CAISA at WASSA 2023 Shared Task: Domain Transfer for Empathy, Distress, and Personality Prediction

Fabio Gruschka and **Allison Lahnala** and **Charles Welch** and **Lucie Flek**
Conversational AI and Social Analytics (CAISA) Lab
Department of Mathematics and Computer Science, University of Marburg
http://caisa-lab.github.io

Abstract

This research contributes to the task of predicting empathy and personality traits within dialogue, an important aspect of natural language processing, as part of our experimental work for the WASSA 2023 Empathy and Emotion Shared Task. For predicting empathy, emotion polarity, and emotion intensity on turns within a dialogue, we employ adapters trained on social media interactions labeled with empathy ratings in a stacked composition with the target task adapters. Furthermore, we embed demographic information to predict Interpersonal Reactivity Index (IRI) subscales and Big Five Personality Traits utilizing BERT-based models. The results from our study provide valuable insights, contributing to advancements in understanding human behavior and interaction through text. Our team ranked 2nd on the personality and empathy prediction tasks, 4th on the interpersonal reactivity index, and 6th on the conversational task.

1 Introduction

Empathy, a fundamental component of interpersonal communication, emerges in broad spectrum of conversational and discourse settings, ranging from informal dialogues to purpose-driven conversations. With an invaluable role for cooperative interactions, modeling empathetic language is a growing area in natural language processing (NLP) research, enabling the improvement of dialogue agent experiences, analysis of online supportive interactions, and the development of educational tools (Rashkin et al., 2019; Lin et al., 2020; Majumder et al., 2020). Despite its significance, predicting empathy poses an immense challenge due to the scarcity of resources and the complexities involved in establishing a gold standard for this nuanced phenomenon (Omitaomu et al., 2022).

A major hindrance to empathy prediction in the NLP field is the lack of accessible, high-quality datasets. Many studies are conducted on sensitive data, which cannot be disclosed publicly (Lahnala et al., 2022b). Though some publicly available datasets exist, built on social media platforms or through specific data collection tasks, they are sparse and each comes with its inherent limitations due to the challenges in the data collection and annotation process (Omitaomu et al., 2022).

Defining empathy in a concrete, measurable way for consistent and relevant gold standard annotations is another formidable challenge. Empathy definitions vary greatly across psychological research, and NLP datasets are often annotated by third parties, rather than the individuals experiencing or receiving empathy. This approach captures specific language aspects identified by an external observer but fails to provide insight into how particular empathetic experiences influence language (Buechel et al., 2018).

In parallel to empathy, personality traits are fundamental to interpersonal relationships and social interactions. The Big Five Personality model (PER) (McCrae and Costa, 1992), a widely studied framework, is instrumental in understanding human social behavior. A comprehensive understanding of human behavior can be achieved by combining empathy, gauged by the Interpersonal Reactivity Index (IRI) (Davis, 1980), and personality traits.

This paper presents our experimental work on the WASSA 2023 Empathy and Emotion Shared Task (Barriere et al., 2023). We predict perceived empathy, emotion polarity, and emotion intensity at the turn-level in a conversation (the CONV task), and empathy and distress at the essay level (the EMP task). We explore the use of adapters, which provide greater efficiency compared to complete fine-tuning, and an adjusted approach derived by Sharma et al. (2020) at the essay level. Additionally, we embed demographic information to predict IRI subscales (Perspective Taking, Personal Distress, Fantasy, and Empathic Concern) and the Big

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 553–557
July 14, 2023 ©2023 Association for Computational Linguistics

Five Personality Traits (Conscientiousness, Openness, Extraversion, Agreeableness, and Stability) using BERT-based models (Devlin et al., 2019).

The paper is structured as follows: Section 2 reviews the task and dataset, provides an overview of the IRI, PER constructs, and empathy in conversations, emphasizing their significance in understanding human behavior. Section 3 describes our implementation. Section 4 presents the results and discussion of our findings. Finally, Section 5 concludes the paper and proposes directions for future research.

2 Task and Dataset

In this section, we describe the dataset and tasks employed in our research aimed at predicting empathy, distress, personality traits, and IRI subscales using the dataset provided by Omitaomu et al. (2022).

The dataset utilized in our experiments comprises empathetic reactions captured in essays and conversation responses to news articles involving harm to a person, group, or other entities. These reactions are documented in essays that range between 300 and 800 characters in length, as well as in conversations that consist of an average of 23 speech turns. The dataset also includes the original news articles and demographic information at the person-level, such as age, gender, ethnicity, income, and education level.

Each essay in the dataset is supplemented with Batson et al. (1987)'s empathic concern and personal distress scores, providing an insight into the individual's empathetic response after engaging with the news article. In addition, the dataset provides McCrae and Costa (1992)'s Big Five personality scores and Inter-Personal Reactivity Index (IRI) scores for the respective user, further enhancing our understanding of their empathetic capacity.

The Big Five of Costa and McCrae (1992) was shown to predict many traits about people, their behavior, and relationships. Each dimension can be rated on a continuous scale, where a person has more or a lesser degree of the qualities associated with that dimension. The following facets are from John et al. (1999):

1. **Neuroticism** - Anxiety, angry hostility, depression, self-consciousness, impulsiveness, vulnerability

2. **Extraversion** - Warmth, gregariousness, assertiveness, activity, excitement seeking, positive emotions

3. **Openness** - Fantasy, aesthetics, feelings, actions, ideas, values

4. **Agreeableness** - Trust, straightforwardness, altruism, compliance, modesty, tendermindedness

5. **Conscientiousness** Competence, order, dutifulness, achievement striving, self-discipline, deliberation

The IRI index is discussed in Davis (1983), who constructed a 28-item survey to measure four aspects of empathy using a 5-point Likert scale. The items (directly taken from the paper) are as follows:

1. **Perspective Taking** – the tendency to spontaneously adopt the psychological point of view of others.

2. **Fantasy** – taps respondents' tendencies to transpose themselves imaginatively into the feelings and actions of fictitious characters in books, movies, and plays.

3. **Empathic Concern** – assesses "other-oriented" feelings of sympathy and concern for unfortunate others.

4. **Personal Distress** – measures "self-oriented" feelings of personal anxiety and unease in tense interpersonal settings.

In the case of conversations, each speech turn has been annotated by a third person for perceived empathy, emotion polarity, and emotion intensity. This offers a comprehensive view of the interaction, enabling a detailed examination of the empathetic exchanges within the conversation.

The tasks of our research involve predicting empathy and emotion polarity & intensity on conversational turns (CONV), empathy and distress scores (EMP), personality traits (PER), and IRI subscale values (IRI). Systems are evaluated based on the Pearson's r correlation between the predicted and actual values in a test set, similar to the approach adopted in the previous edition of the shared task (Barriere et al., 2022).

3 System Description

3.1 Essay-Level Prediction

Domain Adapted Model. In our approach to predict empathy and distress at the essay-level, we adapt the classification model proposed by Sharma et al. (2020) to a regression model. Their original model, was designed for empathy classification, while our goal is to predict empathy and distress in essay texts. To achieve this, we make several modifications to the model, allowing it to handle essay-level predictions.

First, we normalize the labels for empathy and distress scores in the range of 0 to 1. This transformation enables the model to predict continuous values rather than categorical labels.

Next, we modify the model's architecture to accommodate the regression task. We replace the classification layer with a regression layer, which predicts continuous values instead of class probabilities. To train the modified model, we use the mean squared error (MSE) loss function, which measures the average squared differences between the predicted and true empathy and distress scores.

Finally, we fine-tune the adapted model on the Omitaomu et al. (2022) datasets, which contains essay texts along with their corresponding empathy and distress scores. The model learns to predict empathy and distress scores by leveraging the pretrained model's understanding of natural language and adjusting its weights based on the specific context of empathy and distress in essays.

Demographic Embeddings. The demographic embedding layer takes in one-hot encoded demographic information (i.e. gender, education, race, age) as well as income as a single number and concatenates it to the text encoding during the forward pass. This allows the model to utilize demographic features for each individual, which is particularly beneficial for tasks requiring personalized predictions. The demographic embedding layer is initialized using Xavier initialization and is updated during training (Glorot and Bengio, 2010) and has a dimension of 135. The BERT encodings of the article and essay are projected down to the same dimension before being concatenated and passed to the final classification layer.

3.2 Conversation-Level Emotion Prediction

Adapter-Tuning Framework. For our implementation, we use AdapterHub (Pfeiffer et al., 2020), a straightforward framework built on the HuggingFace transformers. We train adapters to predict a conversation's emotional polarity, emotional intensity, and empathy.

We employ a method inspired by the EPITOMEFUSION method implemented by Lahnala et al. (2022a). We used the model of Sharma et al. (2020), which is based on RoBERTa (Liu et al., 2019) to predict empathetic reactions, explorations, and interpretations. We finetuned separate adapters to categorize each of these aspects in the EPITOME dataset. Later, these adapters are merged using the AdapterFusion composition technique (Pfeiffer et al., 2021). An adapter for the prediction of empathy and distress in conversation was trained on top of this, with learning rate $1e^{-4}$, for 10 epochs. This configuration allows the combination of the knowledge from each of the pretrained adapters for the EPITOME tasks and their application in the conversation-level prediction tasks.

Attribute	Pearson Correlation
Personality (PER)	
Conscientiousness	0.3229
Openness	0.3273
Extraversion	-0.1966
Agreeableness	0.2900
Stability	0.1999
Interpersonal Reactivity Index (IRI)	
Perspective Taking	0.1582
Personal Distress	-0.1875
Fantasy	-0.0556
Empathic Concern	0.1804
Overall	0.0239
Empathy Prediction (EMP)	
Empathic Concern	0.3478
Personal Distress	0.4197
Overall	0.3840
Empathy & Emotion in Conversations (CONV)	
Emotion Polarity	0.7832
Emotion Intensity	0.6858
Empathy	0.6523
Overall	0.7071

Table 1: Pearson correlations for the personality, IRI, and empathy prediction tasks post-phase essay-level results with the task embedding model.

4 Results and Discussion

Our submissions to the post-evaluation phase on the test dataset have yielded promising results, as showcased in Table 1.

In the domain of personality prediction, as displayed in the top portion of Table 1, our task embedding model has performed particularly well in predicting certain aspects of the Big Five Personality traits. Specifically, it has demonstrated strong predictive power for the traits of Conscientiousness (r=0.3229), Openness (r=0.3273), and Agreeableness (r=0.29). However, the model has shown a negative correlation for Extraversion (r=-0.1966), indicating a need for further refinement in this area.

Turning to the Interpersonal Reactivity Index (IRI) prediction, as seen in the middle of Table 1, the performance of our model has been more varied. While the prediction of the Perspective Taking subscale showed a modest positive correlation (r=0.1582), Personal Distress exhibited a negative correlation (r=-0.1875). This might suggest that the model currently struggles with accurately capturing the nuances of distress experienced by individuals. The model also demonstrated a low correlation for the Fantasy (r=-0.0556) subscale, though our best performance was on the Empathic Concern (r=0.1804) subscale.

At the essay level, our approaches have shown encouraging results for empathy and distress prediction, as evidenced by the second to last portion of Table 1. The Domain Adapted (Sharma) approach, in particular, has excelled in this task, yielding an average Pearson correlation of 0.3478 for Empathy and a notable 0.4197 for Distress. These results underline the efficacy of this approach in gauging empathy and distress from written texts.

Lastly, as we move to the conversation level prediction in bottom of Table 1, our adapter approach has demonstrated satisfactory performance. The model has been particularly successful in predicting emotional polarity (r=0.7832), emotional intensity (r=0.6858), and empathy (r=0.6523) in the conversation. These results affirm the potential of our adapter approach in effectively capturing the empathetic and emotional dynamics within conversational exchanges. We believe there is much room for improvement at the conversation level and through the use of adapters. Our model was relatively simple and future work should explore other adapters and architectures to more effectively transfer knowledge from related tasks. We only took individual turns in the conversation into account and future work would benefit from providing the model with additional context from the conversation history.

5 Conclusion

In this paper, we have presented our methodologies and findings from predicting a range of empathy-related features in text, specifically in essays and conversation responses from the Omitaomu et al. (2022) dataset.

We developed and evaluated a series of models, each addressing unique aspects of the prediction tasks. At the essay level, we employed a domain-adapted model based on the work of Sharma et al. (2020), modified to perform regression instead of classification, effectively predicting empathy and distress scores.

Our results across different measures are encouraging. The demographic-embedding approach performed quite well in predicting the conscientiousness, openness, and agreeableness aspects of the Big Five Personality traits. In contrast, the performance on the Interpersonal Reactivity Indices was less impressive. The domain-adapted model excelled in predicting empathy and distress at the essay level. At the conversation level, our adapter approach achieved satisfactory results in predicting emotional polarity, emotional intensity, and empathy.

Code for our systems and experiments are publicly available at `https://github.com/caisa-lab/wassa-shared-task-2023`.

References

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. Wassa 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Salvatore Giorgi. 2023. Wassa 2023 shared task: Predicting empathy, emotion and personality in interactions and reaction to news stories. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*.

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.

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.

Paul T Costa and Robert R McCrae. 1992. Neo pi-r professional manual: Revised neo personality inventory (neo pi-r) and neo five-factor inventory (neo-ffi). *Odessa, FL: Psychological Assessment Resources*.

Mark H Davis. 1980. A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, 10:85.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Allison Lahnala, Charles Welch, and Lucie Flek. 2022a. CAISA at WASSA 2022: Adapter-tuning for empathy prediction. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 280–285, Dublin, Ireland. Association for Computational Linguistics.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022b. 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.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An empathetic neural chatbot.

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. *arXiv preprint arXiv:1907.11692*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Robert R McCrae and Paul T Costa. 1992. The five-factor model of personality and its relevance to personality disorders. *Journal of Personality Disorders*, 6(4):343–359.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. 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.

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.