Beyond Annotations: Labelling empathy from the listener's perspective

Zhilin Wang
University of Washington
zhilinw@uw.edu

Pablo Torres
University of Cambridge
pelt2@cam.ac.uk

Abstract

Third-party annotation is an approach that is often used to label empathy in text, alongside self-reported. However, annotations are often labor-intensive - which limits dataset size - and inaccurate given the annotators' position outside of a conversation. Similarly, labelling text based on responses to self-reported questionnaires results in highly noisy labels that are further confounded by socio-cognitive biases. To alleviate these limitations, we propose a novel approach to labelling empathy based on whether the listener of the text finds the text able to effectively understand and address their concerns. We show that our approach creates a dataset that is 8 times larger than any existing dataset and is supported by an established measure of empathy. Further analysis of our dataset also shows that the significant predictors of empathy are theoretically grounded and aligned with significant predictors in other empathy datasets.

1 Introduction

Empathy is defined as understanding a person from their perspective rather than one's own and indirectly experiencing a person's feelings, perceptions, and thoughts (American Psychological Association, 2021). Current datasets on empathy are obtained through either third-party annotations or self-reports by the speakers (a form of distance supervision). However, both approaches have their limitations.

1.1 Limitations of existing approaches

Annotations have high labor costs, constraining dataset size to a maximum of 10,143 samples and limiting the classes of models that can fit adequately. Moreover, annotations may not capture empathy accurately. This is because empathy concerns understanding a person from their frame of reference. Third parties who are not engaged in the conversation (typically volunteer annotators or crowdworkers) may not be in the best position to determine whether texts are empathetic because third parties do not have full information about the listener's frame of reference, as shown in Figure 1. A direct consequence of this is that annotators differ in their inference about the cognitive and affective effects that the text has on its intended audience (and hence the empathy of the text), leading to an imperfect inter-annotator agreement of 69-89% (Xiao et al., 2015; Khanpour et al., 2017; Sharma et al., 2020).

Figure 1: Different approaches to label empathy of text

On the other hand, labels of empathy based on speakers' self-reported questionnaires tend to be highly noisy. This is because attributes calculated from questionnaires reflect the general attributes of a person rather than the specific attributes communicated by a particular piece of their writing. Further noise in such labels comes from speakers' lack of self-understanding (Wilson and Dunn, 2004), compounded by speakers' self-serving biases such as wanting to appear socially desirable (Müller and Moshagen, 2019). Participants might answer that they are "good at predicting how someone will feel" because this is a socially valued attribute regardless of their true ability.

1.2 Our approach

Help-seeking post
My friend is upset but she won't tell me why. I asked what I can do to help but she doesn't want me to do anything. What do I say/do?

Empathetic comment
Give her space but let her know you respect her request. Say that when and if she wants to share, you're there to help.

Un-empathetic comment
You ignore it until she is grown up enough to communicate.

Table 1: Examples of empathetic and un-empathetic comments to a help-seeking post

To overcome such limitations, we introduce a new approach towards labelling empathy of text from the listeners'/recipients' perspective. Specifically, we obtained text from the r/Advice subreddit[1] where people ask for advice on issues they encounter, such as problems with family and friends, difficulties at school/work as well as troubles in pursuing one's interests and hobbies. Other users can then comment on these posts to attempt to help the post authors.

In response to these comments, r/Advice allows post authors to mark out comment(s) that they have found helpful[2]. This can indicate which comment(s) the post authors have found to be able to (emotionally and cognitively) understand and address the concerns that they have raised, and are by definition empathetic (American Psychological Association, 2021). Comments that were not labeled as empathetic were deemed to be un-empathetic. To minimize mislabeling of un-empathetic comments, comments on posts with no empathetic comment were excluded. This was done since authors of those posts likely did not actively label comments based on whether they were empathetic. An example is presented in Table 1 to illustrate the difference between empathetic and un-empathetic comments.

Our key contributions are as follows:

1. We introduce and validate a novel approach for labeling empathy in text.

2. We release this dataset, which is 8 times larger than the largest dataset on empathy currently available.

3. We explore this dataset, relating it to both theoretical literature on empathy as well as other empathy datasets.

2 Related work

	N
Ours	92477
Sharma et al. (2020)	10143
Buechel et al. (2018)	1860
Khanpour et al. (2017)	2107
Litvak et al. (2016)	202
Gibson et al. (2015)	348
Xiao et al. (2015)	200

Table 2: Size of Empathy datasets

2.1 Labelling empathy based on third-party annotations

Xiao et al. (2015) and Gibson et al. (2015) annotated the empathy of counselors during counseling sessions for those with alcohol and drug abuse problems. Counseling sessions were taped and converted into text using transcription software. The empathy of each text was labelled on an 8-point Likert scale. Khanpour et al. (2017) labelled the empathy of comments on a cancer survivors network forum in a binary fashion. Sharma et al. (2020) annotated messages on mental health support forums in terms of various aspects of empathy on a 3-point Likert scale.

2.2 Labelling empathy based on self-reported questionnaire

Litvak et al. (2016) collected participants' Facebook posts and labeled their empathy based on the participant's score on a widely-used self-reported empathy question (Davis, 1983). Buechel et al. (2018) asked participants to write short responses to sad news stories and labelled the empathy expressed in such writing based on participants' response on an empathy questionnaire (Batson et al., 1987) that focuses on affective aspects of empathy.

[1] https://www.reddit.com/r/Advice/

[2] This is done using the magic word "helped", which is picked up by AdviceFlairBot

3 Dataset

Empathetic comments refer to those marked out by post authors on r/Advice (i.e. listeners). Text from Reddit was downloaded through the Pushshift Application Programming Interface[3]. Suitable posts and all associated comments from the Advice subreddit were downloaded within 300 days (Apr 2019 - Feb 2020). Comments by the post authors and automated bots were excluded. Among the 24964 posts that were downloaded, there were 92477 comments (41146 empathetic). On average, each comment has 95.8 words (SD=134.5). 10% was randomly chosen as the test set while the remaining 90% was used as the training set. This dataset is 8x larger than existing datasets as shown in Table 2.

4 Validation

To validate our labeling of empathetic comments, we calculated an aggregated metric for each user (i.e. speaker) based on the proportion of their comments found to be empathetic. We then correlated User empathy against an established measure of empathy - the Empathy Quotient (EQ) questionnaire (Wakabayashi et al., 2006) - to determine the external validity of user empathy [4]. Higher scores on the EQ represent higher empathy. The EQ questionnaire contains items on both affective and cognitive aspects of empathy, and has high internal consistency (Cronbach's $\alpha = 0.90$) and test-retest reliability after 12 months ($r = 0.97, p < .001$). Additional details of the questionnaire are available in appendix A.

4.1 Participants

Only users with more than 20 comments were included to minimize the likelihood that their user empathy was biased due to chance events. 508 Reddit users were sent an online questionnaire through Reddit and 91 responded. Gender and age were optional to report. 86 participants reported gender (53 male and 33 female) and 83 reported age (M=33.7, SD=13.8). The mean user empathy is 0.5440 (SD=0.1956). Using a two-sample t-test, the distribution of EQ scores (M=24.45, SD=8.822, N=91) in this study is found to be not significantly different (t(1850)=0.0169, p=0.9866) from the sample (M=23.8, SD=8.75, N=1761) in Wakabayashi et al. (2006), demonstrating the representativeness of our sample.

4.2 Results

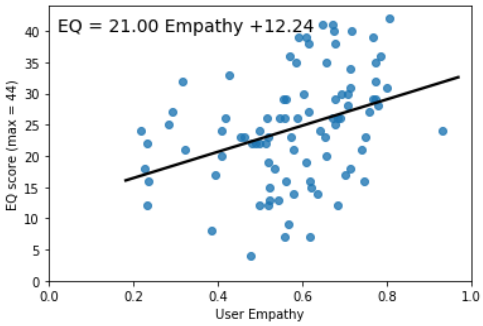

Figure 2: Empathy quotient (EQ) score against User Empathy

As illustrated in Figure 2, there is a moderate correlation effect between EQ and User empathy ($r(91) = 0.359, p < 0.001$). This means that they are likely to measure the same underlying construct of empathy to some degree.

5 Exploration of dataset

5.1 Performance of baseline models

	Micro-F1
BERT	**0.6881**
Logistic Regression	0.6484
Naive Bayes	0.6298
Support Vector Classifier	0.6394
Random Forest	0.6526

Table 3: Performance of baseline models on dataset. Details of their preparation are in Appendix B

To explore the potential for the dataset to be useful in training models to distinguish between empathetic and non-empathetic comments, we trained several baseline models. The performance of baseline models on this task is relatively low but similar to the performance on three existing data-sets on empathy (Khanpour et al., 2017; Gibson et al., 2015; Sharma et al., 2020). The relatively low performance of baseline models on this task suggests that while recognizing empathy in language is trivial for typically-developing humans, they remain challenging for machines. This suggests that alternative approaches such as integrate commonsense

[3] https://pushshift.io/

[4] We decided against comparing the empathy labels generated by our approach against those generated through annotations. This is because the extent of discrepancy between labels generated by the two approaches does not elucidate whether our approach is useful given that neither approach can be seen as the gold-standard in this context.

Direction	Themes	Words	Examples
Positive	Polite, friendly sounding words	personally, friend, glad, welcome, feels, hey	Me, **personally**...I'd let it slide. He'd be That's okay I'm just **glad** that you were able to maybe text her? Be like **hey**, just wanted to say
	Optimistic sounding words	good, luck hope, hopefully yes, learned, helped forward, strong,	session with your therapist. **Good luck** **hope** something I say can help you a little! And **yes** that is dangerous and quite work that you can look **forward** to.
	Words addressing the post author directly	you	I really think **you** deserve better. **You** sound like I understand that **you** really like these guys as long as **you** feel **you** are making the best of
Negative	Words indicating negative emotions	victim, kill, rid bad, depression	to be labelled as a **victim**. She might be afraid of I was internalizing every **bad** thing that happened
	Words that trivialize the problem faced by the post author	dealt, wish easy, promise advice, told	it's the latter, as I **dealt** with when I was like it seems like the **easy** solution to your situation. The best **advice** I can give you though

Table 4: Thematic categories for significantly predictors of Empathy

knowledge (Sap et al., 2019; Bosselut et al., 2019) might be explored to better capture the highly complex relationship between language and empathy.

5.2 Significant predictors of empathy

To understand what characterizes empathy in our dataset, significant predictors of empathy ($p < 0.05$) based on the Logistic Regression model were extracted and analysed[5]. Thematic categories based on these predictors are in Table 4 while their word clouds are available in Appendix C.

The first overarching theme is positive and friendly words. Empathy is positively predicted by polite, friendly-sounding and optimistic-sounding words but negatively predicted by words that indicate negative emotions. This supports the positive correlation that has been found between empathy and the attributes of friendliness (Melchers et al., 2016) and optimism (Hojat et al., 2015). Affect-related words (such as happy, sad, tears) were previously found to be significantly predictive of empathy (Gibson et al., 2015).

A second overarching theme is words that suggest attempts to understand the perspective of others. Empathetic comments do so by addressing post authors directly, instead of trivializing the difficulties that they face. This supports literature on the association between empathy and attempts to understand the viewpoints of others (Baron-Cohen and Wheelwright, 2004). Furthermore, terms indicating an inclination to find out more about the perspective of others (e.g. "do you think", "it sounds like" and "you think about") were also identified in a different empathy dataset (Xiao et al., 2015).

Overall, the overarching themes that are predictive of empathy in our dataset are supported by literature on empathy as well as the significant predictors in other empathy datasets. This further established the validity of our approach of labelling empathy

6 Conclusion

We propose a robust approach to label empathy in text and show that our proposed method correlates with an established method of measuring empathy. Our dataset is not only 8 times larger than current empathy datasets but can also capture the construct of empathy more adequately from the perspective of the listener. This can enable the development of models that can more effectively detect empathy (in human-human communication) and express empathy (in human-machine communication). Beyond empathy, the listener's perspective might also be useful to label other text attributes such as social bias (Sap et al., 2020), microaggression (Breitfeller et al., 2019) and interpersonal intimacy (Pei and Jurgens, 2020).

[5]The dataset used to extract the most significant predictors is slightly different. Only one comment was sampled from each post and author to overcome the problem that the covariance matrix was originally non-invertible

7 Ethics and Broader Impact

This project has been approved by an Institutional Review Board. The use of Reddit data in this project is in alignment with the Reddit End User License Agreement and the Terms of Use for Developers. Because part of the project requires participants to respond to questionnaires, we made sure that the items were phrased sensitively so that no unintended harm would be caused. We also guided participants to make informed decisions about their participation, giving them the opportunity to withdraw any time, during and after the completion of the questionnaire. The collected information, which does not include personally identifiable information was stored securely with access restricted to the research team. We anticipate that this project can accelerate the development of models that can better detect and express empathy in language.

References

American Psychological Association. 2021. American psychological association dictionary of psychology: Empathy.

Simon Baron-Cohen and Sally Wheelwright. 2004. The empathy quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, 34(2):163–175.

C. Daniel Batson, Jim Fultz, and Patricia A. Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of Personality*, 55(1):19–39.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

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

Mark H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1):113–126.

James Gibson, Nikolaos Malandrakis, Francisco Romero, David Atkins, and Shrikanth S. Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Proceedings of Interspeech*.

Mohammadreza Hojat, Michael Vergare, Gerald Isenberg, Mitchell Cohen, and John Spandorfer. 2015. Underlying construct of empathy, optimism, and burnout in medical students. *International Journal of Medical Education*, 6:12–16.

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

Marina Litvak, Jahna Otterbacher, Chee Siang Ang, and David Atkins. 2016. Social and linguistic behavior and its correlation to trait empathy. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 128–137, Osaka, Japan. The COLING 2016 Organizing Committee.

Martin C. Melchers, Mei Li, Brian W. Haas, Martin Reuter, Lena Bischoff, and Christian Montag. 2016. Similar personality patterns are associated with empathy in four different countries. *Frontiers in Psychology*, 7.

Sascha Müller and Morten Moshagen. 2019. True virtue, self-presentation, or both?: A behavioral test of impression management and overclaiming. *Psychological Assessment*, 31(2):181–191.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

Ashish Sharma, Adam Miner, David 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)*, pages 5263–5276, Online. Association for Computational Linguistics.

Akio Wakabayashi, Simon Baron-Cohen, Sally Wheelwright, Nigel Goldenfeld, Joe Delaney, Debra Fine, Richard Smith, and Leonora Weil. 2006. Development of short forms of the empathy quotient (EQ-short) and the systemizing quotient (SQ-short). *Personality and Individual Differences*, 41(5):929–940.

Timothy D. Wilson and Elizabeth W. Dunn. 2004. Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, 55(1):493–518.

Bo Xiao, Zac E. Imel, Panayiotis G. Georgiou, David C. Atkins, and Shrikanth S. Narayanan. 2015. "rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLOS ONE*, 10(12):1–15.

A Empathy Quotient Questionnaire

The short form of Empathy Quotient (EQ) questionnaire (Wakabayashi et al., 2006) is an established measure of empathy. Items originate from the long form of Empathy Quotient questionnaire (Baron-Cohen and Wheelwright, 2004). The short form was chosen to reduce the time taken to answer the questionnaire and increase the response rate. The short form is a 22-item forced-choice self-report questionnaire that can be answered on a four-point Likert Scale (Strongly Agree, Agree, Disagree, Strongly Disagree). Questions include "I often find it difficult to judge if something is rude or polite", "I can pick up quickly if someone says one thing but means another", and "I am good at predicting how someone will feel". Each response can give 0, 1 or 2 points, leading to a maximum total EQ score of 44. A higher score represents greater empathy.

B Baseline models

BERT The default pre-trained BERT English-base-uncased WordPiece tokenizer was used. We fine-tuned a BERT Sequence Prediction model (English-base-uncased)[6]. All hyper-parameters used to train the model were default[7] except adjusting the maximum sequence length to 512 tokens, batch-size per GPU to 8 and epoch number to 2. Training took 4 hours on a single Nvidia P100 GPU.

Others Text was split up into individual words and lower-cased. The number of times each word occurred in each text was then counted. Words that occurred fewer than ten times altogether were removed to minimize the effects of misspelled or rare words. Logistic Regression, Linear Support Vector Classifier, Multinomial Naive Bayes and Random Forest models were trained (accessed from https://scikit-learn.org/stable/) All hyperparameters were default except adjusting the number of estimators in the Random Forest model to 100. Training took negligible time (¡ 0.5 hours) on CPU.

[6] 12-layer, 768-hidden, 12-heads, 110M parameters accessed from https://github.com/huggingface/transformers

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

C Word clouds of significant predictors of Empathy

Size of words are directly proportional to their significance of correlation.

Figure 3: Significant positive predictors of empathy

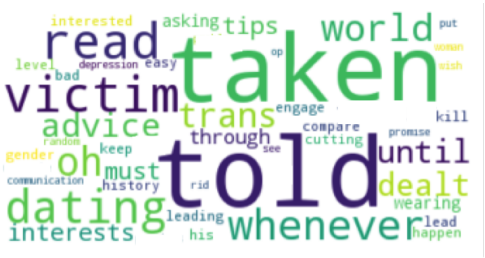

Figure 4: Significant negative predictors of empathy