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

Existing approaches of labelling empathy in text can be inaccurate and noisy. To address this, we propose a novel approach to label empathy based on whether the listener of conversational text finds written responses to their help-seeking posts able to effectively understand and address their concerns. We show that our approach is grounded in theoretical understandings of empathy, relate it to an established psychological measure of empathy and create a dataset that is 8 times larger than any existing dataset. Further analysis also shows that the significant predictors of empathy in our dataset are supported by literature on empathy and aligned with significant predictors in previously published 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). This consensus definition is informed by the dual emphasis on affective and cognitive aspects of empathy in theoretical literature (Davis, 1983; Baron-Cohen and Wheelwright, 2004; Zhou et al., 2003). Affective empathy concerns appreciating others' emotional states (Hoffman, 2000) while cognitive empathy centers around understanding how others are likely to think based on their experiences (Leslie, 1987). 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 significant limitations which cause them to capture empathy inaccurately or noisily, as detailed in Section 2.

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.

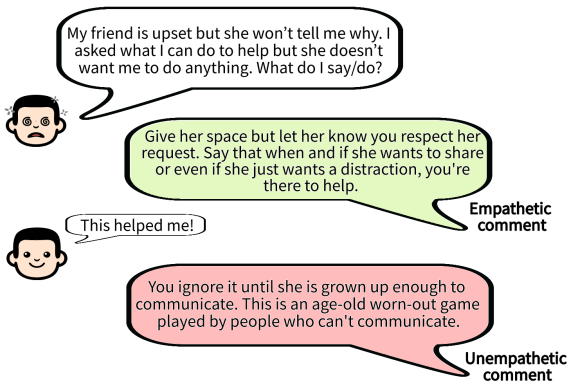

Figure 1: Examples of empathetic and un-empathetic comments to a help-seeking post

In response to these comments, r/Advice allows post authors to mark out comment(s) that they have found helpful[2]. Providing helpful comments not only requires commenters to emotionally and cognitively understand the post author's experiences, but also address the post author's concerns in a way that considers their current cognitive and emotional states. Both of these aspects require cognitive and affective empathy, and hence these comments can be seen as empathetic. Further discussion on the establishment of helpful comments as a marker of empathy can be found in Appendix A. Comments to posts with at least one empathetic comment, but were not themselves labeled as empathetic are labelled as un-empathetic. This inclusion criterion minimizes the mislabelling of comments to posts

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

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

whose authors did not actively participate in labelling comments. An example is presented in Figures 1 to illustrate the difference between empathetic and un-empathetic comments.

Our key contributions are as follows:

1. We introduce a novel approach for labeling empathy in text, and validate it by relating it to theoretical understandings of empathy and an established psychological measure of empathy.

2. We demonstrate that significant predictors of empathy in our dataset are supported by empathy literature and aligned with predictors in previously published empathy datasets.

3. We will make this dataset openly available, which will be 8 times larger than the largest dataset on empathy currently available.

2 Related work

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. 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.

Annotations have high labor costs, constraining dataset size thus far to a maximum of around 10,000 samples. Moreover, annotations may not capture empathy accurately. This is because empathy concerns understanding how a a person thinks and feels from their frame of reference. Third parties who are not engaged in the conversation (typically volunteer annotators or crowdworkers) cannot effectively determine whether such texts demonstrate an understanding of how the listener think and feel since third parties do not know about the listener's frame of reference, as shown in Figure 2. Therefore, third-parties make errors in labelling the empathy of text when they perceive the text differently from the intended listeners.

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.

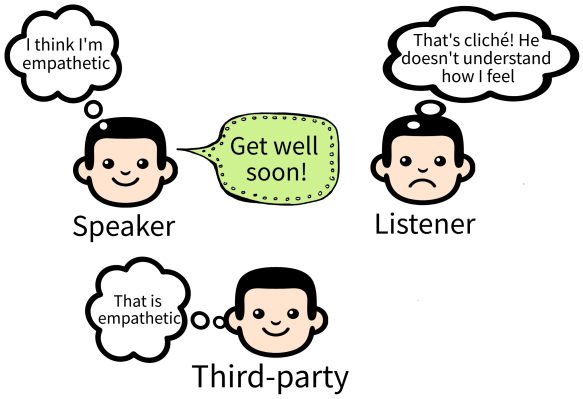

Figure 2: Different approaches to label empathy of text. Speaker/Listener refers to roles and text is in written form.

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.

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

[3] https://pushshift.io/

92477 comments (41146 empathetic). On average, each comment has 95.8 words (SD=134.5). Training/validation/test split was 80-10-10. This dataset is 8x larger than existing datasets as shown in Table 1.

Approach	Dataset	Samples
Listener	**Ours**	92477
Third-party (Annotations)	Sharma et al. (2020)	10143
	Khanpour et al. (2017)	2107
	Gibson et al. (2015)	348
	Xiao et al. (2015)	200
Speaker (Self-reported)	Buechel et al. (2018)	1860
	Litvak et al. (2016)	202

Table 1: Empathy datasets

4 Validation

To validate the labeling of empathetic comments, we calculated an aggregated metric (User Empathy) 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 psychological measure of empathy - the Empathy Quotient (EQ) questionnaire (Wakabayashi et al., 2006) - to determine the external validity of User Empathy. Higher EQ scores 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 on our choice of validation strategy and approach are available in appendix B.

4.1 Participants

Only users with more than 20 comments were included to minimize the likelihood that their user empathy was biased due to limited observations. 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

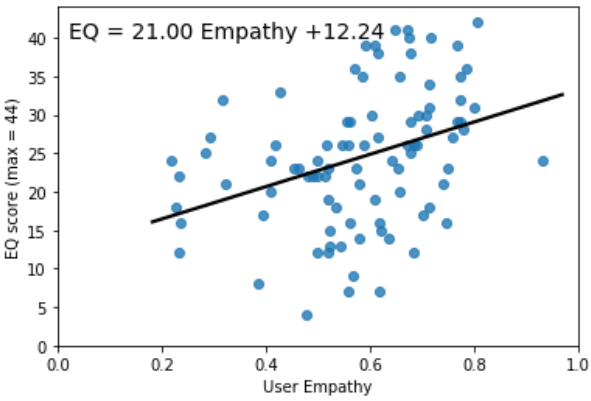

Figure 3: Empathy quotient (EQ) score against User Empathy

As illustrated in Figure 3, there is a moderate correlation effect between EQ and User empathy $(r(91) = 0.359, p < 0.001)$. Such correlation further supports our theory-grounded approach as an adequate measure of empathy.

5 Exploration of dataset

5.1 Performance of baseline models

	Micro-F1 (σ)
BERT	**69.2 (0.60)**
Logistic Regression	65.4 (0.55)
Naive Bayes	63.0 (0.44)
Support Vector Classifier	63.5 (0.59)
Random Forest	65.1 (0.60)

Table 2: Performance of baseline models on test set. Details of their preparation are in Appendix C

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 and report their micro-average F1 scores. The performance of baseline models on this task is relatively low but similar to the performance on three existing datasets 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. Techniques such as commonsense reasoning (Sap et al., 2019; Bosselut

Direction	Themes	Words	Examples
Positive predictors	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 predictors	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 patronize 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 3: Thematic categories for significantly predictors of Empathy

et al., 2019) can be explored in the future 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.[4] Thematic categories that were inductively generated from these predictors are shown in Table 3 while their word clouds are available in Appendix E.

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 sad and tears) were previously found to be significant predictors 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 patronizing 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 predictors 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 those of existing empathy datasets.

6 Conclusion

We propose a theory-grounded approach to label empathy in text based on whether the listeners in conversation text finds written responses to their help-seeking posts able to effectively understand their concerns. Not only does our approach correlate with an established psychological measure of empathy, predictors of empathy in our dataset are also supported by literature and aligned with predictors in existing empathy datasets. Finally, we have produced and plan to openly release an empathy dataset that is 8 times larger than current empathy datasets. Our work enables the development of models that can more effectively detect empathy in human-human communication and express empathy in a variety of dialogue systems, ranging from automated tutoring systems, social chit-chat agents and task-oriented dialogue.

[4]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*.

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.

James N. Butcher, Giselle A. Hass, Roger L. Greene, and Linda D. Nelson. 2015. *Using the MMPI-2 in forensic assessment.* American Psychological Association.

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

Nancy Eisenberg, Richard A. Fabes, and Tracy L. Spinrad. 2007. *Prosocial Development*, chapter 11. American Cancer Society.

Howard Gardner. 2003. *Multiple intelligences: the theory in practice*. BasicBooks.

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*.

Daniel Goleman. 2020. *Emotional intelligence*. Bantam Books.

David M. Greenberg, Varun Warrier, Carrie Allison, and Simon Baron-Cohen. 2018. Testing the empathizing–systemizing theory of sex differences and the extreme male brain theory of autism in half a million people. *Proceedings of the National Academy of Sciences*, 115(48):12152–12157.

Y. Groen, A. B. M. Fuermaier, A. E. Den Heijer, O. Tucha, and M. Althaus. 2015. The empathy and systemizing quotient: The psychometric properties of the dutch version and a review of the cross-cultural stability. *Journal of Autism and Developmental Disorders*, 45(9):2848–2864.

Martin L. Hoffman. 2000. Empathy and moral development.

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.

Keith Jensen. 2016. Prosociality. *Current Biology*, 26(16):R748–R752.

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.

Vladimir Kosonogov. 2014. The psychometric properties of the russian version of the empathy quotient. *Psychology in Russia: State of the Art*, 7:96–104.

Alan M. Leslie. 1987. Pretense and representation: The origins of "theory of mind.". *Psychological Review*, 94(4):412–426.

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.

Michael Quinn. Patton. 2002. *Qualitative research and evaluation methods 3rd. ed.* Sage Publications.

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.

Qing Zhou, Carlos Valiente, and Nancy Eisenberg. 2003. Empathy and its measurement. *Positive psychological assessment: A handbook of models and measures.*, page 269–284.

A Further discussion on the establishment of our measure as a marker of empathy

To ensure that our measure assesses only empathy, we established that our measure does not assess other constructs that can be confounded with empathy. For example, empathy could be confounded with people's level of pro-sociality. Pro-social behavior refers to behavior that intends to benefit other people or society as a whole (Eisenberg et al., 2007; Jensen, 2016) Therefore, pro-social behavior is about how often a person attempts to help others (by commenting) rather than their ability to *effectively* help others through empathetically understanding their issue and thereby responding appropriately. Empathy could also be confounded with emotional intelligence and interpersonal intelligence. Emotional intelligence (Goleman, 2020) involves self-awareness, self-regulation, motivation, empathy and social skills. The overlapping construct of interpersonal intelligence (Gardner, 2003) concerns sensitivity to others' moods, feelings, temperaments, motivations as well as abilities to empathize, communicate and cooperate with others. Among the skills that constitute emotional/interpersonal intelligence, helpfulness mostly indicates about a person's empathy because it is unrelated to one's self-awareness, self-regulation, motivation as well as aspects of social skills that do not tap on empathy (*e.g.* initiating social interactions). Therefore, our measure is most representative of empathy.

B Details of validation approach and strategy

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), which is well-cited (>3500 citations) and demonstrates good validity in large (>500,000) and culturally-diverse samples (Kosonogov, 2014; Groen et al., 2015; Greenberg et al., 2018). The short form was chosen to reduce the time taken to answer the questionnaire and thereby 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.

Comparison with manual annotations of empathetic text We decided against comparing the empathy labels generated by our approach against those generated through manual annotations. Manual annotations have not been shown to be valid measures of empathy from a psychological perspective, despite being often used for Natural Language Processing. These are usually based on face validity (whether a comment seem to be empathetic to a minimally-trained non-domain-expert: graduate student/crowd worker), which despite having validity in its name, is not useful in assessing the overall validity of a measure (Patton, 2002). A famous example in which face validity has been counterproductive is the Minnesota Multiphasic Personality Inventory (Butcher et al., 2015) for which useful items often do not appear relevant to psychopathological conditions that they seek to measure. Similarly, manual annotations of empathy by third parties who have little to no information about post author's perspective can be counterproductive. This is evidenced by the lack of attempt by existing annotation-based Empathy datasets (Khanpour et al., 2017; Gibson et al., 2015; Sharma et al., 2020; Xiao et al., 2015) to test for convergent validity with psychological measures of empathy. Therefore, it is not useful to validate our approach by comparing against manual annotations, when a well-established alternative - Empathy Quotient (Wakabayashi et al., 2006) - is available.

C Baseline models

Each model was run with 5 different random seeds.

BERT Pre-trained BERT English-base-uncased WordPiece tokenizer was used. We fine-tuned a BERT Sequence Prediction model (English-base-uncased version with 12-layer, 768-hidden, 12-heads, 110M parameters accessed from `https://github.com/huggingface/transformers`). BertAdam optimizer was used with 0.1 epoch for warmup and learning rate of $2 * 10^{-6}$ following a search within {1,2,5} * 10^n, $-6 \geq n \geq -4$ using F1 as criterion. Maximum sequence length was 512 tokens, batch-size was 8 and epoch number was 2. Training took 4 hours on a 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.

D Performance of baseline models (Validation set)

	Micro-F1 (σ)
BERT	**69.5 (0.52)**
Logistic Regression	65.1 (0.12)
Naive Bayes	62.9 (0.40)
Support Vector Classifier	63.5 (0.34)
Random Forest	65.2 (0.33)

Table 4: Performance of baseline models on validation set.

E Word clouds of significant predictors of Empathy

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

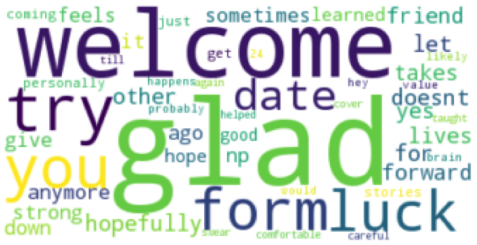

Figure 4: Significant positive predictors of empathy

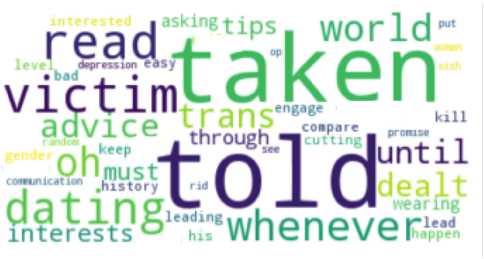

Figure 5: Significant negative predictors of empathy