CAISA at WASSA 2022: Adapter-Tuning for Empathy Prediction

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
{allison.lahnala,welchc,lucie.flek}@uni-marburg.de

Abstract

We build a system that leverages adapters, a light weight and efficient method for leveraging large language models to perform the task Empathy and Distress prediction tasks for WASSA 2022. In our experiments, we find that stacking our empathy and distress adapters on a pre-trained emotion classification adapter performs best compared to full fine-tuning approaches and emotion feature concatenation. We make our experimental code publicly available.[1]

1 Introduction

Empathy is an important interpersonal function in communication settings from conversations between friends and family, to educational, medical, or other goal-oriented dialogues. In natural language processing research, automatic empathy recognition and generation are explored for motivations such as improved experiences with open-domain dialogue agents (Rashkin et al., 2019; Majumder et al., 2020; Lin et al., 2020), analyzing supportive interactions in online forums (Zhou and Jurgens, 2020; Sharma et al., 2020; Lahnala et al., 2021), and for the development of educational and evaluative tools for counselor training (Gibson et al., 2015; Pérez-Rosas et al., 2017; Zhong et al., 2020) in addition to other educational domains (Wambsganss et al., 2021). Yet empathy prediction is a challenge for current language technologies due to resource availability and difficulty defining a gold standard for the complex phenomenon.

The lack of proper resources for empathy modeling limits the ability of the NLP community to more widely explore it. Many studies, for instance, are on sensitive data that cannot be made public. There are some datasets that are publicly available that are built on social media platforms, or through specific data collection tasks, however, these are few and far between, and each have limitations due to inherent challenges in the collection and annotation process.

A general challenge with studying empathy is how to define the concept concretely enough to obtain consistent and relevant gold standard annotations, as there are many highly varied definitions in psychology research (Cuff et al., 2016). Furthermore, empathy datasets in NLP are almost always annotated by others rather than the person having an empathetic experience (Buechel et al., 2018) or the person on the receiving end. Such annotations thus indicate particular aspects of language identified by a removed observer, rather than provide insight into the effect that particular empathetic experiences have on language.

Toward this issue, Buechel et al. (2018) developed the EMPATHETICREACTIONS dataset, which contains empathic concern and personal distress ratings based on self-evaluations of individuals' own empathetic experiences at the time of writing the text. These reactions are short essays in which the author describes their feelings as they would to a friend after reading an article meant to evoke empathy. This data may then enable analysis into the way the empathetic experiences impact or relate to produced language. The EMPATHETICREACTIONS dataset is used for predicting empathy and distress in the WASSA 2022 Shared Task, enabling a large group of people to research empathy prediction on a standard and public set of data.

In this paper, we present our experiments for empathy and distress prediction as part of WASSA 2022. We explore adapters for the task since it is more efficient than full fine-tuning, which so far has not been explored for empathy prediction. Following work on domain transfer, we also build a system leveraging additional empathy data, as the amount of empathy data is still sparse.

[1] https://github.com/caisa-lab/wassa-empathy-adapters

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 280 - 285

2 Background

The ability to recognize empathy in text is important for advancing language technologies from dialogue agents to computational social science tools. As such, there is a growing body of research on automatic empathy recognition. Many studies concern highly sensitive and important scenarios such as counseling and medical dialogues (Sharma et al., 2020) or are crisis-related (Zhang et al., 2020) but such resources are protected and cannot be made public. However, there are a number of recently proposed empathy datasets available to the public, which are consolidated by means such as collecting and labeling social media (Sharma et al., 2020; Zhou and Jurgens, 2020), or through collection tasks (Rashkin et al., 2019; Buechel et al., 2018).

Annotating empathy involves a host of challenges. Most datasets are annotated by someone who did not take part in the writing or conversation, requiring them to interpret how the author felt, rather than acquiring this information from the authors directly. Also, there are various definitions of empathy across fields. Generally, NLP has considered *emotional empathy*, despite the prevalence of other components of empathy in psychology (Cuff et al., 2016). There have been valuable efforts to build resources for empathy identification, each operating upon different perspectives of empathy.

Sharma et al. (2020)'s EPITOME dataset, contains support-seeker and responder post pairs from Reddit and has multi-faceted empathy labels on the responder posts. The responder posts are annotated with the degree of three different aspects of empathy (interpretations, emotional reactions, and explorations), 0 for absent, 1 for weak, and 2 for strong. As this scheme contains distinct labels for both emotional and cognitive aspects of empathy, this dataset is a valuable resource for pursuing empathetic modeling beyond emotional aspects.

Zhou and Jurgens (2020) introduced a dataset post-response pairs from Reddit where the post contains an expression of distress and the response is a condolence. While the final dataset contained one empathy score, the annotation process was strictly guided by a multi-faceted definition of empathy, the *appraisal theory* (Lamm et al., 2007; Wondra and Ellsworth, 2015). Under this definition, the degree of empathy is how closely the responder's appraisal of another person's situation matches the person's appraisal of their own situation.

Rashkin et al. (2019)'s EMPATHETICDIALOGUES dataset contains conversations grounded in one of 32 emotions. During data collection, participants were instructed to converse with each other. Dialogues contain emotion labels but not empathy labels. Welivita and Pu (2020) further annotated empathetic intents in this dataset.

Buechel et al. (2018) built the EMPATHETICREACTIONS dataset based on Batson's Empathic Concern – Personal Distress Scale (Batson et al., 1987). Under this view, there are two aspects of empathetic reactions, the level a personal distress experienced by the reactor ("suffering with something") and the level of empathy ("feeling for someone") while maintaining self-other separation. Here, empathy involves emotional feelings such as compassion, warmth, and tenderness, whereas distress involves those such as worry, alarm, and grief.

These datasets may differ stylistically due to their different domains. Having this diversity is valuable so that we can study how empathetic communication may vary across contexts. However, as the volume of data across these datasets is still limited, it is important to understand if they can be leveraged together despite their differences.

3 Task and Dataset

This paper describes our system submitted for Track 1 of the WASSA 2022 task which concerns empathy and distress prediction in Buechel et al. (2018)'s dataset of empathic reactions to news stories. Empathetic reactions are captured in essays written by people who were asked to read an article that involves a harmful situation a write a response. Participants were asked to rate their empathy after reading an article before writing their response. These ratings were self-measured using Batson's Empathic Concern - Personal Distress Scale (Batson et al., 1987), which contains multiple items that were averaged in order to obtain the gold ratings for empathy and distress.

The task of Track 1 of WASSA 2022 was to predict the numerical values for empathy and distress on a continuous scale for the essays. Systems were evaluated by Pearson's r correlation between the predictions and the actual values in a test set. WASSA provided an extension of the dataset to include the original news articles, demographics (age, gender, ethnicity, income, education level) and personality information. The extension also included emotion labels obtained using pretrained emotion detection models.

4 System Description

Adapters offer a lightweight tuning strategy alternative to full fine-tuning (Houlsby et al., 2019). With adapter-tuning, new initialized layers are inserted at each layer of the original pretrained network, and the new weights are fine-tuned while the original network's weights remain fixed. Adapters have been shown to effectively perform at near state-of-the-art levels while drastically improving efficiency (Houlsby et al., 2019; Pfeiffer et al., 2020b, 2021).

As reported by the WASSA 2021 task (Tafreshi et al., 2021), the most robust systems for empathy and distress modeling involved fine-tuning of transformer models such as RoBERTA (Liu et al., 2019) and ELECTRA (Clark et al., 2020). In our experiments, we attempt an adapter tuning approach (Houlsby et al., 2019) motivated by their efficiency, and compare to full fine-tuning.

Furthermore, we experiment with leveraging a different empathy dataset, EPITOME (Sharma et al., 2020). This dataset contains support-seeker and responder posts on Reddit (as described in § 2).

Full fine-tuning. For our full fine-tuning approaches, we fine-tune RoBERTa using `roberta-base` from the HuggingFace library (Wolf et al., 2020) for separate models predicting the essay's empathy and distress scores. Our most basic model RoBERTa is trained only on the essay text.

The second model EmoRoBERTa is fine-tuned with emotional features, by leveraging the sentence-level emotion tags provided for the shared task, particularly the labels from the transformer model. For each essay, we concatenate each sentence's emotion tag to the sentence. We define these emotion tags as special tokens when tokenizing the text (e.g., [sadness]). We also include a separator token between each sentence after the emotion tag. To obtain these labels for the test dataset, we trained an adapter for roberta-base to predict these labels. This classifier attained 83.9, 83.8, and 80.2 for accuracy, weighted F1, and macro F1 respectively on the dev dataset.

For our final full fine-tuning approach EpitomeFT we leverage the Epitome dataset (Sharma et al., 2020) to obtain implicit empathy features from this other domain and labeling scheme. We fine-tune roberta-base to predict the level of empathy in the emotional reactions, explorations, and interpretations defined in their labeling scheme. The model we submitted for the test set was trained on the aspects consecutively.

Adapter-tuning. For our implementation we leverage AdapterHub (Pfeiffer et al., 2020a) which is a simple framework built on HuggingFace `transformers`. For our approach we train Tasks Adapters for a RoBERTa model to predict the empathy and distress scores for an essay.

EpitomeFusion: First we fine-tune three separate adapters to classify the degree of each of the three aspects of empathy in the EPITOME dataset. Then, we combine these adapters using AdapterFusion composition (Pfeiffer et al., 2021). This setup allows for combining the knowledge of each of the pre-trained adapters for the EPITOME tasks in order to leverage them in the WASSA empathy and distress prediction tasks. A classification head for the WASSA tasks is added on top of the fusion layer, and then trained.

EmotionStack: Following the procedure by Poth et al. (2021) to identify a similar adapters trained on a similar dataset, we identified a pre-trained emotion adapter available on AdapterHub.[2] This adapter was trained by Poth et al. (2021) on a dataset of English tweets (Saravia et al., 2018) with Ekman's six basic emotion labels (Eckman, 1972); the same emotion labels as in EmpatheticReactions dataset. Using this adapter is an alternative to using emotions explicitly labeled for the target dataset.

To leverage the knowledge of this pretrained adapter, we use the stacked composition setup presented by Pfeiffer et al. (2020b) (see Fig. 1 [3]), by stacking our task adapter, i.e. empathy or distress prediction, on the emotion adapter. The empathetic reaction essays are first input into the emotion adapter, and its output and residual are input to the empathy task adapter. Thus, the empathy task adapter is essentially obtaining predictions of Ekman's six emotions for the essays. While training the empathy adapter, the emotion adapter remains frozen.

5 Results and Discussion

Results from our submissions to the post-evaluation phase on the test dataset are presented in Table 1. The EmotionStack outperformed all other models on the test dataset on both empathy and distress

[2] https://huggingface.co/AdapterHub/roberta-base-pf-emotion

[3] https://docs.adapterhub.ml/adapter_composition.html

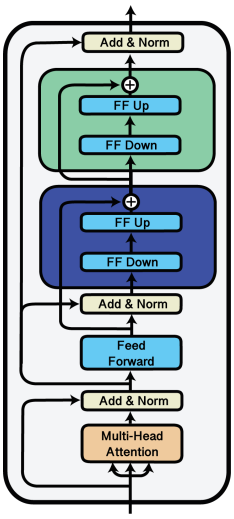

Figure 1: Stacked adapter composition.

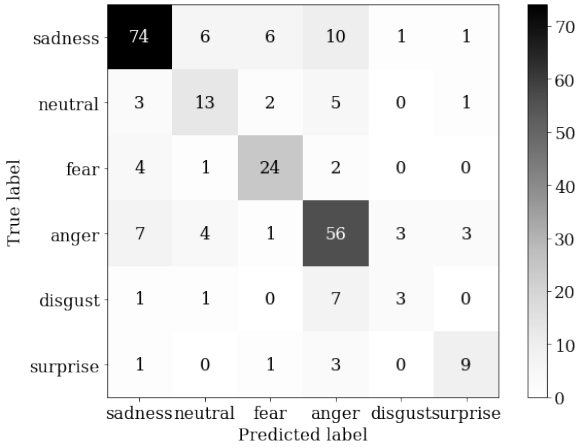

Figure 2: Confusion matrix of emotion predictions on dev dataset.

Model	Emp	Dis	Avg
EMOTIONSTACK	**0.524**	**0.521**	**0.523**
EPITOMEFUSION	0.472	0.496	0.484
ROBERTA	0.505	0.463	0.484
EMOROBERTA	0.478	0.493	0.486
EPITOMEFT	0.476	0.382	0.430

Table 1: Empathy and Distress prediction results on the test dataset.

detection. On average, the results of EPITOMEFUSION are comparable to the full fine-tuning approaches, namely ROBERTA and EMOROBERTA, by slightly outperforming on distress detection and underperforming on empathy prediction. EPITOMEFT performed worst on average due a particularly low score on distress prediction.

While we only explored the EMP track's tasks of empathy and distress prediction, the performance of the EMOTIONSTACK inspired us to submit predictions for the EMO track, predicting emotions. We used the same model, only changing the label set-up from predicting one value to predicting the six emotion categories–sadness, neutral, fear, anger, disgust, and surprise. This approach ranked highly with a macro F1-score of **0.604**. A confusion matrix for our classifier is shown in Figure 2.

The results of the adapter approach are exciting as it allievates the heaviness of full fine tuning. Adapters make it easy to leverage knowledge from other tasks learned on other datasets. In particular, we observe positive effects from using the pretrained emotion adapter on these tasks, which likely provides important emotional information relevant to empathic concern and personal distress.

However, we see no improvement from using the EPITOME data. Similarly, recent work found separate empathy types were found to have different effects on toxicity reduction (Lahnala et al., 2022). In preliminary experiments, we fine-tuned on only one of these aspects at a time, as we were interested in whether they have distinct effects and whether one or a combination of them is particularly well suited for our tasks. Further work is needed to definitively understand the effect of EPITOME and it's aspects on empathy and distress detection in the EMPATHETICREACTIONS. Given the sparsity of public empathy data, it is imperative for future work to better understand how the existing datasets can complement each other.

6 Conclusion

We presented our models for empathy and distress prediction on the EMPATHETICREACTIONS dataset for the WASSA 2022 shared task. We found that a stacked adapter composition with the WASSA task adapter stacked on a pre-trained emotion adapter (EMOTIONSTACK) outperformed other methods. This approach mitigates the costs of full fine-tuning while achieving comparable results. Furthermore, this method required no additional features beyond the empathetic reaction text. We further discussed challenges of researching empathy in natural language processing. In future work, we could explore incorporating the personal features provided for the shared task. We plan to further explore the use of different empathy datasets together for empathy prediction.

References

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.

Kevin Clark, Minh-Thang Luong, Quoc Le, and Christopher D Manning. 2020. Pre-training transformers as energy-based cloze models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 285–294.

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

Paul Eckman. 1972. Universal and cultural differences in facial expression of emotion. In *Nebraska symposium on motivation*, volume 19, pages 207–284. University of Nebraska Press Lincoln.

James Gibson, Nikolaos Malandrakis, Francisco Romero, David C Atkins, and Shrikanth S Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Sixteenth annual conference of the international speech communication association*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Allison Lahnala, Charles Welch, Béla Neuendorf, and Lucie Flek. 2022. Mitigating toxic degeneration with empathetic data: Exploring the relationship between toxicity and empathy. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (Forthcoming)*.

Allison Lahnala, Yuntian Zhao, Charles Welch, Jonathan K Kummerfeld, Lawrence C An, Kenneth Resnicow, Rada Mihalcea, and Verónica Pérez-Rosas. 2021. Exploring self-identified counseling expertise in online support forums. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4467–4480.

Claus Lamm, C Daniel Batson, and Jean Decety. 2007. The neural substrate of human empathy: effects of perspective-taking and cognitive appraisal. *Journal of cognitive neuroscience*, 19(1):42–58.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13622–13623.

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.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435.

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.

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

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to pre-train on? efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605.

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.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. 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.

Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting cognitive and emotional empathic writing of students. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4063–4077, Online. Association for Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Joshua D Wondra and Phoebe C Ellsworth. 2015. An appraisal theory of empathy and other vicarious emotional experiences. *Psychological review*, 122(3):411.

Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020. Quantifying the causal effects of conversational tendencies. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

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.