Predicting Empathic Accuracy from User-Designer Interviews

Steven Nguyen* **Daniel Beck*** **Katja Hölttä-Otto****
*School of Computing and Information Systems, The University of Melbourne, Australia
**Department of Mechanical Engineering, The University of Melbourne, Australia
psnguyen@student.unimelb.edu.au d.beck@unimelb.edu.au
katja.holttaotto@unimelb.edu.au

Abstract

Measuring empathy as a natural language processing task has often been limited to a subjective measure of how well individuals respond to each other in emotive situations. Cognitive empathy, or an individual's ability to accurately assess another individual's thoughts, remains a more novel task. In this paper, we explore natural language processing techniques to measure cognitive empathy using paired sentence data from design interviews. Our findings show that an unsupervised approach based on similarity of vectors from a Large Language Model is surprisingly promising, while adding supervision does not necessarily improve the performance. An analysis of the results highlights potential reasons for this behaviour and gives directions for future work in this space.[1]

1 Introduction

User interviews are an important part of modern product development frameworks as meeting user needs defines success in Engineering Design. Typically these interviews, conducted between a potential user and a designer, are used to either gather knowledge about the user's problem or their experiences with current products, or to gain feedback on the product as it is being developed. However it remains a question as to whether these processes improve user understanding and lead to good outcomes, and the factors which contribute to these.

One such factor regards whether or not designers are able to understand the user during these interviews - this is referred to as 'empathic understanding' (Surma-aho and Hölttä-Otto, 2022). If a designer is able to grasp the user's experiences and thoughts, does this necessarily lead to better outcomes? To answer this question, Chang-Arana et al. (2020) developed a method borrowed from the social sciences to quantitatively measure empathic understanding through interviews. The method requires laborious manual annotation, involving the original user-designer pair and additional raters.

In this paper, we propose to use natural language processing (NLP) approaches to automate the measurement of empathic understanding in interviews, especially due to the advent of out-of-the-box Large Language Models (LLMs). This can not only streamline the process of analysing interviews in Engineering Design but also provide a test bed for automatically measuring empathy in conversations, an open problem in NLP. Automated evaluation in this way may be useful more broadly in other fields, where empathy is highly valued, such as teaching.

2 Background and Related Work

Work measuring empathy in NLP has been explored, with open domain dialogue data such as `EmpatheticDialogues` (Rashkin et al., 2019) existing as benchmarks for the task. Much work has been done detecting how empathy is expressed in dialogues in a variety of contexts from healthcare (Sharma et al., 2020; Xiao et al., 2015) in both speech and text, as well as in online communities (Zhou and Jurgens, 2020). However the theme of these works is primarily focused on empathy in the emotional sense. That is, there is a large focus on studying how individual express empathy towards others through dialogues. A common example is choosing the 'right' emotional words to comfort another individual in distress, guiding work in generating empathetic responses (Welivita et al., 2021).

On the other hand, in a review, Lahnala et al. (2022) points out that tasks revolving around *cognitive empathy* are not as prevalent in the NLP literature. While empathy is a complex concept, loosely we can distinguish between *emotional empathy* as

[1]Code used for our experiments is available at https://github.com/owowouwu/empathic-accuracy. Data is available under request to Katja Hölttä-Otto, katja.holttaotto@unimelb.edu.au.

Proceedings of The 21st Annual Workshop of the Australasian Language Technology Association, pages 125–129
November 29 - December 1, 2023

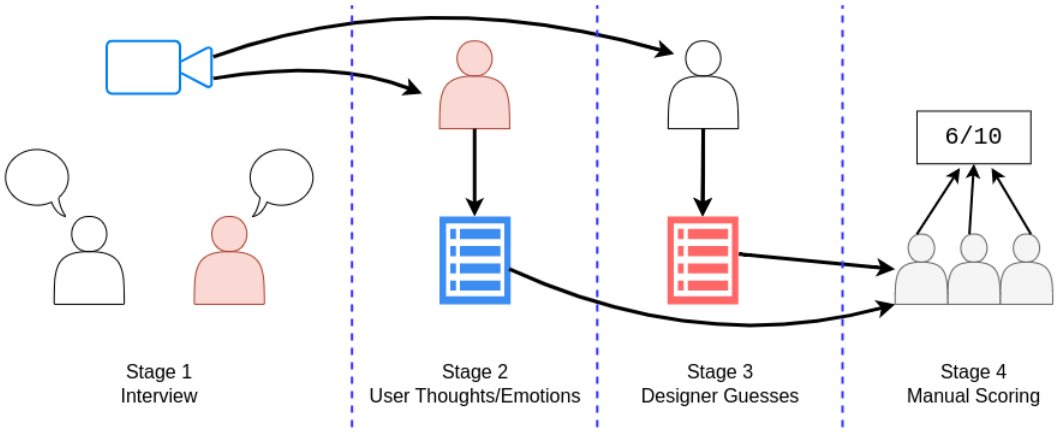

Figure 1: Overview of the collection of empathic accuracy ratings for interviews. Stage 1 represents the original user-designer interview, which is recorded. The same user then write their thoughts in Stage 2, with the designer aiming at guessing these thoughts in Stage 3. The annotation process finishes at Stage 4, where a set of human raters (3 in this case) assign a score to each aligned thought and guess, with the final empathic accuracy score being the average of these ratings.

processing and responding to another's emotions effectively, and *cognitive empathy* as being able to infer their thoughts in a broader sense (Cuff et al., 2016). One may be able to identify how another person is feeling and act appropriately, but may not necessarily know what the other person is thinking. A key distinction between our work and more common tasks involving 'empathy' in NLP is that we primarily try to measure cognitive empathy from pairs of thoughts.

3 Data

The dataset was collected from user-designer interview experiments in Salmi et al. (2023). Figure 1 gives an overview of the annotation process for empathic accuracy ratings. Each interview was recorded in video format. Interviewees were played back the recording and were asked at any time to pause the video and write down their thoughts. The same recording was played back to the interviewer, where they were tasked with guessing the user's thoughts in those moments.

In total, 46 users were interviewed by 3 designers, although not every user and designer were paired. Each instance of the dataset is indexed by a (user, designer) pair and contains a timestamped sentence pair - one being the user's thoughts at that particular moment, and the other being a guess of the user's thoughts by the designer at the same moment. Each pair is rated by 3 judges with a three-level Likert scale ($\{0, 1, 2\}$), with the average taken as a score indicating the accuracy of the designer's prediction. The designer is also tasked with predicting the user's self-evaluated tone of speech at that moment.

Designer/User- EEPK-01 Time : 10:52

USER: I was: Thinking that a "very improved version" of GPS's would really enhance driving experience.

DESIGNER: S/He was: not sure she explained herself well

rater1=0 rater2=0 rater3=0

user_tone = 0
designer_tone = -

Designer/User- EEPK-02 Time : 4:40

USER: I was concentrating. Lots of brainwork

DESIGNER: S/He was: trying to think how to formulate his answer

rater1=2 rater2=0 rater=1

user_tone = 0
designer_tone = 0

Figure 2: Example instances of data.

In this work, we focus on automating the rating stage (Stage 4 in Figure 1). Each instance contains a sentence pair (a user thought paired with a designer guess) as the input and the averaged rate given by the judges as the output. Figure 2 shows two such instances as an example. The inputs were preprocessed by removing text indicating the subject ("he/she/I was:") at the start of the string. We also rescaled the ratings to the unit interval. Table 1 details the statistics of our dataset.

4 Methods

All our models use Sentence BERT (Reimers and Gurevych, 2019, SBERT) as the LLM backbone, generating two embedding vectors for each pair

Designer	Instances	Avg. Score
1	120	0.414
2	129	0.519
3	200	0.398
All	449	0.437

Table 1: Summary statistics of our dataset. For each designer, we report their corresponding total number of sentence pairs and its average similarity score.

of user thought and designer guess. Experiments with in-domain supervised models were performed using 10-fold cross validation. For a sound comparison, we use the same 10-fold setup for the unsupervised and out-of-domain models, using only the testing folds for evaluation.

Unsupervised. Our first approach does not employ any training: we calculate the cosine similarity between the two embedding vectors and report the result as the rating.

In-Domain Supervised. Here we employ a standard cross-validation procedure, using 9 folds as training data. Our main approach finetunes a SBERT regression model following the original "siamese" method from (Reimers and Gurevych, 2019), which uses the cosine similarity between the embedding vectors as the regression output. In addition, we also employed SBERT as a feature extractor and two off-the-shelf regressors as additional models: a Gaussian Process (Rasmussen and Williams, 2006, GP) with an RBF kernel and a Multilayer Perceptron (MLP). Each input uses the concatenation of the SBERT vectors obtained from the user thought and the designer guess, plus the vector obtained from their absolute difference. On average, each training set contains 400 pairs.

Out-of-Domain Supervised. Finally, we also tested with a supervised approach trained on out-of-domain data. The rationale is that the rating can be framed as a Semantic Textual Similarity problem (Corley and Mihalcea, 2005, STS). This raises the question of whether we can employ existing STS data to create a good regressor without requiring any initial ratings for training. For these experiments, we used the widely available STS-B (Cer et al., 2017) dataset, containing approximately 6000 pairs. We used the same models as in the in-domain experiments.

5 Results

Our main results are shown in Table 2, using both Pearson's correlation and Root Mean Squared Error as evaluation metrics. As expected, the finetuned model on in-domain data gives the best performance. However, notably, it is not significantly better than the unsupervised model, potentially due to the limited amount of training data. This is further evidenced by the poor performance of the off-the-shelf regressors.

The models trainedd on the out-of-domain STS-B data did not outperform the unsupervised approach for any regressors. We believe this is due to significant differences in the STS-B and the Interview data. While both can be interpreted as sentence similarity, the pairs present in STS-B are much shorter and use simpler language, compared to the more complex sentences present in our dataset. While we were aware of this important domain difference, we still expected the performance to be better than the unsupervised approach, but our findings showed otherwise.

It is important to note that a Pearson score of 0.66 already demonstrates good prediction performance. Performance improvements could be obtained by adding in-domain training data and further model tuning. However, these results are already promising from an application perspective and could potentially lead to a reduction in human labour for obtaining empathic accuracy scores.

6 Qualitative Analysis

Here we will conduct further analysis on our data to understand the performance of our models under our task. We summarise three findings that could lead to further improvements in the prediction task.

Lack of Non-textual Context Textual similarity tasks rely on the meaning and context within the sentence itself, but in our case did not contain the extra information that raters may have when scoring pairs of text. The thoughts are often written down in an ad-hoc and conversational manner, containing *implied* information around the topic or interview itself that is able to be inferred by the raters, but which models which rely on complete information fail to do. This causes a mismatch between true scores and predicted scores. Our first instance in Table 3 shows this, as the designer is implicitly referring to the "AI system" that the user is mentioning, and is thus scored highly by the

Model	STS-B Test		Interviews	
	Pearson↑	RMSE↓	Pearson↑	RMSE↓
Unsupervised				
Cosine Similarity w/ SBERT	0.836	0.225	0.662 ± 0.060	0.227 ± 0.015
In-Domain Supervised				
Gaussian Process	-	-	0.562 ± 0.102	0.234 ± 0.022
Multilayer Perceptron	-	-	0.481 ± 0.153	0.263 ± 0.033
Finetuned SBERT	-	-	$\mathbf{0.680 \pm 0.050}$	$\mathbf{0.215 \pm 0.019}$
Out-of-Domain Supervised				
Gaussian Process	0.828	0.171	0.534 ± 0.074	0.240 ± 0.016
Multilayer Perceptron	0.800	0.191	0.515 ± 0.123	0.252 ± 0.028
Finetuned SBERT	0.858	2.424	0.618 ± 0.061	0.226 ± 0.017

Table 2: Summary of results. RMSE denotes root mean squared error. For the interview data, we report the average and standard deviation over 10 folds.

rater, but SBERT fails as in a vaccuum these two sentences do not have the same meaning without knowing what the designer refers to.

User: *thinking that this is quite hard to do in some kind of ai system*
Designer: *its technically hard to detect pedestrians*
True Score: 0.833
Predicted Score: 0.156

User: *you could just ask me what you want me to provide*
Designer: *feeling confused about the question and didn't know what answers the interviewer wants*
True Score: 0.833
Predicted Score: 0.249

Table 3: Example predictions for interview data.

Inconsistent Points of View Within our data it is often the case that the two pairs of text are written from two different points of view, resulting in sentences that may have similar content, but have different meaning. However they may still be rated highly because the designer, in their own writing, has effectively guessed the user's thoughts, even if they are not writing the thoughts from the perspective of the user.

Judge Scoring Our methods also tend to overestimate the scores in cases where the context or topic that both the designer and user are thinking of are the same, but the actual user text was different. For example, because the interviews were related to driving, both the user and designer wrote down thoughts related to driving, but these thoughts did not necessarily contain the same idea. In these cases, the human judges tended to more harshly assign scores of 0 whereas our system tended to provide a more soft assignment. This is a common problem of standard regression models, which are unable to predict extreme values outside a certain range. Future work should carefully consider how to penalise the scores based on how the two sentences diverge in actual meaning.

7 Conclusion

We introduce a novel task of predicting an individual's cognitive empathy as scored by their ability to predict, in text, the thoughts of another individual using a dataset from design engineering interviews. Using this data we demonstrate the performance and limitations of current state of the art models on our task. Our analysis shows that this problem poses unique challenges due to the unique structure and missing context of user written thoughts.

Initial directions for future work are based on our analysis in Section 6. Incorporating context from interview transcripts is an important direction, as well as improved regression models that can better predict extreme values. A more challenging, longer term goal is the prediction of empathic accuracy *directly from interviews*, without requiring user thoughts and designer guesses. This would

effectively bypass Stages 2 and 3 in Figure 1, drastically reducing annotation costs and potentially enabling real-time empathy feedback *during an interview*. We believe this is a much harder problem, but that nevertheless would lead to benefits to not just our task in engineering design, but lead to novel advances in other tasks in NLP.

Acknowledgements

This research was supported by The University of Melbourne's Research Computing Services, the Petascale Campus Initiative and a CIS-ME grant (2022) from The University of Melbourne.

References

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14. Association for Computational Linguistics.

Álvaro M. Chang-Arana, Matias Piispanen, Tommi Himberg, Antti Surma-aho, Jussi Alho, Mikko Sams, and Katja Hölttä-Otto. 2020. Empathic accuracy in design: Exploring design outcomes through empathic performance and physiology. *Design Science*, 6:e16.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18. Association for Computational Linguistics.

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

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158, Abu Dhabi, United Arab Emirates. 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 Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Carl Edward Rasmussen and Christopher K. I. Williams. 2006. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press. OCLC: ocm61285753.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Aleksi Salmi, Jie Li, and Katja Holtta-Otto. 2023. Automatic Facial Expression Analysis as a Measure of User-Designer Empathy. *Journal of Mechanical Design*, 145(3):031403.

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.

Antti Surma-aho and Katja Hölttä-Otto. 2022. Conceptualization and operationalization of empathy in design research. *Design Studies*, 78:101075.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A Large-Scale Dataset for Empathetic Response Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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):e0143055.

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)*, pages 609–626, Online. Association for Computational Linguistics.