal of improving empathy in human-machine interactions. Specifically, we are interested in understanding whether the presence, or absence, of certain emotion interaction patterns between conversing partners can help determine the degree of empathy present in a conversation. Though much future work is needed, our work in this paper suggests that empathy levels in conversations can be inferred from the presence or absence of such patterns. Thus, future dialog corpus used to train and evaluate language models on empathy could be automatically selected and built from a variety of sources based on emotion interaction patterns, with the assistance of an emotion classifier model. Going forward, this would significantly reduce the time and human resources needed to curate dialog datasets such as the one created in Rashkin et al.[11]

2 Related Work

Empathy Identification: Recent work on empathy in affective computing has been focused on analyzing empathetic behavior in specific domains, including call center conversations[1], counseling therapy sessions[9], and online health community message boards[4]. Often, these frameworks rely on a combination of linguistic, acoustic, and other features[1, 9], rather than text-only data. Due to the domain-specific data used, it is also unclear whether performance could transfer to more general domains. Notably, Khanpour et al.[4] present a dataset of messages from online health communities with sentence-level binary empathy annotations. Their CNN-LSTM model achieved an F-1 score of 0.78. However, the corpus and model are not publicly available.

Emotion Annotation: Datasets for emotion recognition was first introduced in Affective Text[12], where news headlines were classfied into six emotion categories using the Ekman taxonomy (i.e., anger, disgust, fear, joy, sadness, and surprise) based on biological responses[3]. Additional datasets for emotion annotation have been introduced over the years, typically manually labelled and relatively small in size. Datasets used include tweets[6], movie subtitles[7], dialogues[5, 10], and Reddit comments[2]. Of these, Demszky et al.'s GoEmotions (GE) corpus based on Reddit comments is one of the largest with fine-grained emotion labels for 58k comments in 27 emotion categories and contains schema mapping to sentiment and Ekman taxonomy, which enables transfer learning to other datasets.

Dialogue Corpus: Several existing dialogue datasets contain utterance-level emotion annotation. Two of the largest are DailyDialog (DD), which is composed of 13k dialogues and 103k utterances obtained from crawling educational websites entended for ESL learners[5], and MELD, which contains 1.4k dialogues and 13k utterances from the Friends TV series.[10] Both datasets are based on the Ekman taxonomy and neither have any annotation related to empathy. The EmpatheticDialogues (ED) dataset introduced by Rashkin et al.[11] was constructed by hiring 810 US workers through Amazon Mechanical Turk where participants are prompted with one of 32 emotion contexts (e.g., surprised, excited, angry, proud) for each conversation, and instructed to converse in an empathetic manner. Participant provide self-evaluation (scale 1-5) of each conversation on dimensions of empathy, relevance, and fluency. However, utterance-level emotion annotations are not available.

3 Methods

For this paper, we are interested in determining whether the presence, or absence, of certain emotion interaction patterns between conversing partners could help determine the degree of empathy present in a conversation. To do this, we need to annotate each utterance (i.e., each turn by a speaker in a conversation) in the ED corpus with an emotion label, analyze the interaction patterns between speakers, and revise the ED corpus based on any patterns we find that could signal the presence or absence of empathy. If the revised corpus shows improved ratings on human and automatic measures for empathy versus the original corpus, it would suggest viability and applicability for our methodology of using emotion interaction pattern as a data-selection tool to build dialogue datasets used to train digital agents on empathy.

Our work starts with 1) data-cleaning and replication of BERT fine-tuned measures on retrieval as reported by Rashkin et al.[11] on ED (1.1) and by Demszky et al[2] on GE (1.2), 2) performing emotion annotation for each utterance of the ED dialogue corpus from Rashkin et al. using the GE emotion classification framework provided by Demszky et al, additional annotations were predicted for the DD and MELD dialogue databases to verify transfer learning, 3) conducting qualitative analysis of patterns in emotion interaction between conversing partners based on the utterance-level emotion annotations, revising the ED corpus based on our findings in interaction patterns, and 4) testing whether the revised corpus achieves improved measures on human self-evaluation of empathy (4.1), and on automatic measures when fine-tuned for BERT retrieval (4.2).

3.1 Data Preparation and BERT Fine-tuning Replication

3.1.1 EmpatheticDialogues (ED)[11]

Rashkin et al. used a variety of transformer-type models[13] [14], but achieved the best results by tuning BERT models. The base BERT model is pre-trained on a corpus of 1.7 billion Reddit conversations[1], before fine-tuning on the ED dataset. The ED dataset for fine-tuning consists of approximately 25k conversations averaging just under four utterances per conversation. We replicated this process using the pre-trained BERT model and fine-tuned with the training dataset per Rashkin et al. for our Baseline models.

The model architecture is shown in Figure 1. A BERT-based model is to used to encode both the original utterance and the candidate responses (separately). The model chooses the most likely candidate utterance based on a softmax of the dot product of the two encodings. During training the candidate responses are all the other members of the batch (32 in our configuration).[2] The candidates at inference were drawn from the entire EmpatheticDialogues dataset.[3]

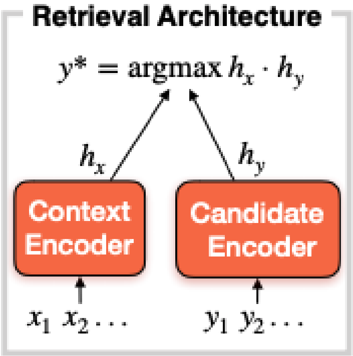

Figure 1: BERT Retrieval Architecture (figure from Rashkin et al.[11])

While replicating the fine-tuning process, we noticed various formatting issues in the dataset that likely prevented some conversations from being used during the fine-tuning process. We fixed these formatting issues to produce a "clean" version of our Baseline model. Automatic measures on BERT retrieval from the original and clean versions of the dataset can be found in Table 3.

3.1.2 GoEmotions (GE)[2]

Demszky et al. utilize a cased base BERT model with fine-tuning to classify the existence of emotions in utterances. While the model is not publicly available, the GE corpus with which to replicate the fine-tuning exercise is provided. The dataset used to fine-tune the base BERT model consists of 58k Reddit comments with manual emotion annotations from 27 emotion categories. We replicated the fine-tuning process and results per Demszky et al., producing a multi-label classifier that takes an utterance and provides probabilities that the utterance represents each of 27 emotions or neutral (no emotion content)[4].

[1] Pre-training based on the Hugging Face PyTorch implementation of BERT at https://github.com/huggingface/pytorchtransformers.

[2] Rashkin et al. trained with a batch size of 256, which likely improved their training efficiency given this constraint. We were limited by using a single GPU (T4) with 16GB of memory, thus the reduced batch size to avoid out-of-memory errors. Our training process takes approximately three hours each time on one NVIDIA T4 GPU with 16GB of memory.

[3] It is worth noting that we relied heavily on the codebase generously made available by Rashkin et al. for purposes of fine-tuning and inference. This has provided an invaluable starting point for our work.

[4] We achieved overall F1-score of 0.50 across the different emotion categories when tested on the test set, consistent with results reported in Demszky et al.

To prepare for utterance-level emotion annotation on the ED, DD, and MELD dialogue datasets, we also reformatted these datasets into a format consistent with data used to train the GE classifier. For example, reformatting involved replacing all English names within the dialogue datasets with the token "[NAME]".

3.2 Dialog Utterance-Level Emotion Annotation

Having replicated the GE emotion annotation classifier and results per Demszky et al.[2], we proceeded to annotate the ED, DD, and dialog corpus on the utterance-level. Each utterance is annotated with a predicted label in one of 27 emotion categories used by Demszky et al. or neutral if lacking emotion content[5], as well as mapped into the six categories of Ekman taxonomy[3] or neutral if lacking emotion content[6]. True label and predicted label were compared where available (for DD and MELD) to analyze discrepancies and validity of the classifier predictions.

3.3 ED Corpus Revision Based on Emotion Interaction Patterns

We conducted qualitative analysis of emotion labels on pairs of consecutive utterances (i.e., one utterance per each conversing partner) in the dialogues for emotion interaction patterns that could indicate the presence or absence of empathy. After analyzing the interaction patterns, we proceeded with the following ED corpus revisions for empathy testing. We are most interested in test results on ED-GE Select List 1 and ED-GE Select List 2, with the other corpus revisions serving as controls for our tests:

Corpus	Dialogue Count	% of ED-Base	Composition
ED-GE Select List 1	17,038	87%	ED-Base conversations, excluding conversations in ED-GE Exclude List 1.
ED-GE Exclude List 1	2,495	13%	Conversations where 1) the emotion "surprise" based on Ekman taxonomy manifests in consecutive utterances, and 2) the emotion "surprise" based on Ekman taxonomy ends the conversation.
ED - Control 1	17,172	87%	Random subset of ED-Base conversations with the same number of total utterances as ED-GE Select List 1
ED-GE Select List 2	11,132	57%	ED-Base conversations, excluding conversations in ED-GE Exclude List 2.
ED-GE Exclude List 2	8,401	43%	All conversations in ED-GE Exclude List 1, and conversations where the lack of emotion content (i.e., "neutral") based on GE taxonomy manifests in consecutive utterances.
ED-Control 2	11,123	57%	Random subset of ED - Base conversations with the same number of total utterances as ED-GE Select List 2
ED-GE Select List 2 Size Adjusted	8,563	43%	Random subset of ED-GE Select 2 conversations with the same number of total utterances as ED-GE Exclude List 2

Table 1: Revised ED corpus based on observed emotion interaction patterns. Note that % of ED - Base values are calculated based on the number of utterances contained in the corpus.

[5] Total of 28 emotion categories per GE taxonomy: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness optimism, pride, realization, relief, remorse, sadness, surprise, and neutral,

[6] Total of seven categories per Ekman taxonomy: neutral, anger, disgust, fear, joy, sadness, and surprise.

3.4 Empathy Rating of Revised ED Corpus

3.4.1 Human Self-Evaluation Measures

Participants of ED provide ratings on each of their conversations with another partner on three dimensions: empathy, relevance, and fluency on a scale of 1-5 (1: not at all, 5: very much). Definitions for the three dimensions are provided below, per Rashkin et al.[11]:

- **Empathy:** did the response show understanding of the feelings of the person talking about their experience?
- **Relevance:** did the response seem appropriate to the conversation? Were they on-topic?
- **Fluency:** could you understand the responses? Did the language seem accurate?

We will evaluate the revised ED corpus on all three dimensions, but will focus on average rating of the corpus on the empathy dimension. As both parties to a conversation was required to provide ratings, a corpus containing more empathetic conversations should lead to increased self-evaluation scoring on empathy on average.

3.4.2 Automatic Measures

We focus on the retrieval task (i.e., retrieving the correct response to an utterance) as the main automatic measures for fine-tuning efficacy on empathy. We evaluate the effectiveness models fine-tuned with revised training corpus using the metrics below:

- **BLEU:** compares model response to the actual response and measures correspondence[8]
- **P@1,100:** precision accuracy of the model at choosing the correct response out of a hundred randomly selected examples in the test set

As participants for ED were instructed to converse empathetically, higher ratings on the retrieval task generally suggests increased empathetic behavior of the model, after controlling for the size of training data.

4 Results and Analysis

4.1 Emotion Interaction Patterns in Dialogues

4.2 Results for Human Self-Evaluation Measures

Statistics for human self-evaluation scores by ED participants on different ED datasets revised using our emotion interaction patterns are presented in Table 2. While human self-evaluation scores are subjective, the revised datasets report higher self-evaluation scores on all dimensions, including empathy, vs. ED-Base dataset and datasets from the corpus using our data selection framework.

Corpus	Dialogue Count	Empathy	Relevance	Fluency
ED-Base	19,533	4.709 ± 0.675	4.796 ± 0.570	4.837 ± 0.514
ED-GE Select List 1	17,038	4.717 ± 0.662	4.805 ± 0.552	4.844 ± 0.501
ED-GE Exclude List 1	2,495	4.648 ± 0.748	4.729 ± 0.674	4.794 ± 0.593
ED-GE Select List 2	11,132	$\mathbf{4.733 \pm 0.635}$	$\mathbf{4.815 \pm 0.535}$	$\mathbf{4.853 \pm 0.484}$
ED - GE Exclude List 2	8,401	4.677 ± 0.723	4.770 ± 0.612	4.816 ± 0.550

Table 2: Mean and standard deviation of human self-evaluation scores by ED participants on revised ED datasets. Conversations selected based on emotion interaction patterns in utterance-level emotion annotations generated by GE show higher average self-evaluation scores with lower standard deviations on empathy, relevance and fluency, when compared with the Base ED corpus and conversations excluded based on the emotion interaction patterns. Bold: best result for the evaluation category.

4.3 Results for Automatic Measures

An overview of automatic evaluation results on the response retrieval task using BERT models fine-tuned on the different revisions of the ED corpus can be found below in Table 3. As noted in Rashkin et al., fine-tuning the BERT model pretrained on a large Reddit corpus with the ED dataset led to a improvements in both BLEU and P@1,100 scores (6.21 vs. 5.97 and 65.92% vs. 49.94%, respectively), with the base comparison being to the BERT model that had been pre-trained on Reddit but not fine-tuned on the ED dataset.

Due to limited computing resources, we were forced to use a lower batch size (32 vs. 256 used by Rashkin et al.) when performing BERT fine-tuning, and we were not able to precisely replicate the original results of Rashkin et al. The meaningfully lower batch sizes used for training likely impacted training results on all models, and we use scores of ED-Base - Replicated (6.10 for BLEU and 63.87% for P1,100) as our new Baseline for comparing model performances.[11]

Model	Training Corpus for Fine-Tuning	% of ED-Base	Avg. BLEU	P@1,100
0a	Pretrained BERT - As Reported	100%	5.97	49.94%
0b	ED-Base - As Reported	100%	6.21	**65.92**%
1	ED-Base - Replicated (Baseline)	100%	6.10	63.87%
2	ED-Base - Cleaned	100%	**6.25**	63.75%
3	ED-GE Select List 1	87%	**6.22**	**63.37**%
4	ED-GE Control 1	87%	6.15	63.27%
5	ED-GE Select List 2	57%	5.99	**62.17**%
6	ED-GE Control 2	57%	**6.13**	61.52%
7	ED-GE Select List 2 Size Adjusted	43%	**6.10**	**61.73**%
8	ED-GE Exclude List 2	43%	6.05	60.79%

Table 3: Summary of model outcomes for response retrieval. AVG BLEU: average of BLEU-1,-2,-3,-4. P@1,100: precision retrieving the correct test candidate out of 100 test candidates. The two "As Reported" models include results from Rashkin et al.[11], while eight new models are evaluated for BERT retrieval. These models differ only in the corpus used to fine-tune a base BERT model. ED-Base - Replicated represents our attempt to replicate the Rashkin et al. results using unmodified ED corpus; ED-Base - Cleaned uses the ED corpus cleaned for formatting issues. Composition of the other datasets used for fine tuning are explained in Table 1. Bold: best result for the evaluation category controlled for training data size.

Interestingly, a notable substantial improvement to the Base model from our various dataset manipulation was in simply cleaning up the original Rashkin et al. finetuning dataset (as described above). Fixing its formatting and thus adding more training data improved the BLEU score from 6.10 to 6.25 (though did not improve the p@1,100 score).

Models fine-tuned using our revised ED datasets (i.e., Models 3 and 5) achieved higher scores vs. the ED-Base - Replicated Baseline, but lower scores vs. ED-Base Cleaned. However, the lower scores vs. ED-Base Cleaned is likely due to the reduced training data size available in the revised ED datasets. When controlled for training data size, models based on our revised ED datasets generally achieved higher scores vs. control datasets. For example, Model 7 is trained using a random subset of ED-GE Select List 2 (conversations selected into the revised corpus using our emotion interaction findings) to adjust the size of the dataset so that it is the same as ED-GE Exclude List 2 (conversations excluded from the revised corpus) used in Model 8 - both contain 43% of the original ED corpus based on number of utterances. As expected, Model 7 achieved higher Avg. BLEU and P@1,100 scores vs. Model 8. The only exception to the results is the lower BLEU score of Model 5 when compared to the control - Model 6, we could not determine the cause of this lower BLEU score and it is an anomaly requiring furthur investigation.

5 Conclusion and Future Work

One obvious avenue of further research would be to apply further scoring measures on the different models we developed above. The automated measures employed to judge the effectiveness of the model (BLEU and p@1,100) seem blunt instruments for assessing progress. Rashkin et al. ultimately used human reviewers to rate the empathetic nature of the dialogues in their test sets. We did not have the resources to do this with our retrieved responses (to compare to a baseline of the pretrained BERT model), but it seems possible that this review technique might generate more nuanced evaluations of the different models.

Our work does suggest that you can use emotion patterns in dialogue to identify more empathetic language. This should allow further work to filter larger dialogue datasets and add them to a dataset like ED, thus allowing a more effective finetuning of BERT (or presumably any other language model). Given the increasing use and importance of dialogue models, it bears further research to improve their ability to generate language that is deemed empathetic. Finding an automated way to select empathetic dialogue to grow fine-tuning datasets is worthy of future effort.

A Example Dialogues from the EmpatheticDialogues (ED) Dataset

My coworker is allowed to work remotely, but I am not...
I work remotely, I wish that you could do something like that as well.
I do too,it is unfortnuate because unbiased, I am our best performer, so it is curious
Sometimes in life, it is not about performance. Some people just get a shiner spoon.

People get rejected all the time but it shouldn't be an excuse to let it dissuade you from reaching your goals.
What happened?
Got rejected from a place I wanted to work, not once but three times
I am sorry to hear that. I hope you find a better opportunity. Did you know why they rejected you?
Thanks, they gave me the same generic lines you always hear.

I never thought things would go like this.
What happened for you to feel that way?
I was at the beach and a hurricane changed track and was coming straight at us. My boyfriend didnt want to leave because we were on the second day only of our vacation.
Oh that does sound terrible. What did you guys decide to do?
Well we ended up leaving I was too scared to stay and was really upset.

What a difference a year makes. Last year one evening my family was at home when a tree fell on the house and broke through the ceiling.
That's very scary. I hope no one got hurt.
We were OK, though the tree broke through only a few feet away from my daughter.
So happy everyone was fine!! Everything else can be fixed.
Indeed. We were out of the house for five months while repairs were being done, but now the house is better than ever.
So good to hear. Might want to trim some trees lol

References

[1] Firoj Alam, Morena Danieli, and Giuseppe Riccardi. "Annotating and modeling empathy in spoken conversations". In: *Computer Speech Language* 50 (July 2018), pp. 40–61. ISSN: 0885-2308. DOI: 10.1016/j.csl.2017.12.003.

[2] Dorottya Demszky et al. "GoEmotions: A Dataset of Fine-Grained Emotions". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020). DOI: 10.18653/v1/2020.acl-main.372.

[3] Paul Ekman et al. "Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion". In: *Journal of personality and social psychology* 53 (Nov. 1987), pp. 712–7. DOI: 10.1037/0022-3514.53.4.712.

[4] Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. "Identifying Empathetic Messages in Online Health Communities". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Asian Federation of Natural Language Processing, Nov. 2017, pp. 246–251. URL: https://www.aclweb.org/anthology/I17-2042.

[5] Yanran Li et al. *DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset*. 2017. arXiv: 1710.03957 [cs.CL].

[6] Saif Mohammad and Felipe Bravo-Marquez. "Emotion Intensities in Tweets". In: Jan. 2017, pp. 65–77. DOI: 10.18653/v1/S17-1007.

[7] Emily Ohman et al. "Creating a Dataset for Multilingual Fine-grained Emotion-detection Using Gamification-based Annotation". In: Jan. 2018, pp. 24–30. DOI: 10.18653/v1/W18-6205.

[8] Kishore Papineni et al. "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135.

[9] Verónica Pérez-Rosas et al. "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)*. Association for Computational Linguistics, July 2017, pp. 1426–1435. DOI: 10.18653/v1/P17-1131.

[10] Soujanya Poria et al. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019). DOI: 10.18653/v1/p19-1050.

[11] Hannah Rashkin et al. *Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset*. 2019. arXiv: 1811.00207v5 [cs.CL].

[12] Carlo Strapparava and Rada Mihalcea. "SemEval-2007 Task 14: Affective Text". In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, June 2007, pp. 70–74. URL: https://www.aclweb.org/anthology/S07-1013.

[13] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[14] Yinfei Yang et al. "Learning Semantic Textual Similarity from Conversations". In: *Proceedings of The Third Workshop on Representation Learning for NLP*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 164–174. DOI: 10.18653/v1/W18-3022. URL: https://www.aclweb.org/anthology/W18-3022.