EmpatheticFIG at WASSA 2024 Empathy and Personality Shared Task: Predicting Empathy and Emotion in Conversations with Figurative Language

Gyeongeun Lee * **Zhu Wang*** **Sathya N. Ravi** **Natalie Parde**
Department of Computer Science
University of Illinois at Chicago
{glee87, zwang260, sathya, parde}@uic.edu

Abstract

Recent research highlights the importance of figurative language as a tool for amplifying emotional impact. In this paper, we dive deeper into this phenomenon and outline our methods for *Track 1, Empathy Prediction in Conversations* (CONV-dialog) and *Track 2, Empathy and Emotion Prediction in Conversation Turns* (CONV-turn) of the WASSA 2024 shared task. We leveraged transformer-based large language models augmented with figurative language prompts, specifically idioms, metaphors and hyperbole, that were selected and trained for each track to optimize system performance. For Track 1, we observed that a fine-tuned BERT with metaphor and hyperbole features outperformed other models on the development set. For Track 2, DeBERTa, with different combinations of figurative language prompts, performed well for different prediction tasks. Our method provides a novel framework for understanding how figurative language influences emotional perception in conversational contexts. Our system officially ranked 4th in the 1st track and 3rd in the 2nd track.

1 Introduction

The computational study of empathy[1] is crucial to enabling and advancing the development of innovative and resourceful tools for social good in various settings, ranging from online conversations to clinical therapy (Eysenbach et al., 2004; Elliott et al., 2018). At a broader level, recognizing emotional needs and appropriately responding to them are essential for successful interactions, making this an important step in chatbot development. However, despite the recent surge in interest in automated empathy detection (Barriere et al., 2023; Lee and Parde, 2024; Giorgi et al., 2024; Lee et al., 2024), research focusing on empathetic dialogue involving back-and-forth conversations (such as that by Rashkin et al. (2018)) still remains scarce. The Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA) 2023 (Barriere et al., 2023) and 2024 (Giorgi et al., 2024) provides the opportunity to explore this domain further with the *Empathic Conversations* dataset (Omitaomu et al., 2022).

Analyzing linguistic features is crucial in understanding how language is used to convey empathy and emotion in dialogue. Figurative language (non-literal language; see § 2.3.2 for more information), and particularly metaphor, has been shown to enhance the performance of emotion prediction models (Dankers et al., 2019), hold more emotional charge than literal language (Citron and Goldberg, 2014), and strengthen expressions (Mohammad et al., 2016; Li et al., 2023). From our own recent work, we found that figurative language prompts improved empathy detection performance when using the domain-specialized *AcnEmpathize* dataset (Lee et al., 2024). Therefore, we hypothesized that identifying the use of figurative language in *Empathic Conversations* could likewise provide deeper insight into the study of empathy and emotion. We thus propose methods to encode figurative language prompts into large language models (LLMs) for emotion and empathy prediction in conversations.

In this shared task, we participated in two of the four tracks:

- **Track 1**: *Empathy Prediction in Conversations* (CONV-dialog), focusing on predicting perceived empathy at the dialogue level.
- **Track 2**: *Empathy and Emotion Prediction in Conversation Turns* (CONV-turn), focused on predicting perceived empathy, emotion polarity, and emotion intensity at the speech-turn level in a conversation.

*Equal contribution

[1]We follow Davis et al. (1980)'s definition of *empathy* as the ability to understand and respond to the experiences and feelings of others.

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 441–447

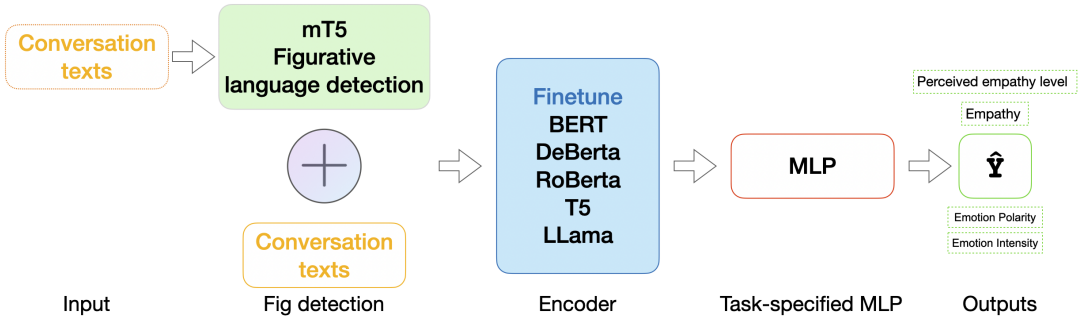

Figure 1: EmpatheticFIG architecture. We combined figurative language prompts with conversation text and passed the combined input through various text encoders. We applied different pre-trained LLMs, such as BERT and DeBERTa. Then, we used task-specific MLPs to obtain the outputs.

In both tracks, we implemented transformer-based models with additional figurative language prompts. We applied combinations of figurative language prompts that yielded the best performance on the development sets. In Track 1, we used metaphor and hyperbole features with BERT (Devlin et al., 2018). In Track 2, we used idiom and hyperbole features together for empathy prediction, hyperbole for emotional polarity, and metaphor for emotion intensity with DeBERTa (He et al., 2021).

2 System Description

2.1 Track 1: Empathy Prediction in Conversations (CONV-dialog)

The goal of this track was to predict the perceived empathy of one person towards another based on their conversation. For each conversation ID, we were given the speaker ID of Speaker 1 and their perceived empathy from the other person, Speaker 2. We extracted all texts from Speaker 2 in each conversation and combined them to predict the empathy level perceived by Speaker 1, implementing this as a multi-class classification problem.

2.2 Track 2: Empathy and Emotion Prediction in Conversation Turns (CONV-turn)

In this track, we used the text of each speaker in each conversation to predict turn-level annotated emotion, emotional polarity, and empathy. Since the target labels were provided for each turn, preprocessing the text was unnecessary. We encoded the conversations and figurative language prompts at the turn level, subsequently passing them through the task-specific multilayer perception (MLP) layer to obtain the output scores.

2.3 EmpatheticFIG

We propose EmpatheticFIG, a framework that incorporates figurative language prompts for empathy and emotion prediction in conversations. It draws inspiration from our earlier work (Lee et al., 2024) and that of others emphasizing the importance of metaphor to emotional expression (Citron and Goldberg, 2014; Mohammad et al., 2016; Dankers et al., 2019; Li et al., 2023). The main architecture is illustrated in Figure 1.

We used mT5 (Xue et al., 2021) to extract figurative language prompts (described further in §2.3.2) and appended these features to the conversation texts. These combined inputs were then processed by pre-trained LLMs, including BERT (Devlin et al., 2018), DeBerta (He et al., 2021), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020), and Llama (Touvron et al., 2023). Finally, we applied task-specified MLP layers for Track 1 and for different tasks in Track 2 to generate outputs.

2.3.1 Models and Hyperparameters

We conducted extensive experiments using different LLMs and hyperparameter settings. In Track 1, the backbone models were BERT-base, DeBerta-v3-base, RoBERTa-base, and Llama-3-8b. In Track 2, we fine-tuned BERT, DeBerta, RoBERTa, and T5-small. We fine-tuned BERT, DeBerta, RoBERTa, and T5 on conversation texts with figurative language prompts for both tracks. When using Llama, we utilized Parameter-Efficient Fine-Tuning (Hu et al., 2021, PEFT) to adapt the model weights for perceived empathy level prediction for Track 1.

For both tracks, we searched for the optimal hyperparameters from the following sets: learning rates ranging from {3e-4, 1e-4, 5e-5, 1e-5, 5e-6}, batch sizes from 2 to 64 depending on the model, and training epochs ranging from {1, 3, 5, 10, 30,

Table 1: The combined total number of texts containing each type of figurative language in the training and development datasets in Track 1 (CONV-dialog) and Track 2 (CONV-turn). Each text may include multiple types of figurative language.

	Track 1	Track 2
# Texts	1,037	12,080
# Idiom	552 (53%)	6,378 (53%)
# Metaphor	267 (26%)	1,472 (12%)
# Hyperbole	120 (12%)	1,795 (15%)
Total Figurative	939 (91%)	9,645 (80%)

50}. The AdamW (Loshchilov and Hutter, 2017) optimizer was used with a weight decay of 0.01.

2.3.2 Figurative Language Prompts

Since empathy is primarily conveyed through language, we focused our investigation on the performance benefits of incorporating automatically extracted linguistic constructs (figurative language phenomena) from the text to assist in predicting empathy and emotion labels. Figurative language is non-literal language (Paul, 1970) that serves to compare, exaggerate, and add nuanced meaning; oftentimes, it is used as a vehicle to simplify complicated or abstract ideas.

We applied the multi-figurative language detection method from Lai et al. (2023) to identify metaphor, idiom, and hyperbole (three distinct types of figurative language, explained later in this section) in *Empathic Conversations*. Lai et al. (2023)'s approach requires the use of a pre-trained LLM; we specifically employed mT5 (Xue et al., 2020) using predefined prompts of the format:

```
Which figure of speech does this
text contain? (A) Literal (B)
[Task] | Text: [Text]
```

We filled `[Task]` using `metaphor`, `idiom`, and `hyperbole`, respectively. Using this approach, we then one-hot encoded the labels for each detected type of figurative language if at least one type was present in the text. Finally, we appended a description to the original conversation text that reads, *The text contains <label>.*, where "label" represents the type of figurative language. Again, note that each text may contain more than one type of figurative language (which were appended separated by commas).

In Table 1, we summarize the distributions of figurative language types in the combined training and development datasets for Track 1 and Track 2. We find that the datasets for both are rich in figurative language—in total, figurative language is present in approximately 91% and 80% of the samples, respectively. Below, we briefly define and provide examples for each type of figurative language identified.

Table 2: Track 1 development set results. The best combinations of figurative language prompts for each model are shown here. "-" denotes models without figurative language prompts. Bold indicates the highest Pearson correlation coefficient (Pearson r).

Model	Fig. Features	Pearson r
RoBERTa	-	0.164
DeBERTa	-	0.158
Llama	-	0.0713
BERT	-	0.185
RoBERTa	metaphor	0.198
DeBERTa	all	0.185
Llama	-	0.0713
BERT	**metaphor, hyperbole**	**0.242**

Metaphor. Metaphoric expressions frame conventional ideas in more accessible terms by assigning new meanings (Lakoff and Johnson, 1980). For example, the phrase "fight these natural disasters" from *Empathetic Conversations* personifies natural calamities as adversaries one must battle.

Idiom. Idiomatic expressions tie abstract ideas or meanings to more concrete anchors and often involve cultural context (Nunberg et al., 1994). For example, the phrase "pull through the hard times" from *Empathetic Conversations* uses idiomatic language to convey the act of enduring and surviving difficult times.

Hyperbole. Hyperbolic expressions are often used to exaggerate a statement or phenomenon (Claridge, 2010), as in "thousands of years of human progress." This phrase from the shared task dataset likely emphasizes the significance of human progress over a long period, rather than literally referring to exactly thousands of years.

3 Results and Discussions

To investigate the role of figurative language in empathy and emotion prediction in conversations, we implemented EmpatheticFIG for Track 1 and Track 2. We conducted experiments using different combinations of figurative language prompts and generally found that many models with idiom features performed worse, perhaps due to the high prevalence of idioms in the dataset. We present

Table 3: Track 2 development set results. The performance of the baseline models is displayed in rows 1-4, with DeBERTa as the top-performing baseline model. We also present the results of different combinations of figurative language prompts using DeBERTa in rows 5-11. Bold text indicates the best Pearson r for each task. Different figurative language prompts improved different types of emotion and empathy predictions.

Model	Empathy	Emotion Polarity	Emotion Intensity
T5-small (Raffel et al., 2020)	0.556	0.63	0.658
BERT (Devlin et al., 2018)	0.619	0.735	0.653
RoBERTa (Liu et al., 2019)	0.626	0.739	0.658
DeBERTa (He et al., 2021)	0.633	0.759	0.66
DeBERTa + idiom	0.632	0.745	0.655
DeBERTa + metaphor	0.632	0.751	**0.666**
DeBERTa + hyperbole	0.633	**0.765**	0.659
DeBERTa + idiom + metaphor	0.62	0.745	0.644
DeBERTa + metaphor + hyperbole	0.6311	0.75	0.66
DeBERTa + idiom + hyperbole	**0.661**	0.748	0.622
DeBERTa + all	0.656	0.761	0.635

the experimental results on the development and test sets for Track 1 and Track 2 in Tables 2 and 3, respectively. To assess the impact of figurative language prompts, we also conducted ablation experiments using various combinations of figurative language types. Moreover, we evaluated the performance of these features across different LLMs. We discuss the results for Tracks 1 and 2 in more detail in §3.1 and §3.2, respectively.

3.1 Results for Track 1 (CONV-dialog)

Table 2 presents the development set results for Track 1, showcasing the averaged Pearson correlation values from three runs for each setting. Models incorporating figurative language prompts consistently outperformed the baseline models, suggesting that these features positively impact perceived empathy level predictions. Metaphors generally improved the performance for perceived empathy level prediction and seemed to enhance the ability of language models to capture the nuanced emotional cues in empathetic conversations.

We applied the settings that yielded the best performance on the development set to the test set, specifically using BERT with metaphor and hyperbole features, a learning rate of 5e-5, a batch size of 8, and an epoch length of 3. However, we observed a drop in performance on the test set (see results in Appendix A.1), potentially due to class imbalance and overfitting. We leave further investigations of test set performance for future work.

3.2 Results for Track 2 (CONV-turn)

We present the results for each task in Track 2 in Table 3. DeBERTa was the best performing model among all the baseline models. We observed that task performance varied depending on the specific figurative language prompts used on the development set. DeBERTa with idiom and hyperbole features notably improved empathy predictions. DeBERTa with hyperbole achieved the highest performance on emotion polarity predictions, while DeBERTa with metaphor slightly enhanced predictions for emotion intensity.

We found that different figurative language prompts provided varying levels of information and impact on different task predictions. For example, hyperbole contributed to improved emotion polarity, while idioms and hyperbole enhanced performance in empathy prediction for turn-level conversations. In contrast, metaphor had less impact on empathy and emotion predictions at the turn level (Track 2) compared to its impact at the dialogue level in Track 1. The test set performance (see Appendix A.2) also shows consistent results.

4 Conclusion

Our team, EmpatheticFIG, participated in Track 1 (Empathy Prediction in Conversations) and Track 2 (Empathy and Emotion Prediction in Conversation Turns) of the WASSA 2024 shared task. Our system architecture involved fine-tuning various LLMs, such as BERT and DeBERTa, with one or more combinations of figurative language types—idioms, metaphors, and hyperbole. The results showed that incorporating figurative language prompts was beneficial for predicting empathy and emotion in conversations. Our method provides unique insights into adapting figurative language prompts into LLMs.

5 Limitations and Future Work

We could experiment with larger, more complex models and perform extensive hyperparameter tuning. Additionally, we could verify the detection of figurative language expressions in the dataset and conduct a deeper analysis of their usage to better understand the reasons behind the performance. Furthermore, we could explore more types of figurative language beyond those we have already investigated.

Acknowledgments

We thank the anonymous reviewers for their feedback, and the WASSA 2024 shared task organizers for running this task. This work was partially supported by the National Science Foundation under Grant No. 2125411. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Valentin Barriere, João 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.

Francesca MM Citron and Adele E Goldberg. 2014. Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of cognitive neuroscience*, 26(11):2585–2595.

Claudia Claridge. 2010. *Hyperbole in English: A corpus-based study of exaggeration*. Cambridge University Press.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229.

Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy. *JSAS: catalog of selected documents in psychology*, 10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399.

Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita Stern. 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *Bmj*, 328(7449):1166.

Salvatore Giorgi, João 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, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. *arXiv preprint arXiv:2306.00121*.

George Lakoff and Mark Johnson. 1980. Metaphors we live by. *University of Chicago, Chicago, IL*.

Gyeongeun Lee and Natalie Parde. 2024. AcnEmpathize: A dataset for understanding empathy in dermatology conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 143–153, Torino, Italia. ELRA and ICCL.

Gyeongeun Lee, Christina Wong, Meghan Guo, and Natalie Parde. 2024. Pouring your heart out: Investigating the role of figurative language in online expressions of empathy. In *Proceedings of the 262nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, page Accepted, Bangkok, Thailand. Association for Computational Linguistics.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2023. The secret of metaphor on expressing stronger emotion. *arXiv preprint arXiv:2301.13042*.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the fifth joint conference on lexical and computational semantics*, pages 23–33.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

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.

Anthony M Paul. 1970. Figurative language. *Philosophy & Rhetoric*, pages 225–248.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Appendix

A.1 Test results for Track 1

We illustrate our test results for Track 1, noting a significant performance drop on the test set. The performance on the dev set was unstable and sensitive to hyperparameters. We will leave the investigation of the test set settings and improving the robustness of our system, EmpatheticFIG, for future work.

Table 4: Track 1 test set results.

	Perceived Empathy	
	r	p
FraunhoferSIT	0.193	0.127
ConText	0.191	0.130
Chinchunmei	0.172	0.173
EmpatheticFIG	0.012	0.923

A.2 Test results for Track 2

The performance of our model on the test set in Track 2 showed similar pattern to our performance on the development set. Our model performed best in predicting emotion polarity, empathy, and emotion intensity (in this order) and achieved an average Pearson r of 0.610.

Table 5: Track 2 test set results.

	Average	**Empathy**		**Emotion Polarity**		**Emotion Intensity**	
	r	r	p	r	p	r	p
ConText	0.626	0.577	0.000	0.679	0.000	0.622	0.000
Chinchunmei	0.623	0.582	0.000	0.680	0.000	0.607	0.000
EmpatheticFIG	0.610	0.559	0.000	0.671	0.000	0.601	0.000
Last_min_submission_team	0.595	0.534	0.000	0.663	0.000	0.589	0.000
hyy3	0.590	0.544	0.000	0.644	0.000	0.581	0.000
Empathify	0.588	0.541	0.000	0.638	0.000	0.584	0.000
empaths	0.477	0.534	0.000	0.422	0.000	0.473	0.000
FraunhoferSIT	-0.007	0.034	0.125	-0.018	0.409	0.032	0.141
Zhenmei	-0.030	-0.027	0.223	-0.020	0.356	-0.043	0.051