Last-min-submission at WASSA 2024 Empathy and Personality Shared Task: Enhancing Emotional Intelligence with Prompts

Svetlana Churina
Department of Communications and New Media & Centre for Trusted Internet and Community
National University of Singapore
Singapore
churinas@nus.edu.sg

Preetika Verma & Suchismita Tripathy
Birla Institute of Technology and Science, Pilani
India
f20190088@pilani.bits-pilani.ac.in
f20190554@pilani.bits-pilani.ac.in

Abstract

This paper describes the system for the last-min-submission team in WASSA-2024 Shared Task 1: Empathy Detection and Emotion Classification. This task aims at developing models which can predict the empathy, emotion, and emotional polarity.

This system achieved relatively good results on the competition's official leaderboard. The code of this system is available here.

1 Introduction

Empathy is a warm, tender, and compassionate feeling directed toward a suffering target. It is a crucial aspect of human interaction, playing a significant role in promoting optimal well-being and fostering social connections.

The Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis (WASSA) has organized a 'Shared Task on Empathy Detection, Emotion Classification, and Personality Detection in Interactions.' The aim of this task is to develop models capable of predicting empathy, emotion, personality, and inter-personal index. The provided dataset includes essays written in reaction to news articles where harm has occurred to a person, group, or entity. Additionally, the dataset contains conversations between two users, showcasing their empathetic reactions to the same articles. Each speech turn in these conversations has been annotated with perceived empathy, emotion polarity, and emotion intensity.

Moreover, the dataset offers personality information, including the widely used Big Five (OCEAN) personality traits and the Interpersonal Reactivity Index. Previous research has found that women tend to exhibit higher empathy scores and that there is a negative association between age and empathy. Consequently, demographic information is also provided in the dataset.

The WASSA shared task 2024 (Giorgi et al., 2024) is divided into four tracks:

- Track 1: Empathy Prediction in Conversations (CONV-dialog)
- Track 2: Empathy and Emotion Prediction in Conversation Turns (CONV-turn)
- Track 3: Empathy Prediction (EMP)
- Track 4: Personality Prediction (PER)

We are participating in Track 2. This was newly introduced in WASSA 2023 (Barriere et al., 2023). It involves predicting the perceived empathy, emotion polarity, and emotion intensity at the speech-turn level in a conversation. This task requires a nuanced understanding of the conversational context and the ability to accurately assess the emotional content and empathetic responses within each turn.

The remainder of the paper is structured as follows: Section 2 includes system description, Section 3 talks about experimental results and Section 4 provides conclusion.

2 System Description

2.1 Feature extraction

The dataset (Omitaomu et al., 2022) for Track 2 has been provided with manual annotations regarding Emotion, Emotional Polarity, Empathy, and Self-Disclosure. It has been found that empathetic text is rich in pronouns, emotional, understanding, seeing, and feeling words (Shi et al., 2021). In this context, we extracted additional features of the text to gain a better understanding of empathy.

LIWC (Linguistic Inquiry and Word Count) quantifies language use by measuring the proportion of words in various categories in a given piece of text. These categories include linguistic categories (such as prepositions and pronouns), psychological processes (such as emotion, cognition,

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 425–429

and social), specific topics (such as words related to time, leisure, and money), and punctuation (such as commas and question marks). Using LIWC, we extracted semantic features such as pronoun usage, words related to sadness, politeness, and more. The most relevant features with their correlations can be found in Table 1. We can see, that empathy is strongly correlated with negative politeness (feature politeness HASNEGATIVE) as well as compassion.

Predictors	*Corr*
Empathy	
compassion	0.47
feature politeness HASNEGATIVE	0.37
allsubj	0.28
inflammatory	0.26
NEGEMO	0.236
reasoning	0.232
SAD	0.2
feature politeness 1st person start	0.19
SOCIAL	-0.3
turn id	-0.31
YOU	-0.27
TIME	-0.249
likely to reject	-0.289
Emotion	
compassion	0.446
feature politeness HASNEGATIVE	0.434
allsubj	0.335
toxicity	0.343
inflammatory	0.32
NEGEMO	0.245
SOCIAL	-0.290
TIME	-0.248
YOU	-0.246
likely to reject	-0.22
Emotion Polarity	
feature politeness HASNEGATIVE	0.475
toxicity	0.34
inflammatory	0.294
NEGEMO	0.268
SAD	0.195
respect	-0.37
POSEMO	-0.349
turn id	-0.337
YOU	-0.238

Table 1: Table of extracted features with their orrelations

We found no significant correlation between the demographic features provided in Track 4 and the target scores. Therefore, we are not considering those features further.

2.2 GPT-3.5 turbo finetuning

We used zeroshot prompting with GPT-3.5 turbo (Brown et al., 2020) finetuned on the training dataset. The finetuning was done using OpenAI API for 3 epochs with default temperature. The data was structured as system prompt, prompt and completion trios as follows:

Role: System, **Content:** "You are given a conversation between two people, along with some additional sentiment analysis scores of the last dialog of the conversation."
Role: User, **Content:** <Prompt>
Role: Assistant, **Content:** <Expected response, with scores for Emotion, Emotion Polarity and Empathy>

We did not provide a validation dataset separately for finetuning, and instead combined the train and dev finetuning datasets we generated for final results generation on the test dataset.

2.3 Prompting details

The results of fine-tuning the GPT model heavily depend on the quality and structure of the prompts. For optimal performance, prompts should be carefully crafted and thoroughly tested. In our work, we explored the following approaches to determine the most effective method for our task:

- **Simple Instruction**: The prompt instructs the model to provide scores for 'empathy,' 'emotion,' and 'emotional polarity,' followed by the text to classify.
- **Simple Instruction with Text First**: This prompt is similar to the simple instruction prompt, but the text to classify is provided first, followed by the instruction.
- **Detailed Instruction**: The prompt describes the task goal in detail, explaining what each score means and providing the range of the scores.
- **Simple Instruction with Examples**: After the simple instructions, the prompt includes three samples, providing examples of text with different polarities of scores.

- **Detailed Instruction with Examples**: This is similar to the above, but uses detailed instructions instead of simple ones.

Each prediction is expected to be done on one dialog, as per the dataset. However, we noticed that often, the sentiment analysis for a dialog works better when the previous few dialogs or utterances are also provided to set up context. Using this, we set up prompts providing 2, 5, 10 previous dialogs (of the same conversation) along with each dialog for which the model is expected to predict the required scores.

- This prompt structure was used to generate finetuning data for the provided training dataset, excluding dialogs that did not have the required number of previous dialogs at all.
- Similar prompts were generated for the dev and test datasets. For dialogs that did not have 2, 5 or 10 previous dialogs, we provided as many previous dialogs as available).

We noticed that models finetuned with 2 previous dialogs had too little context for accurate analysis, and models with 10 previous dialogs seemed to get confused/distracted with the extra information provided. 5 previous dialogs (i.e. a total of 6 dialogs per prompt) was ideal, providing just enough information to predict scores.

Using prompts structured with upto 5 previous dialogs, providing the conversation snippet before the instruction, in addition to asking the model to predict all 3 scores in one go (i.e. emption, emotion polarity and empathy), we also tried modifying the instruction to ask the model to only predict one score at an time. Hence, we finetuned 3 specialised models that predict emotion, polarity and empathy separately. Contrary to what we expected however, these models had lower accuracy than the combined model which predicts all 3 scores at once.

Since text features have been extracted and showed improvement for simple models, we created a prompt variation that includes upto 5 previous dialogs, self-disclosure and features that showed a high correlation with the target scores:

- **Features before the conversation snippet**: Before giving the conversation snippet, the features (i.e. LIWC features and self-disclosure as obtained from the dataset) are provided along with explanatory feature names.
- **Features after the conversation snippet**: After giving the conversation snippet, the features (i.e. LIWC features and self-disclosure as obtained from the dataset) are provided along with explanatory feature names. Providing the conversation snippet first seems to help the model better understand the additional information we provided.

Adding LIWC extracted features was decreasing the performance of the model, so we excluded these from our final system.

2.4 Datasets used for finetuning

To augment our training data, we sought additional datasets containing emotion, empathy, or emotional polarity scores. One such dataset is the Emotional Reactions Dataset (Sharma et al., 2020), which provides empathy levels for response posts in the context of seeker posts. This dataset categorizes empathy into three levels: 0 (no empathy) to 2 (high empathy).

Due to the differing scoring systems between this dataset and our original dataset, we normalized the empathy scores to match the range of our required data. Despite this adjustment, fine-tuning our best-performing GPT model with the additional data resulted in a significant drop in performance, with scores decreasing from approximately 0.7 to around 0.3. This decline may be attributed to the differing scoring systems, which could have led to a mismatch in empathy levels after normalization.

3 Experimental Results

3.1 Classical ML approaches

We derived embedding vectors of size 1536 from the **text-embedding-3-small** model using the Embeddings endpoint provided by OpenAI. We creates two sets of embedding inputs, providing the complete utterance history as additional input for second. These were used to train various classical ML models such as Random Forest, RNN, LSTM, and Bi-LSTM. We observed that providing the utterance history increased the average score for all models. The results are present in table 2.

3.2 Adapter-based Finetuning

We fine-tuned an XXL version of the DeBERTAV2 (He et al., 2021) model with 1.5B parameters loaded from a pretrained checkpoint *deberta-v2-xxlarge* on Huggingface. LoRA (Hu et al., 2022)

Model	Emotional polarity	Emotion	Empathy	Average
RNN (without utterance)	0.6895	0.5672	0.5608	0.6058
RNN (with utterance)	0.7021	0.5745	0.5754	0.6173
LSTM (without utterance)	0.7157	0.5814	0.5780	0.6250
LSTM (with utterance)	0.6959	0.5954	0.6026	0.6313
Bi-LSTM (without utterance)	0.7101	0.5875	0.5657	0.6211
Bi-LSTM (with utterance)	0.7085	0.5881	0.5966	0.6311
Random Forest (without utterance)	0.5588	0.4374	0.5075	0.5012
Random Forest (with utterance)	0.5686	0.4574	0.5113	0.5125

Table 2: Pearson coefficients for different models using GPT embeddings

adapters were used to fine-tune the model for 5 epochs without adding the utterance history. The results are present in table 3.

Model	Emotion	Polarity	Empathy	Average
DeBERTAV2 with LoRA	0.5976	0.7312	0.6383	0.6557

Table 3: Pearson coefficients for finetuning DeBERTAV2 with LoRA

3.3 Finetuning GPT

Fine-tuning GPT-3.5-turbo using OpenAI API gave better results than the previous approaches.We experimented with different styles of prompting and controlled the number of previous dialogues while providing the utterance history.

3.3.1 Controlling length of utterances

Utterance history comprises the previous dialogues spoken in the conversation. The conversations had variable sizes. We chose previous *n* turns and found that **n=5** produces the best results. Table 4 has the results for this experiment.

3.3.2 Prompting

We tried out four different ways of prompting described in Table 5. Adding fewshot examples decreased the average scores. For our final model, we used zeroshot prompting with detailed instructions described in Table 6 in the Appendix.

Utterance length	Emotion	Polarity	Empathy	Average
Previous 2	0.6356	0.7918	0.6611	0.6962
Previous 10	0.6519	0.7791	0.6248	0.6853
Previous 5	0.6467	0.8031	0.6653	0.7050
All	0.6215	0.7136	0.6293	0.6548

Table 4: Pearson coefficients for different finetuning GPT with different values of utterance lengths

Prompting style	Emotion	Polarity	Empathy	Average
Simple instruction	0.6443	0.7866	0.6538	0.6949
Detailed instruction	0.6627	0.7880	0.6655	0.7054
Simple instruction + few shot examples	0.6436	0.7845	0.6732	0.7004
Detailed instruction + few shot examples	0.6446	0.7913	0.6593	0.6984

Table 5: Pearson coefficients for different finetuning GPT with fewshot examples and different prompts

4 Conclusion

Empathy and emotion are complex and challenging to predict, largely due to their nuanced nature. Although research in this area is growing, it is still not as extensive as in other domains, leaving significant room for exploration. The limitation of available annotated data further restricts these possibilities. Our experiments indicated that, while adding extra textual features might theoretically enhance empathy detection, LLMs did not significantly improve the scores. However, we found that providing detailed instructions to LLMs increased clarity and resulted in slight improvements. Additionally, we observed that effective empathy and emotion detection requires understanding the background and previous context of the dialogue, underscoring the importance of context in these tasks.

Acknowledgments

This work is supported by the Ministry of Education, Singapore under its MOE AcRF TIER3 Grant (MOE-MOET32022-0001). The travel grant for this research is supported by the Department of Communication and New Media at the National University of Singapore.

References

Valentin Barriere, Jo ao Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of wassa 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, Social Media Analysis*, pages 511–525.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Salvatore Giorgi, Jo ao Sedoc, Valentin Barriere, and Shabnam Tafreshi. 2024. Findings of wassa 2024 shared task on empathy and personality detection in interactions. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, Social Media Analysis*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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. *Preprint*, arXiv:2205.12698.

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.

Shuju Shi, Yinglun Sun, Jose Zavala, Jeffrey Moore, and Roxana Girju. 2021. Modeling clinical empathy in narrative essays. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 215–220.

A Appendix

A.1 Prompt Engineering for GPT-3.5

Type	Prompt
Simple instruction (without features)	Analyze the last dialogue of the conversation and calculate its Emotion, Emotional Polarity, and Empathy scores. You are given a conversation between two people (P1 and P2). <Conversation>
Detailed instruction (without features)	Below is a dialogue between two people regarding a news article. They express their emotions and empathy through the conversation. The Emotion Score is considered to be a measure of how strongly the speaker is feeling the emotions they express (e.g., happy, anxious, sad, angry). The Emotional Polarity Score is considered to be a numerical value rating the type of emotion the speaker is experiencing. It ranges between 1 (positive), 2 (neutral), and 3 (negative). The Empathy Score is considered to be a measure of whether the speaker is taking on the feelings of the suffering victim. If they are, it evaluates how much the speaker seems to put themselves in the shoes of the suffering victim. The value is a numerical score between 1 (not at all) and 5 (extremely). Analyze the last dialogue of the conversation and calculate its Emotion, Emotional Polarity, and Empathy scores. You are given a conversation between two people (P1 and P2). <Conversation>

Table 6: Prompts used on the finetuned GPT models