

# 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 1, 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 BERT-based 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.[12, 1, 5] 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.[12, 1, 5] 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.[12]

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.[12]

## 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[10], and online health community message boards[5]. Often, these frameworks rely on a combination of linguistic, acoustic, and other features[1, 10], 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.[5] 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[13], where news headlines were classified into six emotion categories using the Ekman taxonomy (i.e., anger, disgust, fear, joy, sadness, and surprise) based on biological responses[4]. Additional datasets for emotion annotation have been introduced over the years, typically manually labelled and relatively small in size. Datasets used include tweets[7], movie subtitles[8], dialogues[6, 11], and Reddit comments[2]. Of these, Demszky et al.’s corpus based on Reddit comments is one of the largest with fine-grained emotion labels for 27 emotion categories and contains schema mapping frameworks to sentiment and Ekman taxonomy that enables transfer learning with other datasets.

**Dialogue Corpus:** Several existing dialogue datasets contain emotion annotation. These include notated with emotion include [6] and [11]

Rashkin et al.[12] set out to explore whether the artificial creation of a dataset of empathetic dialogues would allow the large language models to be refined such that when used in a dialogue context they would generate or choose more empathetic responses.

[TO BE DELETED: The EmpatheticDialogues dataset was intentionally constructed to be empathetic as the participants in its generation were instructed to be empathetic. But, it would stand to reason that the dialogues exist on a continuum of empathy. We wondered if it would be possible to use another framework to assess which of these dialogues were more likely to be empathetic and then refine the dataset for purposes of fine-tuning a pretrained BERT[3] model. We determined to use the work of Demszky et al.[2] and their GoEmotions framework to understand the emotion patterns in the EmpatheticDialogues dataset and use the more common patterns as a proxy for empathy, and eliminate those emotional patterns less likely to occur.]

## 3 Methods

Our work starts with Rashkin et al.’s dataset (EmpatheticDialogues[12]), cleaning the dataset and annotating each utterance with emotion annotations using the emotion classification framework (GoEmotions[2]) provided by Demszky et al. Next, we conduct qualitative analysis of patterns in emotion interaction between conversing partners based on emotion annotations, examining EmpatheticDialogues as well as other publicly available dialogue datasets. Finally, we revise the EmpatheticDialogue based on our findings and tests whether the revised dataset achieves improved measures on a) human self-evaluation of empathy for the conversations, and b) fine-tuning for a BERT retrieval architecture.

### 3.1 BERT Fine Tuning and Retrieval

Rashkin et al.[12] focused their work on Transformer networks[14], but experimented with BERT models. In their results, the fine-tuned BERT based models showed better performance, as such, we chose to work with this framework. They used the base BERT model and pre-trained it on a corpus of 1.7 billion Reddit conversations, before fine-tuning on their EmpatheticDialogues dataset. We began with their pre-trained model and then fine-tuned with revised datasets of dialogue that we developed with the intention of them being ‘more’ empathetic.

We focus on the retrieval task as the key objective of the modeling exercise, which is to say retrieving the correct response to an utterance. We evaluate the effectiveness of the models using two different metrics.

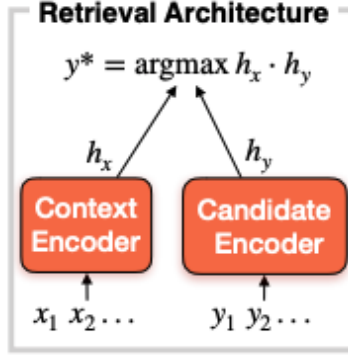


Figure 1: BERT Retrieval Architecture (figure taken from[12])

- P@1,100 - the accuracy of the model at choosing the correct response out of a hundred randomly selected examples in the test set
- BLEU - compares model response to the actual response and measures correspondence[9]

The model architecture is shown in Figure 1. A BERT-based model is 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 (which was 32 in our configuration).<sup>1</sup> The candidates at inference were drawn from the entire EmpatheticDialogues dataset.<sup>2</sup>

### 3.1.1 Replication

The fine-tuning dataset consists of 24,850 conversations with an average of just under four utterances per conversation. We fine-tuned base BERT on the original dataset in order to see if we could replicate the results of the original Rashkin et al. paper. The process of fine-tuning took approximately three hours for each of the datasets we worked with on one NVIDIA T4 GPU with 16GB of memory.

### 3.1.2 GoEmotions1

Using the methodology described in 3.2.1 and 3.2.2 below, we fine-tuned on a subset of the EmpatheticDialogues dataset that represented approximately 87% of the original training set.

### 3.1.3 GoEmotions2

Similarly, we further reduced the training set such that it represented approximately 57% of the original training set.

### 3.1.4 Cleaned Original ED Dataset

In doing our exploratory work we noted that the dataset had various formatting problems that seemed likely to cause them to be not available during training. We modified the original training dataset such that these examples could be utilized in training. [EY to modify]

<sup>1</sup>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.

<sup>2</sup>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. It provided an invaluable starting point.

## 3.2 GoEmotions

### 3.2.1 Generation of Emotion Labels

Demszky et al. [2] utilize a cased base BERT model with fine-tuning to classify the existence of emotions in utterances. While they do not make their model available, they do provide the resources (datasets, etc) with which to repeat their fine-tuning exercise, which we did. This produces a multi-label classifier that takes an utterance and provides probabilities that the utterance represents each of 27 emotions that the methodology utilized.

Having recreated their fine-tuned model, we then took the utterances from the training set of EmpatheticDialogues and produced emotion annotation for each utterance. These emotion tags were then studied in an effort to find patterns that demonstrated empathy, as described further below.

### 3.2.2 Use of Emotions to Segment Dataset

[EY to describe methodology]

## 4 Results and Analysis

Table 1 provides a comparison of the BLEU and p@1,100 scores on the response retrieval task using models fine-tuned on the different datasets described above.

As noted in Rashkin et al., fine-tuning the BERT model (that had been pretrained on a large Reddit corpus) with the EmpatheticDialogues dataset led to improvements in both BLEU and p@1,100 scores (6.21 vs. 5.97 and 65.92 vs. 49.94) (with the base comparison being to the BERT model that had been pre-trained on Reddit but not fine-tuned on the EmpatheticDialogues dataset).

We were not able to precisely replicate the original results of Rashkin et al. using their unmodified dataset for fine-tuning, with lower BLEU (6.10 vs. 6.21) and p@1,100 scores (63.87 vs. 65.92). As described above, we were compelled by compute resources to use meaningfully lower batch sizes (32 vs 256) which seems likely to impact training given the use of the batch for candidate responses to the retrieval problem.

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).

Our two attempts to improve the dataset using emotion pattern matching did not result in improvements. The GoEmotions1 dataset (represent 87% of the original dataset) did improve from the Base - replicated results, but the improvement is due to the cleaning of the dataset above (as we used the clean dataset as a starting point before reducing its scale based on emotion matching). Note its slightly poorer performance when compared with Base - cleaned.

The GoEmotions2 dataset (representing 57% of the original dataset) resulted in lower performance relative to the base models, likely due the substantial reduction in the size of the fine-tuning set reduced its efficacy (even if it was potentially 'more' empathetic).

Fine-Tuning	Avg. BLEU	P@1,100
Pretrained BERT - as reported	5.97	49.94
Base - as reported	6.21	65.92
Base - replicated	6.10	63.87
Base - cleaned	6.25	63.75
GoEmotions1	6.22	63.37
GoEmotions2	5.99	62.17

Table 1: Model outcomes. Reports of two models are reported from [12], while four new models are evaluated. They differ only in the dataset used to fine-tune a base BERT model. AVG BLEU: average of BLEU-1,-2,-3,-4. P@1,100: precision retrieving the correct test candidate out of 100 test candidates.

BERT represents the reported results for a base BERT model pretrained on 1.7B Reddit conversations

Base-as reported

Base-replicated = ED dataset unchanged. Authors attempted to replicate model

Base-cleaned = ED dataset cleaned for formatting issues

GoEmotion = ED dataset reduced to only those emotion pairs most common in empathetic conversations.

Model 1 represented 87% of original training set, while Model 2 represented 57%

Fine-Tuning	Avg. BLEU	P@1,100
Subset (43%) of ED with emotion patterns	6.10	61.73
Subset (43%) of ED without emotion patterns	6.05	60.79

Table 2: Model outcomes adjusted for training size.

The pretrained BERT model was fine-tuned with smaller equally sized datasets in order to remove fine-tuning dataset size effects. Both datasets represented 43% of the original training set, with the first having the emotion patterns that we believe are the strongest signals of empathetic conversation, and the second without. BLEU and P@1,100 scores are as per Table 1.

## 5 Conclusion and Future Work

Our attempt to generate a more empathetic dialogue dataset to use in a fine-tuning exercise by using emotion patterns did not bear fruit. We did, though, note some improvement available from cleaning the original dataset.

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 logical extension of our work would be to use the emotion patterns we noted in the created EmpatheticDialogues dataset to select further fine-tuning data from other dialogue sets available to researchers. Said differently, it might be the case that our reduction of the size of the dataset used to fine-tune the model reduced its effectiveness as fine-tuning towards empathy (even if the dataset better represented empathetic dialogue).

Separately, 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.

## A Example Dialogues from EmpatheticDialogues 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 unfortunate because unbiased, I am our best performer, so it is curious  
Sometimes in life, it is not about performance. Some people just get a shinier 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 didn't 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](https://doi.org/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](https://doi.org/10.18653/v1/2020.acl-main.372).
- [3] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North* (2019). DOI: [10.18653/v1/n19-1423](https://doi.org/10.18653/v1/n19-1423).
- [4] 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](https://doi.org/10.1037/0022-3514.53.4.712).
- [5] 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>.
- [6] Yanran Li et al. *DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset*. 2017. arXiv: [1710.03957](https://arxiv.org/abs/1710.03957) [cs.CL].
- [7] Saif Mohammad and Felipe Bravo-Marquez. “Emotion Intensities in Tweets”. In: Jan. 2017, pp. 65–77. DOI: [10.18653/v1/S17-1007](https://doi.org/10.18653/v1/S17-1007).
- [8] 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](https://doi.org/10.18653/v1/W18-6205).
- [9] 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](https://doi.org/10.3115/1073083.1073135).
- [10] 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](https://doi.org/10.18653/v1/P17-1131).
- [11] 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](https://doi.org/10.18653/v1/p19-1050).
- [12] Hannah Rashkin et al. *Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset*. 2019. arXiv: [1811.00207v5](https://arxiv.org/abs/1811.00207v5) [cs.CL].
- [13] 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>.
- [14] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: [1706.03762](https://arxiv.org/abs/1706.03762) [cs.CL].