Empathy Alignment in Online Communities

JIAMIN YANG, College of Engineering, University of Michigan, Ann Arbor, US

DAVID JURGENS (ADVISOR), School of Information, University of Michigan, Ann Arbor, US

1 INTRODUCTION

The project is inspired by the phenomenon that a growing number of people are talking about their own hard times on social media to seek comfort and advice. But people who replied to those posts may not be empathetic enough in their texts to be helpful. So questions are left in how to be more empathetic in our responses.

In this project, we refer to the appraisal theory, which is a psychology theory that characterizes the trigger of emotions into several dimensions. This will help us find the alignment between conversations where the response is closely related to the post. And this alignment can be a sign of empathy, which is defined as a person feeling the same way as another person who is experiencing some situation.

We developed a codebook for labeling datasets and a website tool to visualize and compare annotations across multiple annotators. Due to the difficulty in gathering golden datasets, we are still actively collecting data. Currently, we have 200+ annotated data, 1700+ unlabeled Reddit comment data and approximately 30M unlabeled per-month Reddit post data. As a reference to future work, we also proposed a model for analyzing the dataset in this paper. The final goal of the paper is to provide more insight into the understanding of empathy and facilitate the development of concerned psychology theories. And the potential application can be a guide in online communities to help people respond to others with more empathy.

2 RELATED WORK

Detecting and analyzing empathy in online communities has become an active research topic. Work has been done in identifying empathy in online discussions on breast and lung cancer, where CNN + LSTM is used as the main structure to predict whether a message is empathetic or not [5]. Sharma et. approached the task of rewriting conversations to reflect more empathy that can be potentially helpful to facilitate online mental health support [9]. In their work, PARTENER, a deep reinforcement learning agent is developed to perform the task of generating texts and arranging texts in the right position, with higher empathy, fluency, context specificity and diversity being rewarded. Empathy has also been analyzed in condolence under the appraisal theory [13] where it is measured by how a person who appraises in the same way as the person who is experiencing a situation in six different dimensions proposed by Smith and Ellyworth [10] (more details in Section 3). Random forest regressor with unigrams and bigrams, and RoBERTa [7] model are used and compared to predict empathy levels.

Authors' addresses: Jiamin Yang, College of Engineering, University of Michigan, Ann Arbor, US, jiaminy@umich.edu; David Jurgens (Advisor), School of Information, University of Michigan, Ann Arbor, US, jurgens@umich.edu.

3 THE APPRAISAL THEORY

We feel emotions when we are experiencing things, but interestingly, the generation of emotions is not restricted to our own experience. We sometimes may feel emotions when something happens to someone else. For example, we may feel nervous for our friends when they are attending some extremely important event, even though we are not part of it. In this case, if our friends also feel nervous, then we are experiencing the same emotion as our friends, which is defined as *empathy*.

In psychology, we refer to the person experiencing the event or in a certain situation as *target*, and the person who is in communication with the *target* and is responding as *observer*.

Several theories have been developed to explain how and why empathy occurs. Hoffman [3] proposed five mechanisms to explain why the observer can feel distress as the target does: mimicry, classical conditioning, direct association, mediated association, and role-taking. Other theories explain empathy from the perspective of neuroscience. Mirror neurons were first discovered in 1992 [2] that were observed to discharge when a monkey was grasping food and seeing people grasping food, which relates to empathy where the same mechanism is triggered when concerning oneself and the others. Perception-action model of empathy [8] shares a similar idea but states that episodic memories, automatic arousal or other representations can also generate our emotions for others.

Those theories help explain empathy where we feel the same as another person when he/she is experiencing something. However, they are insufficient to explain why we can also feel something different from another person when he/she is experiencing something. For instance, we feel embarrassed for someone, whereas he/she may not be aware of the situation [12].

So Wondra and Ellsworth [12] introduced an appraisal theory of empathy based on the appraisal theories of emotion. The appraisal theories of emotion state that emotions are based on how we evaluate, interpret or appraise situations. Smith and Ellsworth found six dimensions that people use to appraise situations [10]:

- *Pleasantness*: How pleasant the situation was.
- *Anticipated effort*: How much effort was needed to deal with the situation.
- *Situational control*: How much the situation was out of anyone's control.
- *Self-other agency*: How much oneself or another person was responsible for the situation.
- *Attentional activity*: How much their attention was drawn to the situation rather than diverted away from the situation, which is akin to the appraisal of novelty in other appraisal models.
- *Certainty*: How certain about what was happening in the situation or what would happen next.

So the appraisal theory of empathy is stated as "empathy occurs ... when the observer appraises the target's situation in the same way that the target appraises it" [12]. Our project will refer to this theory and try to find the alignment between target and observer, where they appraise the same situation with the same dimension.

4 METHODS

4.1 Data

The data we used are Reddit posts and comments data from 2016 to 2021. The data collection procedure contains two stages: data filtering and data annotation. In data filtering, we aim to gather paired data where the target contains texts of high distress and the corresponding observer contains texts of high condolence. In data annotation, we label both target and observer with appraisals and the alignment between them two. More details will be described in the following subsections.

4.1.1 Data Filtering. Due to the large amount of data provided by Reddit API, we selected 35 subreddits that contain the comments of high distress or condolence or both [13] (e.g. 'r/anxiety', 'r/depression'). A full list of subreddits is presented in Appendix A. We then trained Distress and Condolence classifiers separately using `Microsoft/MiniLM-L12-H384-uncased` [11] for filtering target texts of high distress and observer texts of high condolence with both thresholds set as 0.9. After pairing the target and the observer text, an Empathy Classifier [13] is used to further extract post-comment or comment-comment pairs that have high empathy with a threshold set as 2.

4.1.2 Data Annotation. After gathering the paired data, we start the process of annotation. The labels we used contain the six dimensions of appraisals introduced in Section 3 with 2 extra labels: *Advice* and *Objective Experience*. In other words, we used *Advice*, *Anticipated Effort*, *Attentional Activity*, *Certainty*, *Objective Experience*, *Pleasantness*, *Self-Other Agency*, and *Situational Control* to split the texts in both target and observer to spans where each span contains one perspective on how target/observer appraises the situation. For the same label in target and observer (e.g. span labeled as *Pleasantness* in target and span labeled as *Pleasantness* in observer), we also determine whether they are aligned or not by evaluating whether the observer implies the same emotion as the target did on the same issue or situation. The final label is determined by the agreement of at least three annotators.
Due to the complexity and ambiguity of language, we developed a codebook for guiding the annotation process. Some general rules applied to all labels are listed as follows:

- Favor explicit appraisals over implicit or coreferenced appraisals.
- Favor longer spans i.e. label as much of the sentence as possible provided that part contains an appraisal.
- If there are multiple appraisals in the same sentence, label the subparts with the appraisal that dominates.

Next, we provide more specific rules for each label.

Advice. Include expressions where the target asks for advice or the observer provides advice. e.g. "If you don't mind, can you point out which traits sound like BP and what difference you see when your SO is on his meds? I'm hopeful and would like to find out what the medication could help with."

Anticipated Effort. Include expressions concerning efforts such as general expressions on efforts (e.g. need to do something), current or future efforts needed to deal with the situation (e.g. "I'm in therapy, I've been there for at least 7 months"), lacking effort ("I want a way out of the interminable grief and there simply isn't one apart from death") and struggles relating to effort (e.g. procrastination).

Attentional Activity. Nearly equivalent to novelty. Concern more with the occurrence of an event such as suddenness, familiarity, and predictability, rather than the qualities of the event like pleasantness. e.g. "I am not typically an emotional person. I have only really and truly cried 3 times in my adult life, this being the 3rd and by far the worst."

Certainty. Include expressions concerning how certain the target/observer thinks the situation would have happened in the past, or happens at present, or will happen in the future. e.g. "In retrospect, maybe should have broke it off earlier..."

Objective Experience. Include expressions describing the experience of the author that is not an appraisal or the broader context/circumstances in which their story takes place. e.g. "I get good grades, I have lots of friends, I play sports and people here I think like me too!"

Pleasantness. Include expressions with sentimental words. e.g. "I still feel so...numb, and then heartbroken, and then hopeful, and then crushed, ashamed, afraid"

Self-Other Agency. Include expressions describing who should be responsible for a situation. e.g. "The little shit dropped me as a patient and now I can't seem to find a doctor who is willing to prescribe it."

Situational Control. Include expressions concerning how the situation is within or out of control, and the case where people know what to do but are forced to do it another way. e.g. "I feel like we need to just hurry up and move so we can have some space, privacy, nature, quiet, affordable childcare, and I can get back to my writing and my art. But I have to stick it out for another 12 months."

To effectively compare the annotations between different annotators, a span annotation comparison tool is developed specifically for this task. Figure 1 shows the main page of the tool. The left-hand side of the page contains the information of the current text to compare, including the full text and labels that are used. The right-hand side is the main section for showing annotations. For easier comparison, the full text is split into several segments where each time only one segment is shown on the page. Segments can be navigated through top-right arrows. The final annotation is shown on top, followed by annotations from different annotators. The last block is for submitting the final annotations. Each time, a new submission will overwrite the previous one, and will be instantly reflected in the final annotation block at the top.

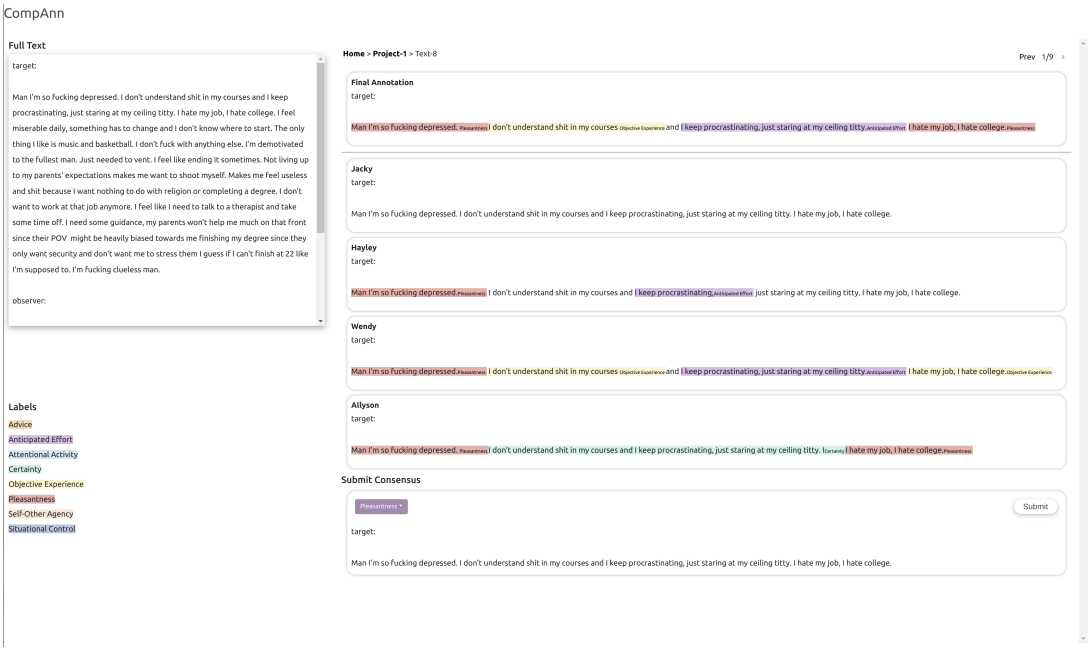

Fig. 1. Main page of Annotation Comparison tool.

Currently, the tool supports the following functions:

- Show full text and labels used.
- Visualize span annotations of each annotator.
- Support labeling and submit final annotations agreed by annotators.
- Navigate different segments for one text.
- Navigate different texts for one project with pagination.

4.2 Modeling

In this section, we present the model designed for detecting the alignment between target and observer. As we are still actively collecting a large enough golden dataset, the model hasn't entered the stage of training. We'll use SpanBERT [4] to recognize each label from both target and observer text. Therefore, there will be a total of 8 SpanBERTs with each tuned for one specific label. The input to SpanBERT is one-hot encoded, where the position of a character is marked as 1 if it is part of the label, and 0 otherwise. The output will have the same format as the input where 1 indicates that the character is predicted as part of the label. The loss function is set to take into consideration of the classification performance of each character and the length of each span so that longer spans are preferred over shorter ones [6]. Then we feed the spans of the same dimension (*Pleasantness* in target is compared with *Pleasantness* in observer, but

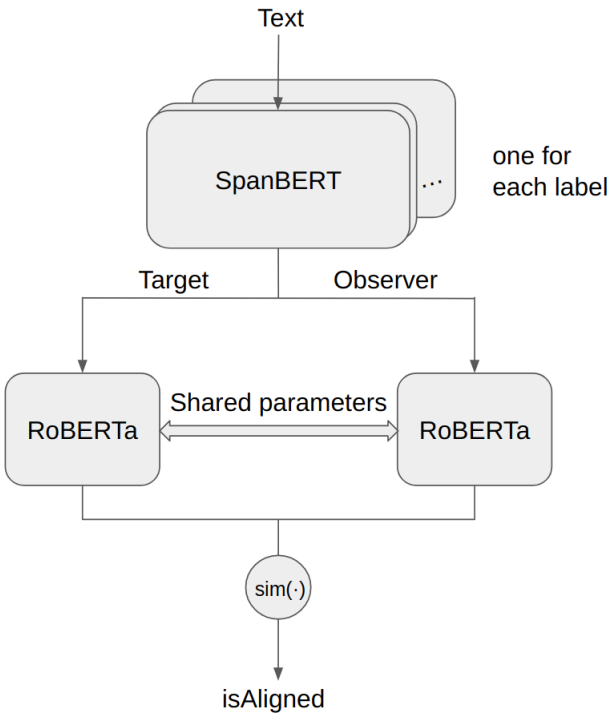

Fig. 2. Model Architecture for detecting spans in target-observer texts using SpanBERT for each label, and comparing alignment between two appraisals of the same dimension using Siamese network with RoBERTa as the main block.

not other dimensions such as *Anticipated effort* in observer) from target and observer to two RoBERTa [7] models to find the alignment. The two RoBERTa models will share the exact same parameters, forming the structure of Siamese Network [1]. By calculating the similarity between those two outputs and setting a threshold, we can predict if two appraisals of the same dimension, one from the target and one from the observer, are aligned. The loss function will be set as cross entropy loss in order to penalize misclassification.

5 RESULTS

5.1 Distress and Condolence Classifier Performance

Table 2 shows the accuracy and F1 score for the Distress and Condolence classifier that are trained for filtering Reddit data mentioned in Section 4.1.1.

Classifier	Accuracy	F1
Distress	0.707	0.692
Condolence	0.799	0.790

Table 1. Distress and condolence classifier performance.

5.2 Data Collected

Currently, we have approximately 30M post-comment pairs data for one month (filtered by Distress and Condolence classifiers), 1700+ comment-comment pairs data for 2019 (filtered by Distress, Condolence and Empathy classifiers), and 246 comment-comment pairs from 2016 to 2017 [13] with annotations.

5.3 Dataset Example

Below, we provide a labeled target text according to the annotation rules.

I'm trying to get divorced and I'm finding it almost impossible.
Anticipated Effort
I have a prime opportunity to open up a checking and savings account in my name only,Objective Experience because my bank just shut down, and we're/I'm forced to create a new account. Situational Control I had work issue me a paper check instead of direct deposit. Objective Experience

Yet the thought of doing so makes me sick with anxiety and guilt. I feel duplicitous. Pleasantness I feel like I'm betraying my husband. Pleasantness

Every aspect of doing this just feels impossible. Situational Control Everything feels so wrong, I feel like everything I'm doing is so awful that I shouldn't be doing it. Pleasantness Thus I have very little impetuous to move forward on anything Anticipated Effort because nothing feels right at all.Pleasantness

Meanwhile I haven't been touched at all in a year in a half,Objective Experience and I'm starting to feel agonizingly physically sick over that too.Pleasantness I fantasize about paying someone just to hold me, stroke my hair, cuddle with me, and caress my skin,Anticipated Effort that's how bad it is.Pleasantness

Ugh. Pleasantness So really not left and leaving. Certainty I'm not getting anywhere. Certainty

Only good thing is I have a job where I work every holiday, and holidays when I was a kid were a nightmare, so I don't feel like I'm missing anything at all. Objective Experience I'm not a Grinch, I like decorating, and I LOVE winter and snow, but ultimately I'm not emotionally invested in the holiday season at all.
Objective Experience

6 DISCUSSION

The annotation appears to be a much more complicated task than expected due to the complexity of languages. For example, when giving the following two small paragraphs:

> " The reason we broke up wasn't because we don't like each other (quite the contrary! I still keep in touch) but because we both know each other well enough to know that *we just wouldn't want to do long distance*."
>
> "The reason we broke up wasn't because we don't like each other (quite the contrary! I still keep in touch) but because we both know each other well enough to know that *it's hard to do long distance*."

They are exactly the same except for the last sentence and they end up with two different labels with the former one being *Situational Control* and the latter one being *Anticipated Effort* because they tend to have a different focus when phrasing. Saying "we just wouldn't want to do long distance" shows more of the control side of a situation where the target expresses his/her willingness to do something. But saying "it's hard to do long distance" will focus more on the difficulty of dealing with the current situation.

Other complexities include:

- *Implicit expression*. "My cat died yesterday I get this." We can infer sadness from the text but there are no explicit sentimental words, which can cause confusion between *Pleasantness* and *Objective Experience*.
- *Ambiguity*. "Reason I'm even telling you guys this is because while I tend to keep most of my real life to myself, this is just something that is too much for me to bottle up I think." "Too much" in the text can mean either the situation is too much for the target to take or the effort needed to deal with the situation is too much. This causes ambiguity between situational control and anticipated effort.
- *Multi-Appraisals*. "I was really concerned that I'm just being selfish and I obviously don't want to hurt them by leaving them." "Selfish" tends to be a word that both indicates appraisal of pleasantness and self-other agency.

Those cases can be solved based on general rules described in Section 4.1.2. Those cases reveal the variety in human language and are important to be discussed and taken into consideration when doing computational analysis.

For future work, we will expand the current dataset and improve the annotation codebook for clearer guidance, try on heuristic methods for labeling datasets, and train and improve the designed model to check the performance and prepare for future analysis.

REFERENCES

[1] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems* 6 (1993). https://proceedings.neurips.cc/paper/1993/file/288cc0ff022877bd3df94bc9360b9c5d-Paper.pdf

[2] G. di Pellegrino, L. Fadiga, L. Fogassi, and et al. 1992. Understanding motor events: a neurophysiological study. *Exp Brain Res* 91 (1992), 176–180. https://doi.org/10.1007/BF00230027

[3] Martin L. Hoffman. 2000. *Empathy and Moral Development: Implications for Caring and Justice.* Cambridge University Press. https://doi.org/10.1017/CBO9780511805851

[4] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *CoRR* abs/1907.10529 (2019). arXiv:1907.10529 http://arxiv.org/abs/1907.10529

[5] 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).* Asian Federation of Natural Language Processing, Taipei, Taiwan, 246–251. https://aclanthology.org/I17-2042

[6] Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing Neural Predictions. *CoRR* abs/1606.04155 (2016). arXiv:1606.04155 http://arxiv.org/abs/1606.04155

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[8] Stephanie D. Preston and Frans B. M. de Waal. 2002. Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences* 25, 1 (2002), 1–20. https://doi.org/10.1017/S0140525X02000018

[9] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach. *CoRR* abs/2101.07714 (2021). arXiv:2101.07714 https://arxiv.org/abs/2101.07714

[10] Craig A Smith and Phoebe C Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology* 48, 4 (1985), 813.

[11] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. arXiv:2002.10957 [cs.CL]

[12] Joshua Daniel Wondra and Phoebe C. Ellsworth. 2015. An appraisal theory of empathy and other vicarious emotional experiences. *Psychological review* 122 3 (2015). https://doi.org/10.1037/a0039252

[13] 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).* Association for Computational Linguistics, Online, 609–626. https://doi.org/10.18653/v1/2020.emnlp-main.45

A SUBREDDIT CATEGORIES

Here, we list the subreddits that are chosen for building the dataset described in Section 4.1:

anxiety, depression, Miscarriage, domesticviolence, widowers, GriefSupport, Petloss, FiftyFifty, SuicideBereavement, ttcafterloss, heartbreak, BreakUps, BreakUp, BipolarSOs, dementia, Alzheimers, ExNoContact, CautiousBB, domesticviolence, CaregiverSupport, abusiverelationships, emotionalabuse, marriageadvice, lastimages, PrayerRequests, OldManDog, seniorkitties, askfuneraldirectors, death, dogpictures, MadeMeCry, cancer, MomForAMinute, sad, happycryingdads.

B DISTRESS AND CONDOLENCE CLASSIFIER HYPERPARAMETERS

Both Distress and condolence classifier are trained using the following hyperparameters:

Model	microsoft/MiniLM-L12-H384-uncased
learning rate	4e-5
number of training epochs	50
training batch size	16

Table 2. Distress and condolence classifier hyperparameters.