A Critical Analysis of EmpatheticDialogues as a Corpus for Empathetic Engagement

ALOK DEBNATH and OWEN CONLAN, ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland

This paper aims to analyze the content and relevance of one of the most popular contemporary training corpora for empathetic conversational agents: EmpatheticDialogues [23]. We provide a detailed qualitative breakdown of the corpus including the corpus creation methodology and point out some critical shortcomings of the corpus. Given the significance of the corpus as the only one of its kind at the moment, we also provide a quantitative comparison of EmpatheticDialogues to other contemporary small-talk corpora including DailyDialog and Persona-Chat, including conversation length, the ratio of conversant interaction, lexical choice, etc. With this analysis, we discuss the merit and implications of indicating a specific small-talk dialogue corpus is more empathetic than other small-talk corpora. Finally, we provide a new lens for developing conversational agents with empathetic engagement capabilities by augmenting existing dialogue datasets.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**; **HCI theory, concepts and models**.

Additional Key Words and Phrases: datasets, empathy, corpus analysis

ACM Reference Format:
Alok Debnath and Owen Conlan. 2023. A Critical Analysis of EmpatheticDialogues as a Corpus for Empathetic Engagement. In *EmpathiCH workshop (EMPATHICH '23), April 23, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3588967.3588973

1 INTRODUCTION

One of the more prominent goals of dialogue systems in recent years has been to develop agents which have human-like conversation capabilities, including fluency and naturalness. An important characteristic of natural small talk[1] is the ability of a conversant to accurately perceive, understand, and respond to the emotional cues of the other participant; commonly referred to as empathy [20, 32, 34].

Empathy itself is a much broader term, encompassing cognitive and behavioural processes including associated concepts such as sympathy and compassion [6, 8, 22]. However, there are three key aspects of empathy that have been used in literature for artificially induced empathetic behaviour:

(1) **Emotion recognition**: the empathetic system recognizes the emotion (sometimes referred to more broadly as *experiences* to include both emotional state and context [5]) conveyed by the user alongside the stimuli or inputs required for the agent to perform its task successfully [17]. In the case of dialogue agents, emotion can either be a latent characteristic of the text, speech, or multimodal input being processed.

[1] small-talk, also referred to as chit-chat in dialogue systems, is the form of conversation where the participants do not have a pre-defined role concerning one another, as opposed to task-oriented dialogue wherein one participant is instructing another on a task and providing information on how to complete it [4].

Manuscript submitted to ACM

(2) **Perspective-taking**: the agent learns and models the user's thoughts and inference processes in a given situation, generally by contextualizing the recognized emotion within the user's preferences, personality, or goals, based on the task the agent is meant to perform or complete [2, 15]. For dialogue agents, perspective-taking relates the recognized emotion to the context of the conversation and other multimodal cues, if available.
(3) **Emotional contagion**: the agent conveys an artificial sense of experiencing either the same or an appropriate emotion to the one being conveyed by the user [13, 32]. Within empathetic conversational agents, emotional contagion is presented by generating an appropriate response to the perceived emotional cues and the contextualization of said emotions within the input or stimulus provided to the agent actively.

Therefore, empathetic engagement in conversation is not a latent characteristic of the dialogue itself, but rather the generation of a contextually and emotionally aware response, including changes in emotional cues during the conversation (emotion recognition), adaptive contextualization of conversation topic (perspective-taking), and the generation of an appropriate response (emotional contagion). The end goal of an empathetic dialogue system is the same as a "non-empathetic" one, i.e. to have a coherent conversation with the user.

One of the most prominent corpora in training dialogue models based on large language models (LLMs) is EmpatheticDialogues [23], a novel dataset of around 25,000 text-based conversations grounded in emotional situations across 32 distinct emotion labels. It is one of the only unimodal corpora on empathetic dialogue engagement for small talk.

In this paper, we analyze the quality and viability of the corpus, to answer the following questions:

(1) Does the corpus consist of empathetic dialogue interactions?, and
(2) How does the corpus differ from other small-talk corpora?

We present an in-depth analysis of the corpus creation methodology and highlight some shortcomings, such as the number and type of emotion labels used for creating the corpus, the association of emotion labels to a text-based conversation, and the relationship between emotion labels and empathy within the corpus. We then qualitatively analyze the corpus, highlighting the training methods used to validate the quality of the corpus. We also provide a statistical comparison of the corpus and other contemporary small-talk corpora including DailyDialog [16] and Persona-Chat [35]. Finally, we discuss our findings across both qualitative and quantitative metrics to better understand the role of a corpus like EmpatheticDialogues in training an agent to respond in an empathetic manner and a possible direction towards developing more empathetic models by changing the lens on existing dialogue corpora.

2 EMPATHETICDIALOGUES: A BREAKDOWN

Rashkin et. al. (2019) [23] claim that humans have been shown to interact with machines naturally and socially, which serves as one of the main motivations for the development of this corpus. In this section, we investigate the corpus creation methodology by isolating the three phases of the process: emotion label extraction, situation generation, and dialogue generation.

2.1 Emotion Label Extraction

Each dialogue interaction in the corpus is based on a situation grounded in one of **32** emotions. This is crucial, as the number of possible identifiable emotions is a contentious question in cognitive and behavioural research [14, 19, 30], and emotion labels are specific to the task and available modalities [11]. For example[2]:

[2]These works are references from [23]

(1) Scherer and Wallbott (1994) [24] references emotion universality in the context of cultural differentiation and shows that across 7 different emotions, the emergent patterns of subjective feeling are expressed in the same way across cultures, based on psychological symptoms and expressive behaviour in humans. The emotions are *joy, fear, anger, sadness, disgust, shame*, and *guilt*.
(2) Strapparava and Mihalcea (2007) [28] presents a SemEval task that uses Ekman's six basic emotion labels along with a positive/negative valence score to categorize newspaper headlines. The task was a short text classification task based on purposefully provocative texts with coarse-grained labels. Ekman's basic emotions include *anger, surprise, disgust, enjoyment, fear*, and *sadness* [7].
(3) Skerry and Saxe (2015) [27] presents neurophysiological experiments of fMRI readings associated with response to emotions elicited by *verbal* stimuli. The study does not establish a list of emotions but rather uses an independent set of experiments. The criteria for the choice of emotions is also compared against other coarse-grained emotion labels, the valence-arousal regression model, and a list of independent emotional event features. They use 20 emotions and show that humans possess the capability to express twenty distinct emotions under their experimental paradigm. The emotion labels used in the paper include *grateful, joyful, hopeful, excited, proud, impressed, content, nostalgic, surprised, lonely, furious, terrified, apprehensive, annoyed, guilty, disgusted, embarrassed, devastated, disappointed*, and *jealous*. Their experiments do not deal with dialogue.
(4) Li et. al. (2017) [16] is one of the most prominent dialogue corpora which uses Ekman's six basic emotions. The corpus comprises conversations by English learners online, a text-based forum for everyday communication which was scraped and then later annotated turn by turn for one of six emotions as well as other annotations.
(5) Mohammad (2012) [18] is one of the first works to use a word-emotion-associated lexicon for detecting emotions in tweets, using hashtags associated with Ekman's six basic emotions.

EmpatheticDialogues is a unimodal corpus annotated with conversations based on situations grounded in 32 different emotions, meant primarily for a dialogue generation task. These labels are *surprised, excited, angry, proud, sad, annoyed, grateful, lonely, afraid, terrified, guilty, impressed, disgusted, hopeful, confident, furious, anxious, anticipating, joyful, nostalgic, disappointed, prepared, jealous, content, devastated, embarrassed, caring, sentimental, trusting, ashamed, apprehensive*, and *faithful*.

Of these 32, labels such as *anticipating, confident, prepared, sentimental, trusting, caring*, and *faithful*, have not been chosen from any of the referenced literature. Labels like *"anticipating"* and *"prepared"* have been studied as individual cognitive and behavioural concepts differentiated from and interacting with, emotion [3, 9]. Whether 32 emotions can be distinguished in a single modality is not studied.

2.2 Emotionally Grounded Situation Generation

Once the emotion labels are identified and extracted, the next step in developing the corpus involved using annotators (MTurkers) to generate situations. After preprocessing, the situations are directly associated with that emotion label. Once generated, a pair of MTurkers are asked to have a conversation between 4-8 utterances long between one another about the situation presented. Note that the workers are not given the emotion label that the situation was generated from [23].

Firstly, the suitability of the situation written out by the worker for a given emotion is under scrutiny. The paper does not detail the studies done to determine confusion between situation and emotion, i.e. will an annotator identify the emotion in a given situation accurately? Will both annotators who are having this conversation respond to the

same emotional cue(s)? The suitability of the prompt for dialogue generation or even the detection of the associated emotion label as (latent) features of a natural dialogue interaction is not evaluated.

Secondly, the combination of emotion labels also invariably leads to confusion in assigning emotions to prompts (afraid vs. terrified; annoyed vs. angry vs. furious). While the paper addresses this and claims that the "goal in providing a single emotion label to have a situation strongly related to (at least one) emotional experience", there is no verification of the provided emotion label being strongly associated with the emotional experience. The authors do suggest that, if necessary, researchers using this corpus can merge similar emotion labels such as *"afraid"* and *"terrified"* into a coarser label if desired. [26, 31].

In conclusion, the use of these fine-grained labels referred to as emotions by the EmpatheticDialogues corpus does not provide the requisite theoretical or empirical background. On the theoretical side, the emotion labels include phenomena outside what is traditionally considered "an emotion" and it is never studied whether or not these phenomena occur as latent features of a dialogue interaction. On the empirical side, the "strong relation" between the generated prompts and the emotion labels is not established.

2.3 From Emotion to Empathy in Dialogue

In this section, we expound upon the development of an empathetic conversation from an emotionally grounded situation, ad provide some insight into the final corpus itself. Once the situations are written out, a pair of MTurkers are provided with the situation prompt (and **not** the emotion label), and are to have a written conversation with each other. This is done to simulate a natural conversation that is based on the emotional cues of the dialogue itself, i.e. that there is no explicit marker of the type of emotion being responded to.

We test this claim of "implicit emotional cues", by calculating the cosine similarity between each word in the prompt sentence and the emotion label (using GloVe embeddings [21]). We find that within the training corpus of 19,306 individual prompts, 2,192 of them have words that have a cosine similarity of 0.7 or more, and 1,015 prompt sentences use the *exact* emotion label in the prompt. For example, for the emotion label "sentimental", the first sentence of the dialogue is: "I always get sentimental when I look at old familly photographs."[3], or for the emotion label "afraid", the first sentence of the dialogue is: "I am so scared of my dog dying", wherein *scared* is a synonym of the emotion label *afraid*. Therefore, in about 11% of the cases, the participants of the dialogue generation process were provided with the "implicit" emotion label, and in 5% of the cases, the participants were directly informed of this label within the first sentence of the dialogue.

Within the dataset creation process, the initiator of the dialogue is known as the *Speaker* and the other participant is referred to as a *Listener*. The Speaker and Listener speak for between 4 to 7 turns with an average conversation length of 4.31 turns. The participants have a conversation based on the situation presented to them. Once collated, these conversations are presented as the EmpatheticDialogues corpus. Note that the dataset collection procedure also included balanced emotion coverage, i.e. workers were asked to provide additional dialogues for emotion labels which were least chosen overall in the dataset creation process.

Once developed, the dataset was used for a variety of generation and retrieval-based experiments involving automated and human metrics. The paper highlights that models trained on the Reddit dialogue corpus [1] and then fine-tuned on the EmpatheticDialogues corpus perform well against the test set of the latter corpus. Moreover, across human judgments, the models which use the EmpatheticDialogues were viewed as more empathetic, fluent, and were seen

[3]the word 'family' is incorrectly spelt in the corpus.

	ED	DD	PC
Size of the corpus	25k	13k	11k
Average Turns Per Dialogue	4.3	7.9	7.4
Average Tokens Per Utterance	15.2	14.6	12.8
Average Sentences Per Turn	1.59	2.05	1.85
Average Sentences Per Dialogue	6.8	12.5	21.5
Average Sentence Ratio (P1:P2)	1.69	1.13	1.21
Average P1 Tokens per Utterance	18.8	15.0	11.9
Average P2 Tokens per Utterance	11.6	14.4	13.7
Lexical Diversity	0.74	0.70	0.65

Table 1. A Comparison of Statistics Across the Corpora: "ED" is EmpatheticDialogues, "DD" is the DailyDialog corpus, and "PC" is the Persona-Chat dataset. P1 refers to Participant 1, the initiator of the conversation, also known as Speaker in EmpatheticDialogues, while P2 Participant 2, the responder, also known as the Listener in EmpatheticDialogues. All corpora are conversations between only two participants. Lexical diversity is computed using log type-token ratio.

to provide more relevant responses. This analysis was done by MTurk workers who were given subsampled test set examples from various models. Note that there is some ambiguity in this response evaluation as well because these workers could not interact with the model, only rate selected responses, which means either: (a) they were provided with a full dialogue between a human and the systems they evaluate, in which case there is a possibility of cherry-picking examples for higher human ratings, or (b) this evaluation only occurs on the first response generated by the model, i.e. there is no evaluation of the ability of the models fine-tuned on this corpus to *remain* empathetic, fluent, or relevant.

3 ANALYZING THE DATASET

In this section, we quantitatively analyze the corpus and provide some insight into its quality and use for fine-tuning dialogue generation models. We do so by first comparing the dataset statistics of this corpus to other popular contemporary dialogue corpora: DailyDialog and Persona-Chat. A quick statistical comparison between the three corpora is provided in table 1.

DailyDialog is an annotated multi-turn dialogue dataset that comprises conversations crawled from various websites where English learners and speakers have general, topic-focused, "daily" conversations. The corpus is then preprocessed for mistakes in spelling and syntax before being annotated for a variety of implicit features, including emotion, topic, and dialogue act. Note that their emotion annotation is specific to each turn of the conversation.

Persona-Chat is a corpus developed in order to train small-talk dialogue models to remain consistent in their personality over conversations and be able to ask and answer questions about personal situations and topics. The aim of corpus development was for the workers to "get to know each other", and therefore includes a variety of topics that include the participant's persona's stances and feelings towards those topics. The corpus is developed by first collating personas, after which MTurk workers are tasked with having a conversation when given these collated multi-line personas.

We highlight a few key points from the comparative statistics presented in Table 1:

(1) **Average Turns Per Dialogue**: On average, we see that EmpatheticDialogues has the lowest number of turns per dialogue, a little more than half the average number of turns across the other two corpora. This metric is an indicator of the average length of the conversation, i.e. models fine-tuned on EmpatheticDialogues are given a shorter dialogue overall. This also implies that appropriately recognizing and responding to emotional situations

can be showcased in approximately 4 turns of a conversation. In fact, the conversations are limited to 8 turns overall with no explanation for this limit.

(2) **Average Sentence Ratio**: The average sentence ratio determines how much the initiator of a conversation (P1) is involved in the conversation when compared to the responder (P2). Note that EmpatheticDialogues has a P1:P2 of 1.9, i.e. in the span of an average of four turns, the initiator is twice as involved in the conversation as the responder. When considering the dataset development methodology, the initiator (Speaker) is tasked with establishing the emotionally grounded situation, as well as providing enough emotional cues for the responder (Listener) to appropriately respond. However, in a dialogue model fine-tuned using this corpus, the model is *always* the responder.[4]

Other comparative statistics also showcase that EmpatheticDialogues is significantly different from other small-talk corpora, for example, the average token ratio across participants which is also skewed to the initiator and the low average number of tokens by the responder. These statistics further reinforce the idea that the participants in each conversation are not equally "responsible" for having an empathetic conversation.

4 DISCUSSION

In this section, we take a step back and focus on the aims and the need for developing an explicit corpus of empathetic engagement. We argue that empathy is not a latent characteristic of text or dialogue, the implications of developing a dialogue corpus for such a phenomenon, and some other possible methods to introduce the study of empathy into dialogue modeling.

4.1 Need for Empathetic Conversational Corpora

Why make a corpus specifically for empathetic conversations? Developing a corpus in any subfield of machine learning either showcases a novel problem or improves upon the shortcomings of existing corpora for an existing problem in machine learning [29]. Work like EmpatheticDialogues identifies a problem in the field of dialogue agents, develops a corpus that addresses this problem specifically, and provides a benchmark to solve this problem.

As defined in Section 1, in the context of an interaction, empathy is the ability to identify, contextualize, and respond appropriately to emotional cues. Within this paradigm, EmpatheticDialogues dataset attempts to capture exactly this phenomenon: an emotion label, a situation grounded in that emotion label, and a conversation based on that situation which showcases how an agent should respond when faced with that emotion-laden situation. Apart from validating the relationship between each step, the corpus does exactly what any novel corpus in machine learning aims to do: establish a new direction of research and provide a benchmark for the novel task of generating empathetic conversations. Therefore, developing this corpus comes with two key assumptions:

(1) empathy is a characteristic of dialogue that is captured in this corpus, and
(2) other corpora do not capture empathy as a phenomenon in dialogue.

We challenge the first assumption by pointing out that there is no existing method to validate whether an exchange is empathetic or unempathetic. Empathy is not a latent characteristic of the textual component of a conversation, but

[4]With regards to model training, corpora often remove the participant labels before they are tokenized, so the entire dialogue is provided to the model as a single paragraph of continuous text. Most modern dialogue models are trained and fine-tuned as purpose-specific large language models. However, it is still interesting to note that the dataset has the responder only half as engaged as the initiator when it was developed with the intent of empathetic engagement.

rather a response to the content, context, and cues of emotions and emotional states presented in the dialogue. There exists no method to validate that EmpatheticDialogues captures an empathetic conversation.

We challenge the second assumption by reinforcing the lack of a method to validate this corpus. Since it can not be validated that this corpus explicitly captures empathy as a concept in dialogue, we also can not claim that the EmpatheticDialogues corpus captures a phenomenon that is missing from other dialogue corpora. While it can be claimed that the high scores on human evaluation of the models should imply the model's ability to respond with empathy (i.e. validating that the corpus does, in fact, capture empathy well), we argue that their evaluation is limited to a single response (cf. from Section 2.3), which is not indicative of model performance throughout a longer dialogue interaction. Furthermore, the human evaluation question: "did the responses show understanding of the feelings of the person talking *about their experience*?" specifically focuses on the agent's ability to respond to an experience or event, which is the only type of engagement captured in the corpus.

In addition, there remains the open question of the requirement of human-like empathetic engagement by dialogue systems, given the vast differences in the type of conversations they are involved in vis-a-vis the user [12]. While discussions about the differences between expectations of human and artificial conversants and the associated empathetic engagement is of academic interest, it is beyond the scope of this paper.

4.2 Incidental Empathy

From the second assumption in Section 4.1, we present an interesting hypothesis, the *Incidental Empathy Hypothesis*. In the absence of our ability to validate the comparative degree of emotion-awareness and empathy presented by one corpus over another, the Incidental Empathy Hypothesis states that *Established small-talk or chit-chat corpora are just as empathetic as "empathetic" dialogue datasets.* Since small-talk corpora are transcripts of natural conversations, they inherently provide a degree of emotional understanding associated with empathetic engagement. While they are not generated for responding to emotions specifically, by being natural conversation, we establish that the corpora likely capture the attributes relevant to an empathetic corpus.

For example, DailyDialog is a corpus of real human interactions. Due to the variety of topics within the corpus, it is not focused on situations where conversation participants necessarily felt or portrayed a certain emotion explicitly. However, there is no merit to the claim that conversations in DailyDialog are not empathetic. Similarly, the Persona-Chat corpus provides transcripts of conversation participants attempting to both provide and model the other's persona. It is impossible to have such a conversation about participants getting to know each other better without empathetic engagement. We further argue that most conversational corpora do entail some recognition and attribution of participants' emotions and contextualization within the topic of conversation. Identifying and isolating trends of accurate emotion identification and appropriate response in context (such as: "when should a model exhibit mirroring behavior as opposed to emotion accommodation and validation?") is an interesting question for future research.

5 CONCLUSION

In this paper, we provide a critical analysis of the EmpatheticDialogues corpus. First, we provide a breakdown of the corpus creation methodology, which includes:

- a literature review of the emotion labels themselves, raising questions about the validity and references for some of the emotion labels such as *"prepared"* and *"faithful"*,

- a qualitative analysis of the situations grounded in those emotion labels to assert that there is limited evidence to prove the correlation between the prompt written out by annotators and the assigned emotion label, and
- an experimental analysis about the overt provision of implicit cues for the conversation participants within the first dialogue of the conversation in 11% of the training corpus (by using the exact emotion label or a synonym within the prompt).

We then compare some basic statistics of the EmpatheticDialogues corpus to other contemporary small-talk corpora such as DailyDialog and Persona-Chat and showcase the differences in the corpora, including the much smaller number of turns and total dialogue content, as well as the large skew of participation between the initiator and responder in a dialogue. Finally, we discuss the nature of empathy in conversation, and how the corpus does not tackle empathy as a concept in conversation, but rather as a small facet of empathetic engagement grounded in certain emotional situations.

In Section 1 we ask two questions. Through this paper, we endeavour to answer them as follows:

(1) Does the corpus consists of empathetic dialogue interactions?: Through our analysis, we find that the corpus consists of some empathetic dialogue interactions under the constraint that empathy can only be shown over a single retrospective incident linked to a single emotion which does not change throughout the conversation.
(2) How does the corpus differ from other small-talk corpora?: The corpus differs from other small-talk corpora as it has a shorter conversation length, higher participation skew towards the initiator, a higher lexical diversity, and several erroneous conversations. Since empathy is notoriously difficult to validate, it is hard to claim that this corpus is indeed more empathetic than any other small-talk corpus which has empathy as a natural byproduct of human interaction.

There is undoubtedly a need for conversational agents to be made more aware of the emotions and emotional state of a user during their interaction and tune their responses based on the recognized emotion [10, 25, 33]. However, the current methodology of developing a dataset to achieve this is difficult to validate and comes with several shortcomings, including the inability to claim that one dataset is fundamentally more empathetic than another and how. This corpus aimed to improve the model's ability to understand emotional cues in conversation. However, the corpus presents a compilation of conversations about situations where certain emotions were elicited in dialogue. This is only one facet of empathy and does not include an understanding of changes in the emotional state during the conversation. Therefore, we argue that the EmpatheticDialogues corpus is limited to being a corpus of short conversations about overtly emotional situations. There is more to empathetic engagement than that.

ACKNOWLEDGMENTS

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Trinity College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme. The authors also thank the anonymous reviewers for their insightful reviews and recommendations for improving the paper.

REFERENCES

[1] Katie Elson Anderson. 2015. Ask me anything: what is Reddit? *Library Hi Tech News* 32, 5 (2015), 8–11.
[2] Minoru Asada. 2015. Development of artificial empathy. *Neuroscience research* 90 (2015), 41–50.
[3] Cristiano Castelfranchi and Maria Miceli. 2010. Anticipation and emotion. In *Emotion-oriented systems: The humaine handbook*. Springer, 483–500.

[4] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter* 19, 2 (2017), 25–35.

[5] Henriette Cramer, Jorrit Goddijn, Bob Wielinga, and Vanessa Evers. 2010. Effects of (in) accurate empathy and situational valence on attitudes towards robots. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 141–142.

[6] Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. 2016. Empathy: A review of the concept. *Emotion review* 8, 2 (2016), 144–153.

[7] Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion* 98, 45-60 (1999), 16.

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

[9] Steven L Gordon. 2017. The sociology of sentiments and emotion. In *Social psychology*. Routledge, 562–592.

[10] Andreea Grosuleac, Stefania Budulan, and Traian Rebedea. 2020. Seeking an Empathy-abled Conversational Agent.. In *RoCHI*. 103–107.

[11] Hatice Gunes and Maja Pantic. 2010. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)* 1, 1 (2010), 68–99.

[12] Richard HR Harper. 2019. The Role of HCI in the Age of AI. *International Journal of Human–Computer Interaction* 35, 15 (2019), 1331–1344.

[13] Rachel Kirby, Jodi Forlizzi, and Reid Simmons. 2010. Affective social robots. *Robotics and Autonomous Systems* 58, 3 (2010), 322–332.

[14] Philip A Kragel and Kevin S LaBar. 2016. Decoding the nature of emotion in the brain. *Trends in cognitive sciences* 20, 6 (2016), 444–455.

[15] Iolanda Leite, André Pereira, Samuel Mascarenhas, Carlos Martinho, Rui Prada, and Ana Paiva. 2013. The influence of empathy in human–robot relations. *International journal of human-computer studies* 71, 3 (2013), 250–260.

[16] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 986–995.

[17] Sascha Meudt, Miriam Schmidt-Wack, Frank Honold, Felix Schüssel, Michael Weber, Friedhelm Schwenker, and Günther Palm. 2016. Going further in affective computing: how emotion recognition can improve adaptive user interaction. *Toward Robotic Socially Believable Behaving Systems-Volume I: Modeling Emotions* (2016), 73–103.

[18] Saif M. Mohammad. 2012. #Emotional Tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (Montréal, Canada) *(SemEval '12)*. Association for Computational Linguistics, USA, 246–255.

[19] Agnes Moors, Phoebe C Ellsworth, Klaus R Scherer, and Nico H Frijda. 2013. Appraisal theories of emotion: State of the art and future development. *Emotion Review* 5, 2 (2013), 119–124.

[20] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 1–40.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[22] Stephanie D Preston and Frans BM De Waal. 2002. Empathy: Its ultimate and proximate bases. *Behavioral and brain sciences* 25, 1 (2002), 1–20.

[23] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5370–5381.

[24] Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology* 66, 2 (1994), 310.

[25] Vincenzo Scotti, Roberto Tedesco, and Licia Sbattella. 2021. A modular data-driven architecture for empathetic conversational agents. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 365–368.

[26] AYAME SHIMIZU and KEI WAKABAYASHI. 2022. EFFECT OF LABEL REDUNDANCY IN CROWDSOURCING FOR TRAINING MACHINE LEARNING MODELS. *Journal of Data Intelligence* 3, 3 (2022), 301–315.

[27] Amy E Skerry and Rebecca Saxe. 2015. Neural representations of emotion are organized around abstract event features. *Current biology* 25, 15 (2015), 1945–1954.

[28] Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*. 70–74.

[29] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*. IEEE, 1521–1528.

[30] Terence J Turner and Andrew Ortony. 1992. Basic emotions: Can conflicting criteria converge? (1992).

[31] Baoyuan Wu, Siwei Lyu, and Bernard Ghanem. 2016. Constrained submodular minimization for missing labels and class imbalance in multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[32] Özge Nilay Yalçın. 2019. Evaluating empathy in artificial agents. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–7.

[33] Özge Nilay Yalçın. 2020. Empathy framework for embodied conversational agents. *Cognitive Systems Research* 59 (2020), 123–132.

[34] Özge Nilay Yalcin and Steve DiPaola. 2018. A computational model of empathy for interactive agents. *Biologically inspired cognitive architectures* 26 (2018), 20–25.

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

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009