Towards Low-Resource Real-Time Assessment of Empathy in Counselling

Zixiu Wu
Philips Research & University of Cagliari
zixiu.wu@philips.com

Rim Helaoui
Philips Research
rim.helaoui@philips.com

Diego Reforgiato Recupero and **Daniele Riboni**
University of Cagliari
{diego.reforgiato, riboni}@unica.it

Abstract

Gauging therapist empathy in counselling is an important component of understanding counselling quality. While session-level empathy assessment based on machine learning has been investigated extensively, it relies on relatively large amounts of well-annotated dialogue data, and real-time evaluation has been overlooked in the past. In this paper, we focus on the task of low-resource utterance-level binary empathy assessment. We train deep learning models on heuristically constructed empathy vs. non-empathy contrast in general conversations, and apply the models directly to therapeutic dialogues, assuming correlation between empathy manifested in those two domains. We show that such training yields poor performance in general, probe its causes, and examine the actual effect of learning from empathy contrast in general conversation.

1 Introduction

As a pillar of psychotherapy, empathy is crucial to effective counselling, owing to its importance in building counsellor[1]-client rapport (Elliott et al., 2011) that can enable more effective interventions and better outcomes (McCambridge et al., 2011; Gaume et al., 2009). In particular, "listening with empathy" is considered a guiding principle (Rollnick et al., 2008) for motivational interviewing (Miller and Rollnick, 2012) (MI), a psychotherapeutic approach widely adopted to elicit positive behaviour change by evoking motivation from clients. Gauging counsellor-side empathy is, therefore, essential to assessing MI integrity (Moyers et al., 2016).

Empathy assessment for MI has conventionally been conducted manually by trained annotators, which requires extensive annotator training and transcript review. Since such a time-consuming and costly setup is difficult to scale up, recent years have seen attempts of automating the process with machine learning, including transcript-based (Xiao et al., 2012; Gibson et al., 2015, 2016), speech-based (Xiao et al., 2014, 2015), and multimodal (Xiao et al., 2016b) methods. Those works are, however, limited in that 1) therapist empathy is only assessed at session-level rather than utterance-level; 2) classical machine learning with heuristic feature engineering is used, while recent deep-learning frameworks have not been utilised for this purpose; 3) the machine-learning-based approaches all assume access to privately-owned sizeable corpora of therapeutic dialogues with empathy annotation at session level, but in reality such well-annotated data are often very limited, even more so at utterance level; and 4) the link between empathy manifested in general conversation and in MI counselling remains unexplored.

In this work, we make the first attempt (to the best of our knowledge) at addressing those limitations while probing the correlation between empathy manifestations in different domains. Specifically, we employ pre-trained language models such as BERT (Devlin et al., 2019) for text-based binary classification of utterance-level therapist empathy, optionally taking the conversation context as input. We consider any counsellor utterance to be empathetic if it shows empathy, and non-empathetic if it does not (ranging from neutral to apathetic). Our models have no access to counselling conversations during their training and validation, as we experiment with learning from contrast of empathy vs. non-empathy in out-of-domain (OOD) training data. To that end, we leverage publicly available datasets of general conversations with heuristic empathy labels (Rashkin et al., 2019; Zhong et al., 2020) for OOD training, investigating the connections between general-conversational empathy and therapeutic empathy, as illustrated in Figure 1.

To benchmark the models, we manually anno-

[1]We use "counsellor" and "therapist" interchangeably in this work.

Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology, pages 204–216

Figure 1: Training a binary empathy classifier on heuristically constructed empathetic vs. non-empathetic utterances in general conversations (i.e. out-of-domain w.r.t. MI), and then testing it on MI conversations. In this case, the empathy contrast for training is *r/OffMyChest* vs. *r/CasualConversation*. The classifier can take only the listener/therapist utterance (**bold**) as input or additionally use the preceding speaker/client utterance (*italic*).

tated utterance-level empathy for a subset of transcribed high- vs. low-quality counselling demonstrations (Pérez-Rosas et al., 2019) that are publicly available. We also build unsupervised baselines for the task by **a)** formulating binary empathy classification as natural language inference (NLI), as proposed by Yin et al. (2019), and **b)** tackling the surrogate task of client-counsellor agreement via NLI, under the assumption that an empathetic reply from the counsellor tends to show accordance with the client utterance in the preceding turn.

Our experiments show that models trained on OOD empathy contrast are not sufficiently accurate predictors of MI empathy/non-empathy, even though the benefit of such training can be observed when compared to training on OOD data without empathy contrast. Upon probing, we argue that more fine-grained (e.g. sentence-level) empathy annotation and prediction could yield better results.

2 Related Work

2.1 Machine-Learning-Based Approaches to Empathy Analysis for MI

Prior work has approached assessment of empathy in MI delivery via speech and linguistic features.

Among text-based methods, Xiao et al. (2012) proposed one of the earliest approaches for utterance-level empathy classification using an n-gram language model. Psycholinguistic norm features are used in addition to other linguistic features in the work of (Gibson et al., 2015). More recently, Gibson et al. (2016) utilised long short-term memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) to generate turn-level behavioural acts that are further processed by a deep neural network to predict session-level empathy.

Speech features have also been examined. Xiao et al. (2014) investigated features such as jitter and shimmer from speech signals, Xiao et al. (2015) studied speech rate entrainment, while Pérez-Rosas et al. (2017) used an array of acoustic and linguistic features to train their multimodal models.

There are also a number of recent studies on data-driven MI behaviour coding based on text (Cao et al., 2019; Tanana et al., 2016; Xiao et al., 2016a; Gibson et al., 2018), speech (Singla et al., 2020), and both (Chen et al., 2019; Flemotomos et al., 2021) , but they are less relevant to this work due to their lack of explicit empathy modelling.

Different from the research listed above, this work addresses utterance-level empathy classification instead of session-level assessment, similar to Wu et al. (2020) which proposes utterance-level prediction of whether the therapist needs to show empathy given the context.

2.2 Data-Driven Text-Based Research on Empathy in General Conversation

Recent years have witnessed a boom of research on data-driven analysis and application of empathy in general conversations.

In terms of empathy analysis for open-domain conversations, Zhou et al. (2021) addressed scoring empathy grounded in specific situations, Welivita and Pu (2020) created a taxonomy of empathetic response intents in social dialogues, while Guda et al. (2021) proposed to take user demographic information into account for empathy prediction.

As therapeutic conversation data is scarce, recent works on empathy analysis have also turned to peer-support dialogues from online communities. Zhou and Jurgens (2020) analysed Reddit[2] conversations for the relationships between condolence, distress and empathy, Hosseini and Caragea (2021) studied empathy seeking and providing with dialogues from a cancer survivor network, and Sharma et al. (2020) proposed an empathy framework of reaction-interpretation-exploration for conversations from mental-health-related online forums.

While early general empathetic chatbots (Zhou and Wang, 2018; Lubis et al., 2018) were mostly based on recurrent neural networks and produced emotion-conditioned output, their more recent counterparts are predominantly based on pre-trained language models and leverage emotions in various ways, including emotion detection as an auxiliary objective (Lin et al., 2020), emotion-based mixture-of-experts decoding (Lin et al., 2019), and rewarding response candidates likely to induce positive user emotion (Shin et al., 2020).

3 Data

We leverage[3] two types of data: general conversations and transcripts of MI demonstration videos.

We define an utterance as everything said by an interlocutor in their turn in a 2-person conversation, which is the most widely used definition of utterance in the literature of deep-learning-based conversational intelligence. This differs from some utterance definitions in psychotherapy. For example, an "utterance" in this work is identical to a "volley" as defined in the motivational interviewing skill code (MISC) (Miller et al., 2003), while an "utterance" in MISC is "a complete thought" that "ends either when one thought is completed or a new thought begins with the same speaker, or by an utterance from the other speaker".

3.1 General Conversations

Our general conversation data is from two datasets: **P**ersona-based **E**mpathetic **C**onversation (*PEC*) (Zhong et al., 2020) and *EmpatheticDialogues* (*ED*) (Rashkin et al., 2019). Their statistics are listed in Table 1. For each 2-interlocutor dialogue, we consider the initiator of the conversation as the **speaker** and the other as the **listener**.

PEC consists of general conversations crawled from 3 subreddits: *r/Happy*[4] (*r/H*), *r/OffMyChest*[5] (*r/OMC*), and *r/CasualConversation*[6] (*r/CC*). Reddit users exchange happy experiences and thoughts in *r/H*, share emotional stories that cannot be told easily in *r/OMC*, and simply talk casually in *r/CC*. Since the original *PEC* dataset includes conversations between more than two participants and some conversations are actually subsets of other conversations (e.g. a 2-turn conversation that in effect constitutes the first 2 turns of a 4-turn conversation), we retain only the non-subset conversations that are between 2 interlocutors, in order to align with the counsellor-client nature of therapeutic conversations, and the filtered *PEC* contains around 56% of the conversations in the original one.

EmpatheticDialogues (abbreviated as *ED*) is comprised of 23.1K general conversations from MTurker pairs. The speaker of each dialogue was first given an emotion label (e.g. "Afraid"), then described a situation where they had felt the emotion before (e.g. "I've been hearing noises around the house at night"), and finally initiated the conversation about this situation with a listener.

3.1.1 Empathy vs. Non-Empathy

We divide the general conversation data into 2 parts: empathetic-listener conversations and non-empathetic-listener ones. Specifically, we assign "empathetic" labels to all the listener utterances of the dialogues in *r/H*, *r/OMC* and *ED*, and "non-empathetic" to the counterparts in *r/CC*.

For *PEC*, the heuristic empathy labelling is based on the annotator ratings from the original paper that suggest comments (i.e. listener

[2]Reddit (https://www.reddit.com/) is an online platform comprised of subforums (known as **subreddits**), each with a specific topic for Reddit users to discuss.

[3]Identifiable information (e.g. names, dates) was replaced with placeholders prior to the experiments.

[4]https://www.reddit.com/r/happy/

[5]https://www.reddit.com/r/offmychest/

[6]https://www.reddit.com/r/CasualConversation

	r/Happy ‡			*r/OffMyChest* ‡			*r/CasualConversation* ¶			*EmpatheticDialogues* ‡		
Split	train	valid	test	train	valid	test	train	valid	test	train	valid	test
#Conv	113.9K	13.9K	16.0K	94.0K	12.1K	11.7K	530.2K	67.5K	66.9K	17.8K	2.8K	2.5K

Table 1: Statistics of *PEC* (*r/Happy*, *r/OffMyChest*, and *r/CasualConversation*) & *EmpatheticDialogues*. For *PEC*, we utilise 2-interlocutor conversations only. #Conv: number of conversations in the data split. We consider *r/Happy*, *r/OffMyChest* and *EmpatheticDialogues* to consist of mostly empathetic (‡) listener utterances and *r/CasualConversation* to be comprised of predominantly non-empathetic (¶) ones. Note that the statistics of *PEC* are about the filtered dataset as described in Section 3.1. See Table 4 for more details.

utterances) in *r/H* and *r/OMC* are significantly more empathetic than those in *r/CC*, and the inter-annotator agreement on this as measured by Fleiss' kappa (Fleiss, 1971) was "substantial". For *ED*, the empathy labelling is intuitive as the authors explicitly instructed the "listeners" to respond empathetically during the data collection.

We note that our heuristic labelling for *PEC* and *ED* is based on the corpus-level labels given by the creators of the datasets, thus it may not be completely accurate at utterance or sentence level. We nevertheless utilise the heuristic labels for our experiments and leave more fine-grained annotation for future work.

3.2 Motivational Interviewing

Our counselling conversations are from Pérez-Rosas et al. (2019), who collected the first and only (to the best of our knowledge) publicly available dataset of MI conversations. The dialogues are the transcripts of 152 demonstrations of high-quality (MI adherent) and another 101 of low-quality (MI non-adherent) counselling from video-sharing platforms such as YouTube and Vimeo. The original transcripts were obtained with the automatic captioning tool of YouTube, so the conversations have minor transcription errors and are mostly without punctuation. We refer to this dataset as ROLEPLAYMI, and list its statistics in Table 2.

3.2.1 Manual Empathy Annotation

We select a subset of ROLEPLAYMI to manually annotate utterance-level empathy to build a benchmark dataset for our models. The annotation guideline follows the definition of high empathy in MISC: *Counsellors high on the empathy scale show an active interest in making sure they understand what the client is saying, including the client's perceptions, situation, meaning, and feelings.* We ask the annotators to consider an utterance that shows MISC-defined *high empathy* as **empathetic**, otherwise as **non-empathetic**. Thus, non-empathy in this context can range from neutrality to apathy.

	ROLEPLAYMI		ANNO	
MI Quality	High	Low	High	Low
#Conv	152	101	7	14
#T-u	3928	1534	217	214
%(emp.T-u)	n/a	n/a	38.7%	2.3%
%(¬Q.T-u)	n/a	n/a	71.9%	73.8%
p(emp | ¬Q, T-u)	n/a	n/a	**0.50**	**0.03**
p(emp | Q, T-u)	n/a	n/a	0.10	0.00

Table 2: Statistics of ROLEPLAYMI and ANNO. #Conv: number of conversations in the subset. "T-u" is short for "Therapist Utterance(s)". #T-u: number of therapist utterances in the subset. %(emp.T-u): percentage of empathetic therapist utterances. %(¬Q.T-u): percentage of non-question therapist utterances. p(emp | ¬Q, T-u): probability of a non-question therapist utterance being empathetic. p(emp | Q, T-u): probability of a question therapist utterance being empathetic. See Table 5 for more details.

We choose 7 transcripts (217 counsellor utterances in total) from the high-quality subset with negligible transcription errors, and 14 transcripts (214 counsellor utterances in total) from the low-quality one. The 431 selected utterances are presented to 2 human annotators for binary utterance-level empathy annotation. One annotator is a senior researcher that has received formal MI training in the past, and the other is a PhD student that has read in depth about MI (incl. Rollnick et al. (2008)). Their annotations show an inter-annotator agreement of 0.71 measured by Cohen's kappa (Cohen, 1968), indicating "substantial agreement". Finally, the annotators discussed their results and resolved the differences. The annotated MI conversations are denoted as ANNO in the rest of the paper.

As Table 2 shows, 38.7% of the therapist utterances in the high-quality subset are empathetic (i.e. 61.3% non-empathetic), while the number for the low-quality subset is 2.3% for empathetic (i.e. 97.7% non-empathetic), suggesting a marked difference between the empathy levels in high- and low-quality counselling.

We note that our empathy annotation is at utterance-level on the punctuation-free MI tran-

scripts, which means an utterance is marked as empathetic as long as a part of the utterance is so, even though the remainder might not be. More fine-grained annotation would be possible with punctuated utterances, which we leave for future work.

3.2.2 Question & Empathy

Empirically, we observe that questions in MI do not show empathy in general, which is intuitive since the purpose of questions is to gather more information. Indeed, we notice that the vast majority of the examples of open and closed questions provided by MISC are not empathetic.

Therefore, we additionally conduct binary annotation for each therapist utterance in ANNO as to whether the utterance is (predominantly) a question, by marking an utterance as a question utterance if more than half of the tokens in an utterance constitute at least one open or closed question as defined by MISC. For instance, "it's good to see you up and about how are you feeling after your last little hospitalization" is considered a question utterance, since "how are you feeling after your last little hospitalization" is an open question and makes up more than half of the utterance. We denote the non-question subset of ANNO as $\neg$Q.ANNO.

The relationship between empathy and question found in ANNO confirms our observation: a non-question therapist utterance from high-quality counselling is substantially more likely (0.50) to be empathetic than one from low-quality counselling (0.03), while the same does not hold for question therapist utterances: 0.10 for high-quality and 0.00 for low-quality, which indicates that therapist questions are overall very unlikely to be empathetic.

3.3 General-Conversation Empathy vs. Therapeutic Empathy

Comparing ROLEPLAYMI with *PEC* & *ED*, we noticed a pronounced difference between empathy in general conversation and therapy: an MI-adherent therapist tends to express empathy through non-questions (as shown in Table 2), e.g. "The blood sugars have increased some, so you're concerned that things are not as good as they were last time that we talked". Conversely, participants in general conversations often show empathy via questions, e.g. "Oh no! That's scary! What do you think it is?". Thus, analysing sentence-level empathy (instead of utterance-level) could better separate the empathetic and non-empathetic parts, and more overlap between general-conversation empathy and therapeutic empathy may be found in the non-question sentences. This was not possible in our experiments as ROLEPLAYMI is not punctuated, thus we leave it for future work.

We note that another domain difference is that ROLEPLAYMI consists of transcripts of spoken dialogues whereas *PEC* and *ED* contain "written" chat conversations. The difference is smoothed by the high-quality transcription of the ROLEPLAYMI videos and we therefore do not use specific techniques to address the difference, but we plan to investigate this factor further in future work.

4 Binary Empathy Classification

In this section, we first define the task of binary empathy classification, then lay out the out-of-domain empathy contrast strategy behind our supervised models for the task, and finally describe our unsupervised baselines driven by NLI.

4.1 Task Definition

We denote $D^{MI} = \{(u_i^C, u_i^T, e_i)\}$, $i = 1, \cdots, N$ as a collection of {(*client utterance*, *therapist utterance*, *empathy label*)} tuples, where u_i^T is the therapist reply to the client utterance u_i^C, $e_i \in \{\texttt{emp}, \neg\texttt{emp}\}$ denotes if u_i^T shows empathy, and N is the number of such tuples in the dataset. Our task can be formulated as follows: given u_i^T and optionally u_i^C for more context, predict the correct empathy label e_i of u_i^T. We use ANNO as D^{MI}.

4.2 Supervised Learning: Using Out-of-Domain Empathy Contrast

Since our manually annotated subset of ROLEPLAYMI is too small to be a proper training set, we resort to learning from out-of-domain (i.e. non-MI) (OOD) empathy contrast. Specifically, as described in Section 3.1.1 and Figure 1, we utilise all listener utterances in *r/H*, *r/OMC* and *ED* as positive (empathetic) examples and their counterparts in *r/CC* as negative (non-empathetic) examples, as we aim to leverage parallels between general-conversation empathy and psychotherapeutic empathy.

We build 3 empathy vs. non-empathy contrast[7] pairs from general conversations: (*r/H* vs. *r/CC*); (*r/OMC* vs. *r/CC*); (*ED* vs. *r/CC*). For each pair, we sample an equal number of examples from the empathetic (positive) and non-empathetic (negative) subsets to construct a contrast dataset

[7] We use "empathy vs. non-empathy contrast" and "empathy contrast" interchangeably.

$\mathbf{P}^a$	Client: Everyone's getting on me about my drinking. | Therapist: Kind of like a bunch of crows pecking at you.	Relationship
$\mathbf{H}^b$	The therapist is empathetic towards the patient	Entailment
	The client wants to smoke more.	Neutral
	The therapist is not listening to the client.	Contradiction
[a]**P**, Premise. [b]**H**, Hypothesis.		

Table 3: Natural Language Inference, example utterances from Miller et al. (2003)

$D^{Gen} = \{(u_j^S, u_j^L, e_j)\}$, where in each sample the empathy label $e_j \in \{\texttt{emp}, \neg\texttt{emp}\}$ denotes whether the listener response u_j^L is empathetic towards its preceding speaker utterance u_j^S. Our sampling ensures that the 2 classes (i.e. `emp` & $\neg$`emp`) in each pair during training are balanced.

For each contrast pair, we train a 1-utterance general-conversation empathy classifier $cls_{(1)}$ to predict e_j given u_j^L, as well as a 2-utterance counterpart $cls_{(2)}$ to predict e_j given (u_j^S, u_j^L). Finally, we apply the trained $cls_{(1)}$ and $cls_{(2)}$ directly on D^{MI}, using u_i^C as u_j^S and u_i^T as u_j^L.

4.3 Unsupervised Baseline: Text Classification as Natural Language Inference

Natural language inference (NLI) is the task of determining if a **hypothesis** is true (*entailment*), false (*contradiction*), or undetermined (*neutral*) given a **premise**[8] (Table 3). Following Yin et al. (2019) where NLI models prove effective as ready-made zero-shot sequence classifiers, we formulate our empathy classification task as an NLI problem.

Assuming only u_i^T is available, we use it as the premise, and define the 1-utterance empathy hypothesis $h_{(1)}$ as "This text is empathetic.". We then utilise an off-the-shelf NLI model M as an unsupervised 1-utterance empathy classifier $nli_{(1)}^E$ to directly predict a label from {*entailment*, *contradiction*, *neutral*} given $(u_i^T, h_{(1)})$. We consider u_i^T to be classified as an empathetic utterance only if the predicted label is *entailment*.

We also investigate a client-therapist exchange scenario where both u_i^C and u_i^T are provided. The premise p_i is then formatted as "Client: u_i^C | Therapist: u_i^T", and we define the 2-utterance hypothesis as $h_{(2)}$ = "The Therapist is empathetic towards the Client.". We use the same M as an unsupervised 2-utterance empathy classifier $nli_{(2)}^E$ given the input $(p_i,\ h_{(2)})$. Again, only *entailment* is deemed equivalent to categorising u_i^T as empathetic.

4.4 Unsupervised Baseline: Client-Therapist Agreement as Natural Language Inference

It is our observation from MISC as well as ROLEPLAYMI that an empathetic therapist tends to acknowledge the difficulties and feelings of clients, and hence we experiment with NLI-style modelling for client-therapist agreement.

Specifically, we use M as an unsupervised 2-utterance agreement classifier $nli_{C\rightarrow T}^A$ to measure the agreement between u_i^C and u_i^T, using the former as the premise and the latter as the hypothesis. We only interpret an *entailment* prediction from M as the therapist agreeing with the client and hence the therapist empathising with the client.

5 Experiments

5.1 Implementation

For OOD empathy contrast (Section 4.2), we keep the original train/dev/test splits of *PEC* and *ED*. Since the two datasets in each contrast pair can be vastly different in their sizes (e.g. *ED* has only 17.8K training examples whereas *r/CC* has 530.2K), we always sample the positive and negative subsets so that their sizes are identical to that of *ED*, the smallest dataset, which ensures **a)** the two classes are balanced in each pair, and **b)** different cls models are trained with equal amounts of data and their performances are hence comparable.

To minimise the bias in training data caused by such sampling, we train the classifier of each contrast pair 5 times, each time with its own randomly sampled data. Note that this leads to 5 different groups of class-balanced {train, dev, set} datasets for each pair.

We leverage pre-trained language models for all our experiments. BERT (Devlin et al., 2019) is the backbone of our OOD empathy contrast models and its `BERT-BASE-UNCASED` variant is chosen. We add a fully connected layer atop the classification token (`[CLS]`) position of the language model to implement a binary classifier, and train the entire model end-to-end on the empathy contrast pairs. For the backbone M of the unsupervised zero-shot baselines, we use the `BART-LARGE` variant

[8]Definition of NLI: https://paperswithcode.com/task/natural-language-inference

of BART (Lewis et al., 2020) that has been fine-tuned on MultiNLI (Williams et al., 2018). For more details, see Section B.

To measure model performance on ANNO, we choose Matthews correlation coefficient (MCC) since it is robust to class imbalance, taking into account that only 38.7% of the ANNO examples from the high-quality subset are marked as empathetic and the number is only 2.3% for low-quality. We also use MCC to measure test set performance to increase comparability.

5.2 Results

We examine the performances achieved on ANNO by the models introduced in Section 4 , namely the blue bars in the "OOD$_{(1)}$ w/ Contrast" (1-utterance models trained on OOD empathy contrast, i.e. $cls_{(1)}$), "OOD$_{(2)}$ w/ Contrast" (2-utterance models trained on OOD empathy contrast, i.e. $cls_{(2)}$), and "Baselines" subplots of Figure 2. The value of each blue bar indicates the mean MCC of the 5 models from the corresponding pair, and we use the error bar to simply represent +/- one standard deviation from the mean, in order to illustrate the variation among the scores of the 5 models.

Also, we show in Figure 3 the performances of the OOD models on their respective test sets. In the test set of each of the 5 models from a (D_+, D_-) OOD pair, we have N_T random samples from D_+ and another N_T from D_-, where N_T is the size of the original test set of *ED*, in line with our sampling method for the OOD training sets. The mean (bar value) - standard deviation (error bar) representation follows that of Figure 2. By comparing the scores of the 5 models from an OOD setup on their own test sets and on ANNO, it becomes clear how the domain shift from general conversation to MI affects the performance of those models.

We first observe that while each test set in the OOD setups is different as we address class imbalance with random sampling, it is still obvious that the OOD models achieve considerably better scores on their test sets but experience significant drops on ANNO. In particular, *ED* vs. *r/CC* (2) reaches over 0.9 MCC on average on its test sets but only around 0.10 on ANNO. This stops any of the OOD empathy contrast models from being a reliable indicator of therapeutic empathy.

There is also considerable variation in the scores on ANNO (but not on the test sets) of the OOD models from the same empathy contrast pair. For instance, while *r/OMC* vs. *r/CC* (2) reaches 0.17 MCC on average, the standard deviation is 0.03. Further, we find that among the 5 models of the *r/OMC* vs. *r/CC* (2) pair, the MCC can be as high as 0.21 and as low as 0.11 despite that **a)** the 5 models only differ in the randomness of their training data sampling, **b)** the models have negligible variation in their test set performances (Figure 3). This pattern is present in all the OOD models, revealing their brittleness w.r.t. MI empathy classification.

As for the choice between 1-utterance and 2-utterance, the effects are mixed. Specifically, *r/H* vs. *r/CC* and *ED* vs. *r/CC* both have decreased performances on ANNO going from 1-utterance to 2-utterance, while *r/OMC* vs. *r/CC* benefits from this transition. In fact, in terms of the average score, *r/OMC* vs. *r/CC* (2) is the best setup. This could be because a client talks more about negative experiences in a therapy session, not unlike how the typical speaker shares emotional stores in *r/OMC*. In contrast, the speakers in *r/H* are more likely to tell positive experiences, which could explain the performance drop resulting from including the speaker utterance in *r/H* vs. *r/CC* (2).

The unsupervised zero-shot baselines do not fare better in general. $nli^E_{(1)}$ and $nli^E_{(2)}$ score around 0.05 and 0.02, respectively, both below most of the mean scores achieved by the OOD empathy contrast models. This can be attributed to the fact that knowledge gained from NLI tasks are not sufficient for reasoning about complex concepts such as empathy. $nli^A_{C\rightarrow T}$, on the other hand, shows better results and outperforms half of the OOD empathy contrast models, which suggests correlation between client-therapist agreement and therapist empathy. As a probing step, we swap the client and therapist utterances to reverse the premise-hypothesis formulation and observe that it ($nli^A_{T\rightarrow C}$) leads to a substantial drop to -0.04 MCC, further illustrating the aforementioned correlation.

5.3 Analysis

To shed light on the impact of the OOD design choices we made in Section 4, we add a **control** group of OOD models that are trained without empathy contrast for comparison, as shown by the blue bars in the "OOD$_{(1)}$ w/o Contrast", "OOD$_{(2)}$ w/o Contrast" subplots. More specifically, We build 3 pairs: (*r/OMC* vs. *r/H*), (*ED* vs. *r/H*), and (*ED* vs. *r/OMC*), as we consider them **(empathy vs. empathy)** pairs from which an OOD model is not

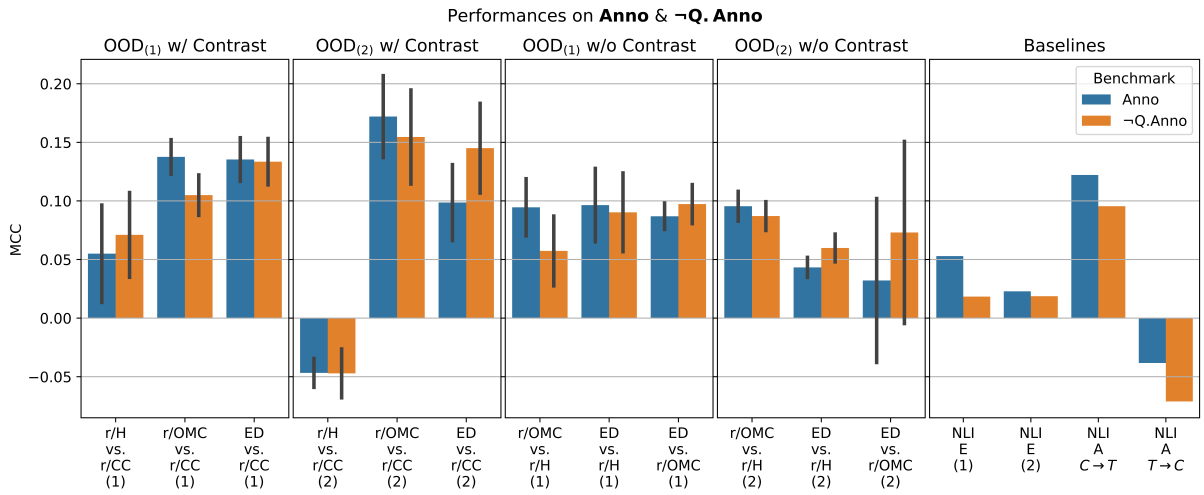

Figure 2: Results of all models on ANNO and ¬Q.ANNO, measured with Matthews correlation coefficient (Matthews, 1975). The names of the baseline models (shown in the rightmost subplot) are re-written in the figure for better visibility, e.g. "NLI\nE\n(1)" instead of $nli^{E}_{(1)}$). The first 4 subplots on the left show the performances of OOD-trained models. The first two show the performances of the 1- (e.g. *r/H* vs. *r/CC* (1)) and 2-utterance OOD models (e.g. *r/H* vs. *r/CC* (2)) trained on data **with** empathy contrast (e.g. *r/H* vs. *r/CC*, which is empathy vs. non-empathy), while the third and fourth show the performances of the 1- and 2-utterance OOD models trained on data **without** empathy contrast (e.g. *ED* vs. *r/H*, which is empathy vs. empathy). As explained in Section 5.1, for each OOD pair (e.g. *r/H* vs. *r/CC*), we randomly sample from the **class-unbalanced** OOD data 5 times to obtain 5 groups of **class-balanced** {train, dev, set} data, in order to address class imbalance and data selection bias. For each OOD pair, therefore, we train 5 models independently with the training data from their respective groups. Thus, the value of each rectangular bar indicates the mean of the scores of the 5 models from the 5 data groups of the corresponding OOD pair, and the error bar shows +/- one standard deviation from the mean.

able to learn **empathy vs. non-empathy** contrast. Additionally, we inspect the performances (orange bars) of all the models on ¬Q.ANNO to understand model behaviour in a less noisy context (i.e. question utterances removed).

Interestingly, the control group models score around 0.11 MCC and are not far behind empathy contrast models such as *r/OMC* vs. *r/CC* and *ED* vs. *r/CC* in the 1-utterance scenario, albeit with similarly large variation in their results. When it comes to 2-utterance, however, the lead of the empathy contrast models (except *r/H* vs. *r/CC*) becomes more obvious, with *r/OMC* vs. *r/CC* scoring over 0.15 MCC in contrast to *ED* vs. *r/OMC* recording less than 0.05. This shows that the benefit of learning from OOD empathy contrast, though small, does exist, and is more pronounced when **a)** compared against learning from no-empathy-contrast OOD data and **b)** more conversation context is taken into account by the models.

Finally, for the OOD contrast models, we notice mixed effects of removing questions from the benchmark dataset. It enables performance gains for *r/H* vs. *r/CC* (1) and *ED* vs. *r/CC* (2) but performance drops for the other OOD empathy contrast models. This shows that despite the annotations indicating that question therapist utterances are predominantly non-empathetic, whether a therapist utterance is a question generally does not substantially impact the empathy prediction of an OOD contrast model. One possible explanation, among others, is that the models simply did not learn to associate question with non-empathy during the OOD contrast training and instead learned to base its classification on semantic cues unrelated to question/non-question. Echoing Section 3.3, we argue that analysing non-questions at sentence level would be less noisy and better predictions would thus be possible, which we leave for future work.

6 Clinical Application & Impact

The motivation for this work was to minimise the annotation effort needed for training an utterance-level classifier of therapeutic empathy/non-empathy, based on the assumption that **1)** pre-trained language models can be fine-tuned to distinguish between empathy and non-empathy in general conversations, and **2)** the fine-tuned model can be leveraged to directly predict therapeutic empathy/non-empathy.

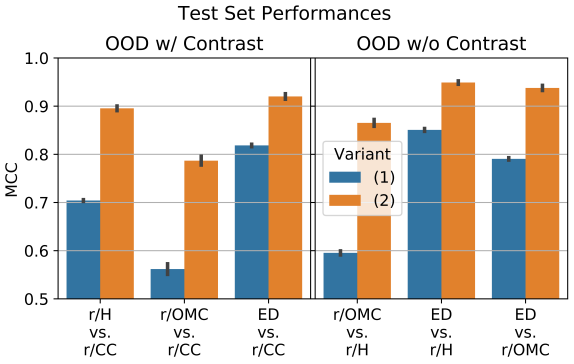

Figure 3: Test set performances (in MCC) of all OOD models. The first subplot on the left shows the test set performances of the 1- and 2-utterance OOD models trained on data with empathy contrast, and the second shows the test set performances of the 1- and 2-utterance models trained on data without empathy contrast. As explained in Figure 2, each OOD pair (e.g. *r/H* vs. *r/CC* (1) / (2)) corresponds to 5 groups of randomly sampled {train, dev, test} data and hence 5 trained models. Thus, the model trained on the training data of a group has a test set score associated with the test data of the group. Therefore, the value of each rectangular bar indicates the mean of the test set scores of the 5 models from the same OOD pair, and the error bar shows +/- one standard deviation from the mean.

Our results, for the most part, show that this simple OOD training approach did not sufficiently perform accurate classification, which limits its application in clinical settings. Compared to supervised learning of session-level empathy on sizeable corpora of well-annotated therapeutic conversations (Gibson et al., 2016), the task of utterance-level empathy classification with no in-domain training is more challenging and the models unsurprisingly fared worse. As discussed, the coarse, heuristic empathy labelling for the utterances in the training data and the domain gap between general conversation and therapeutic dialogue may have contributed considerably to the sub-optimal performance.

Nevertheless, we believe that this work is a meaningful step towards low-resource real-time assessment of empathy in counselling, and that the idea of utilising pre-trained language models for low-resource scenarios related to clinical psychology is still relevant. With smoothed domain gaps and more fine-grained annotation, future work can still use pre-trained language models to leverage parallels between empathy manifestations in general conversation and therapeutic dialogue. For instance, knowledge of empathy vs. non-empathy learned from well-annotated general conversations can serve as a bootstrapping step for empathy vs. non-empathy training on a minimal amount of well-annotated therapeutic conversations, since there can be a small to modest amount of therapeutic dialogue data available for a specialised domain instead of no data at all, which can take advantage of OOD empathy knowledge as a starting point for in-domain fine-tuning and thus maximise the benefit of OOD empathy training.

7 Conclusion

We find that our models trained to learn from empathy vs. non-empathy contrast in general conversation (i.e. out-of-domain w.r.t. counselling) are generally not reliable predictors of empathy/non-empathy in motivational interviewing. Upon probing, we observe that OOD empathy contrast learning is still marginally better than OOD learning without empathy contrast, particularly when more conversation context is available.

In future work, we plan to investigate more fine-grained empathy annotation and prediction, such as at sentence level, where we expect less noise and more accurate predictions. In addition, we will explore few-shot methods for the empathy classification task with out-of-domain empathy contrast training as a bootstrapping step.

Ethics & Privacy

Empathy often involves deeply personal circumstances (e.g. distress & struggle) and computational studies on it therefore warrant ethical consideration. The greatest ethical risk of this work has been privacy implications, as the conversational data we used could contain large amounts of sensitive identifiable information. To mitigate this risk, we experimented with only de-identified data where mentions of information like name, date, and location are replaced with placeholders. As a counterbalance, this study has considerable benefit as the first investigation of using knowledge of general-conversation empathy to support low-resource computational analysis of MI empathy, and the findings can inspire future efforts in making research on therapeutic empathy more accessible.

Acknowledgements

This work has been funded by the EC in the H2020 Marie Skłodowska-Curie PhilHumans project, contract no. 812882. The authors would also like to thank Dr. Mark Aloia for his guidance and support.

References

Jie Cao, Michael Tanana, Zac E. Imel, Eric Poitras, David C. Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5599–5611. Association for Computational Linguistics.

Zhuohao Chen, Karan Singla, James Gibson, Dogan Can, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2019. Improving the prediction of therapist behaviors in addiction counseling by exploiting class confusions. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 6605–6609. IEEE.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and Leslie S Greenberg. 2011. Empathy. *Psychotherapy*, 48(1):43.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Nikolaos Flemotomos, Victor R. Martinez, Zhuohao Chen, Karan Singla, Victor Ardulov, Raghuveer Peri, Derek D. Caperton, James Gibson, Michael J. Tanana, Panayiotis G. Georgiou, Jake Van Epps, Sarah P. Lord, Tad Hirsch, Zac E. Imel, David C. Atkins, and Shrikanth Narayanan. 2021. "am I A good therapist?" automated evaluation of psychotherapy skills using speech and language technologies. *CoRR*, abs/2102.11265.

Jacques Gaume, Gerhard Gmel, Mohamed Faouzi, and Jean-Bernard Daeppen. 2009. Counselor skill influences outcomes of brief motivational interventions. *Journal of substance abuse treatment*, 37(2):151–159.

James Gibson, David C. Atkins, Torrey Creed, Zac E. Imel, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2018. Multi-label multi-task deep learning for behavioral coding. *CoRR*, abs/1810.12349.

James Gibson, Dogan Can, Bo Xiao, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 1447–1451. ISCA.

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 *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 1947–1951. ISCA.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. Empathbert: A bert-based framework for demographic-aware empathy prediction. *CoRR*, abs/2102.00272.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mahshid Hosseini and Cornelia Caragea. 2021. It takes two to empathize: One to seek and one to provide. *Proceedings of the AAAI Conference on Artificial Intelligence*. To appear.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 121–132. Association for Computational Linguistics.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13622–13623. AAAI Press.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5293–5300. AAAI Press.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Jim McCambridge, Maria Day, Bonnita A Thomas, and John Strang. 2011. Fidelity to motivational interviewing and subsequent cannabis cessation among adolescents. *Addictive behaviors*, 36(7):749–754.

William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (misc). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Theresa B Moyers, Lauren N Rowell, Jennifer K Manuel, Denise Ernst, and Jon M Houck. 2016. The motivational interviewing treatment integrity code (miti 4): rationale, preliminary reliability and validity. *Journal of substance abuse treatment*, 65:36–42.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1426–1435. Association for Computational Linguistics.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Stephen Rollnick, William R Miller, and Christopher Butler. 2008. *Motivational interviewing in health care: helping patients change behavior*. Guilford Press.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. 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 2020, Online, November 16-20, 2020*, pages 5263–5276. Association for Computational Linguistics.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2020. Generating empathetic responses by looking ahead the user's sentiment. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 7989–7993. IEEE.

Karan Singla, Zhuohao Chen, David C. Atkins, and Shrikanth Narayanan. 2020. Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3797–3803. Association for Computational Linguistics.

Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment*, 65:43–50.

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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4886–4899. International Committee on Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zixiu Wu, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. 2020. Towards detecting need for empathetic response in motivational interviewing. In *Companion Publication of the 2020 International Conference on Multimodal*

Interaction, ICMI Companion 2020, Virtual Event, The Netherlands, October, 2020, pages 497–502. ACM.

Bo Xiao, Daniel Bone, Maarten Van Segbroeck, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2014. Modeling therapist empathy through prosody in drug addiction counseling. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 213–217. ISCA.

Bo Xiao, Dogan Can, Panayiotis G. Georgiou, David C. Atkins, and Shrikanth S. Narayanan. 2012. Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2012, Hollywood, CA, USA, December 3-6, 2012*, pages 1–4. IEEE.

Bo Xiao, Dogan Can, James Gibson, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2016a. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 908–912. ISCA.

Bo Xiao, Che-Wei Huang, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2016b. A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ Comput. Sci.*, 2:e59.

Bo Xiao, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2015. Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 2489–2493. ISCA.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3912–3921. 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 2020, Online, November 16-20, 2020*, pages 6556–6566. Association for Computational Linguistics.

Ke Zhou, Luca Maria Aiello, Sanja Scepanovic, Daniele Quercia, and Sara Konrath. 2021. The language of situational empathy. *Proc. of the ACM on Human-Computer Interaction*, 1:1–19.

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 2020, Online, November 16-20, 2020*, pages 609–626. Association for Computational Linguistics.

Xianda Zhou and William Yang Wang. 2018. Mojitalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1128–1137. Association for Computational Linguistics.

A Data

We list the complete statistics of the general conversation datasets in Table 4 and those of ROLEPLAYMI in Table 5.

B Implementation Details

All our pre-trained language models are implemented by the HuggingFace framework[9] (Wolf et al., 2019). All our models are implemented in PyTorch[10], while their evaluation is implemented with scikit-learn[11]. For $cls_{(1)}$, the input format to BERT is { `[CLS]` u_m^L `[SEP]` } during training and { `[CLS]` u_i^T `[SEP]` } during testing. Similarly, for $cls_{(2)}$, the input becomes { `[CLS]` u_m^S `[SEP]` u_m^L `[SEP]` } during training and { `[CLS]` u_i^C `[SEP]` u_i^T `[SEP]` } during testing.

During OOD training, we use a learning rate of $1e$-5 and a batch size of 32, and evaluate every 500 steps on the development set. We choose the Matthews correlation coefficient (Matthews, 1975) (`MCC`) as the metric for validation. We stop the training if the performance has not improved in the most recent 10 validations, and select the best checkpoint w.r.t. the development set.

We formulate the input to $nli_{(1)}^E$ as { `[CLS]` u_i^T `[SEP]` $h_{(1)}$ `[SEP]` }, and likewise { `[CLS]` p_i `[SEP]` $h_{(2)}$ `[SEP]` } for $nli_{(2)}^E$, { `[CLS]` u_i^C `[SEP]` u_i^T `[SEP]` } for $nli_{C\to T}^A$, and { `[CLS]` u_i^T `[SEP]` u_i^C `[SEP]` } for $nli_{T\to C}^A$.

[9] https://github.com/huggingface/transformers

[10] https://pytorch.org/

[11] https://scikit-learn.org/stable/

	r/Happy ‡			*r/OffMyChest* ‡			*r/CasualConversation* ¶			*EmpatheticDialogues* ‡		
Split	train	valid	test	train	valid	test	train	valid	test	train	valid	test
#Conv	113.9K	13.9K	16.0K	94.0K	12.1K	11.7K	530.2K	67.5K	66.9K	17.8K	2.8K	2.5K
μ(#S-u./Conv)	1.0	1.0	1.0	1.0	1.0	1.0	1.1	1.1	1.1	2.2	2.3	2.2
μ(#L-u./Conv)	1.0	1.0	1.0	1.0	1.0	1.0	1.1	1.1	1.1	2.1	2.1	2.1
μ(S-u.Len.)	30.8	30.2	30.4	48.9	51.0	47.8	42.8	42.9	43.2	17.6	19.4	21.2
μ(L-u.Len.)	13.3	13.5	13.3	15.7	15.7	15.6	16.9	16.8	16.8	13.7	14.3	14.5

Table 4: Statistics of *PEC* (*r/Happy*, *r/OffMyChest*, and *r/CasualConversation*) & *EmpatheticDialogues*. For *PEC*, we utilise 2-interlocutor conversations only. #Conv: number of conversations in the data split. μ(#S-u./Conv): average number of speaker turns per conversation. μ(#L-u./Conv): average number of listener turns per conversation. μ(S-u.Len.): average speaker utterance length (number of tokens), μ(L-u.Len.): average listener utterance length (number of tokens). We consider *r/Happy*, *r/OffMyChest* and *EmpatheticDialogues* to consist of mostly empathetic (‡) listener utterances and *r/CasualConversation* to be comprised of predominantly non-empathetic (¶) ones. Note that the statistics of *PEC* are about the filtered dataset as described in Section 3.1.

	ROLEPLAYMI		ANNO	
MI Quality	High	Low	High	Low
#Conv	152	101	7	14
#T-u	3928	1534	217	214
μ(#T-u/Conv)	25.8	15.2	31.0	15.3
μ(#C-u/Conv)	25.1	14.5	30.0	14.8
μ(T-uLen).	33.5	31.1	33.2	32.9
μ(C-uLen.)	28.5	20.6	24.4	21.6
%(emp.T-u)	n/a	n/a	38.7%	2.3%
%(¬Q.T-u)	n/a	n/a	71.9%	73.8%
p(emp | ¬Q, T-u)	n/a	n/a	**0.50**	**0.03**
p(emp | Q, T-u)	n/a	n/a	0.10	0.00

Table 5: Statistics of ROLEPLAYMI and ANNO. The abbreviation convention is similar to that in Table 4, while "T-u" is short for "Therapist Utterance(s)" and "C-u" for "Client Utterance(s)". #Conv: number of conversations in the subset. #T-u: number of therapist utterances in the subset. %(emp.T-u): percentage of empathetic therapist utterances. %(¬Q.T-u): percentage of non-question therapist utterances. p(emp | ¬Q, T-u): probability of a non-question therapist utterance being empathetic. p(emp | Q, T-u): probability of a question therapist utterance being empathetic.