Optimizing Empathy Prediction in Text-Based Mental Health Support with Multi-Task Learning and Imbalance Mitigation

Anjitha Divakaran[1], Hyung-Jeong Yang[1*], Seung-won Kim[1], Ji-eun Shin[2], Soo-Hyung Kim[1]
[1]Dept. of AI Convergence, Chonnam National University
Dept. of Psychology, Chonnam National University
*Corresponding author: hjyang@jnu.ac.kr

ABSTRACT

Empathy plays a crucial role in effective mental health support, particularly on text-based platforms where individuals seek understanding and compassion. Building on previous work that predicted empathy in separate models, we propose a multi-task learning approach that jointly models three empathy communication mechanisms: emotional reactions, interpretations, and explorations. By integrating RoBERTa with GRU layers, our model improves the accuracy and efficiency of empathy detection. Additionally, we address class imbalance by incorporating focal loss, which helps the model focus on underrepresented strong empathy signals. Our experiments show that the multi-task model, combined with focal loss, significantly improves performance across all empathy dimensions, making it a promising tool for enhancing real-time empathy feedback in online mental health support systems.

1. INTRODUCTION

According to WHO studies, mental disorders impact 12.2% to 48.6% of individuals over their lifetimes and account for 14% of the global illness burden, yet access to care remains limited, with an average of just nine mental health practitioners per 100,000 people worldwide [1]. The shortage of mental health practitioners, averaging only nine per 100,000 people globally, has fueled the demand for online mental health services. In recent years, the shift towards text-based mental health platforms, such as online peer support forums and therapy apps, has provided millions of people with a space to express their emotions and seek help [2]. Moreover, empathy plays a fundamental role in mental health care, fostering deeper connections and enhancing therapeutic relationships by promoting understanding and compassion, which is essential for meaningful patient-provider interactions.

Research in natural language processing (NLP) has developed computational models to detect empathy in text-based conversations. Sharma et al. introduced a RoBERTa-based bi-encoder model for detecting empathy in online mental health support, focusing on three key mechanisms: Emotional Reactions, Interpretations, and Explorations, which capture different aspects of empathy in response posts [3]. While their approach showed promising results, using separate models for each mechanism limits information sharing across tasks and can lead to inefficiencies.

We propose a multi-task learning model to predict three empathy mechanisms—emotional reactions, interpretations, and explorations—simultaneously. By sharing representations across tasks, the model captures the interactions between these mechanisms. To address class imbalance, we incorporate focal loss, which helps the model focus on difficult-to-classify examples, improving empathy detection for underrepresented instances. This approach balances the training process and enhances the model's accuracy in detecting empathy in mental health conversations.

The main contributions of this paper are as follows:

1. We present a multi-task learning model using RoBERTa and GRU to predict emotional reactions, interpretations, and explorations, improving efficiency and performance.
2. We address class imbalance in empathy detection by applying focal loss, enhancing the model's ability to identify rare empathy instances.
3. We perform experiments comparing rationale-augmented and rationale-free training approaches, demonstrating the improvements brought by our model enhancements.

Through these contributions, we aim to improve the detection of empathy in text-based mental health support, ultimately enabling more effective real-time feedback systems to assist peer supporters in providing compassionate and meaningful help.

2. RELATED WORKS

Empathy detection in text-based conversations has garnered increasing attention in recent years due to the rise of online mental health platforms. Understanding and predicting empathic responses in these settings is crucial for improving the quality of support offered to users. This section discusses prior research on empathy detection, multi-task learning, and handling class imbalance in natural language processing (NLP).

2.1 Empathy Detection in Text-Based Mental Health Support

Empathy detection in text-based mental health support is a growing research area, helping improve peer support and emotional response systems. Sharma et al. developed a RoBERTa-based bi-encoder model to detect emotional reactions, interpretations, and explorations, though it relied on separate models for each mechanism, limiting efficiency [3]. Other studies, like Pérez-Rosas et al., explored empathy in multimodal settings using text, audio, and video, offering insights into how empathy manifests across different modalities [4]. While these studies laid foundational work in detecting empathy in mental health conversations, they relied on separate models for each mechanism, thus limiting task interdependence.

Building on these efforts, our work focuses on improving empathy detection in mental health discussions through a more efficient multi-task learning approach [5][6].

2.2 Multi-Task Learning for NLP

Multi-task learning (MTL) enables multiple related tasks to be learned simultaneously through shared representations, improving generalization and reducing overfitting in NLP tasks like sentiment analysis and named entity recognition [7]. This approach is especially useful when labeled data is limited. For empathy detection, MTL allows a single model to predict emotional reactions, interpretations, and explorations, capturing task dependencies and enhancing efficiency [3]. MTL has demonstrated improved performance in various NLP domains by leveraging shared task knowledge [6][8]. In our study, MTL improves empathy detection by simultaneously predicting multiple empathy mechanisms in mental health conversations.

2.3 Addressing Class Imbalance in NLP

Class imbalance is a common issue in NLP tasks, particularly when predicting rare classes like strong empathy signals. Traditional models often struggle with underrepresented classes, leading to biased results. Methods like oversampling, under sampling, and class-weighted loss have been proposed to address this [8]. More recently, focal loss has emerged as a solution, enabling models to focus on challenging, underrepresented examples [11]. Research by Usuga-Cadavid et al. further demonstrates the effectiveness of focal loss in Transformer-based models, such as RoBERTa, making it a suitable technique for our empathy detection task [12].

3. PROPOSED METHODS

In this work, we propose a multi-task learning approach for empathy detection in text-based mental health conversations. The goal is to predict three empathy communication mechanisms—emotional reactions, explorations, and interpretations—simultaneously, using a shared model architecture. This section describes the architecture of our model, the rationale for selecting this approach, and the techniques used to handle class imbalance.

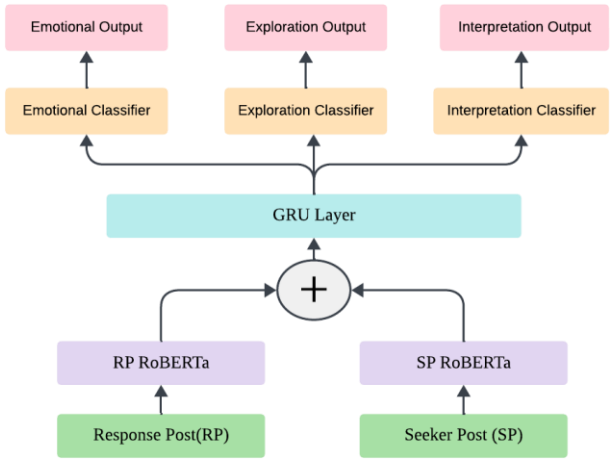
3.1 Multi-Task Learning with RoBERTa and GRU

To capture the complex interactions between the different empathy mechanisms, we employ a multi-task learning (MTL) architecture. By jointly learning these tasks, the model benefits from shared representations across tasks, allowing it to generalize better and reduce overfitting. The proposed model combines the RoBERTa language model and a GRU (Gated Recurrent Unit) to process the textual data.

RoBERTa Encoder: RoBERTa, a transformer-based model, is used to generate contextual embeddings for the input text. Both the Seeker Post (SP) and Response Post (RP) are tokenized and fed into the RoBERTa encoder to capture the contextual meaning of the conversation.

GRU Layer: The outputs from RoBERTa are concatenated with optional rationale embeddings, and the combined sequence is fed into a GRU layer. The GRU is designed to capture sequential dependencies in the text, allowing the model to understand the flow of empathy across the posts.

Task-Specific Heads: After passing through the GRU, the model branches into three separate linear layers, each responsible for predicting one of the empathy mechanisms. These heads predict the emotional, exploration, and interpretation levels for each input.



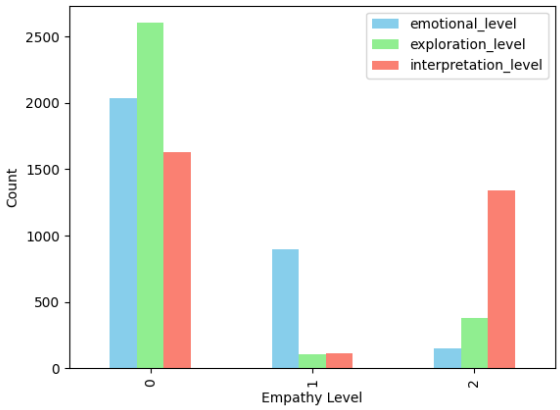
(Fig 1) Proposed Architecture

3.2 Handling Class Imbalance with Focal Loss

One of the key challenges in this task is the class imbalance, particularly the underrepresentation of strong empathy signals as shown in fig 2. To address this, we introduce focal loss, a loss function designed to focus more on difficult-to-classify examples, thereby mitigating the impact of the majority class dominating the learning process.

Unlike standard cross-entropy loss, focal loss reduces the

relative loss for well-classified examples, placing more focus on misclassified instances. This is particularly useful in tasks like empathy detection, where the strong empathy class (Level 2) is significantly rarer than the weak or absent empathy classes (Levels 0 and 1).



(fig 2) Class Distribution of Each empathy level

By applying focal loss, the model gives more weight to rare instances of strong empathy, improving its ability to detect these crucial signals.

4. DATA SET AND RESULTS

In this section, we present the dataset and results of our experiments, comparing the performance of various model configurations across the three empathy communication mechanisms: emotional reactions, explorations, and interpretations. We evaluate the models based on accuracy and F1-score, both of which are crucial metrics for assessing the quality of predictions in the context of class imbalance.

4.1 Dataset

For this study, we used a dataset consisting of posts from online mental health platforms, primarily focusing on seeker-response pairs. The dataset was annotated following the framework established by Sharma et al. [3], where three empathy communication mechanisms: Emotional Reactions, Interpretations, and Explorations are labeled. These mechanisms capture how peer supporters' express empathy in their responses to seekers' posts.

The dataset is structured as follows:

1. Seeker Post (SP): The original post by a user seeking help or emotional support.
2. Response Post (RP): The peer supporter's reply to the seeker's post.
3. Annotations: Each response post is annotated based on the level of empathy expressed in three communication mechanisms:
   - Emotional Reactions (ER): Communicating warmth, compassion, or concern.
   - Interpretations (IP): Communicating an understanding of the seeker's feelings or experiences.
   - Explorations (EX): Probing for more information or encouraging the seeker to explore their feelings.
   - The levels of empathy for each communication mechanism are annotated as:
     - Level 0 (No Communication): No empathy expressed.
     - Level 1 (Weak Communication): Weak empathy expressed.
     - Level 2 (Strong Communication): Strong empathy expressed.

   The dataset is comprised of thousands of seeker-response pairs, with empathy levels annotated by crowd workers trained to identify these mechanisms. The dataset exhibits significant class imbalance, particularly in the distribution of weak and strong empathy levels, which poses a challenge for the model

4.2 Model Configurations

We evaluated four main configurations:

- Simple Train without Rationale (T1): A multi-task model trained without using rationale embeddings.
- Simple Train with Rationale (T2): A multi-task model trained with rationale embeddings.
- Focal Loss without Rationale (T3): A multi-task model trained without rationale embeddings, using focal loss to address class imbalance.
- Focal Loss with Rationale (T4): A multi-task model trained with rationale embeddings, using focal loss to address class imbalance.

In our experiments, we used the RoBERTa-base model with a sequence length of 64 tokens and trained with a batch size of 32. The model was optimized using the Adam optimizer with a learning rate of 1e-5, and focal loss was employed with $\gamma = 2$ to handle class imbalance. We applied a dropout rate of 0.3 to prevent overfitting, and the GRU layer had 256 hidden units across 2 layers. The training was run for 5 epochs, with early stopping based on validation loss, and models were trained on a GPU for faster processing.

4.3 Performance Comparison

The table below summarizes the performance of each configuration across the three empathy mechanisms. We report the accuracy and F1-scores for each mechanism, demonstrating the impact of focal loss and rationale usage on the model's ability to detect empathy levels.

<Table1> Results of different model configurations

Empathy Label / Model		T1	T2	T3	T4
Accuracy	ER	78.01	78.01	80.6	77.8
	EX	93.53	91.59	92.88	91.59
	IP	82.97	81.68	84.26	80.81
F-Score	ER	76.86	77.8	80.23	78.36
	EX	92.35	90.03	92.16	91.52
	IP	81.03	79.81	82.28	78.98

4.4 Analysis of Results

Table 1 depicts the result of each model configuration. The results show that the use of focal loss significantly improves both the accuracy and F1-score for detecting emotional reactions and interpretations, especially in models that do not incorporate rationales. Focal loss effectively addresses class imbalance by focusing on underrepresented empathy signals. Additionally, models without rationale embeddings performed similarly or slightly better than rationale-augmented models, particularly with focal loss, suggesting that the absence of rationales does not significantly affect performance when focal loss is applied.

5. CONCLUSION

This study presented a multi-task learning approach for detecting empathy in text-based mental health conversations, addressing the challenges of predicting emotional reactions, interpretations, and explorations simultaneously. By leveraging a RoBERTa-GRU architecture, our model achieved improved performance over single-task models by sharing representations across empathy mechanisms. Additionally, the use of focal loss helped mitigate the class imbalance issue, particularly in detecting strong empathy signals, leading to more balanced and accurate predictions.

Our results demonstrated that multi-task learning, combined with focal loss, enhances empathy detection in mental health support systems. While rationale-augmented models provide interpretable outputs, rationale-free models offer comparable performance with reduced complexity. Future work could further explore techniques for improving class imbalance handling and expand the application of this model to broader domains within mental health.

ACKNOWLEDGEMENT

This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2023-RS-2022-00156287) supervised by the Institute for Information & communications Technology Planning & Evaluation (IITP). Additionally, it was funded by IITP under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant, and by the National Research Foundation of Korea (NRF) through the Korea government (MSIT) grant (RS-2023-00219107).

REFERENCES

[1] Dhyani A, Gaidhane A, Choudhari SG, Dave S, Choudhary S. Strengthening Response Toward Promoting Mental Health in India: A Narrative Review. Cureus. 2022 Oct 18;14(10):e30435. doi: 10.7759/cureus.30435. PMID: 36407164; PMCID: PMC9671264.

[2] Webb, M., Burns, J., & Collin, P. (2008). Providing online support for young people with mental health difficulties: challenges and opportunities explored. Early intervention in psychiatry, 2(2), 108-113.

[3] 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.

[4] 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, Vancouver, Canada. Association for Computational Linguistics.

[5] Buechel, S., Buffone, A., Slaff, B., Ungar, L.H., & Sedoc, J. (2018). Modeling Empathy and Distress in Reaction to News Stories. Conference on Empirical Methods in Natural Language Processing.

[6] 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.

[7] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. J. Mach. Learn. Res. 12, null (2/1/2011), 2493–2537.

[8] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

[9] Ruder, S. (2017). An Overview of Multi-Task Learning in Deep Neural Networks. ArXiv, abs/1706.05098.

[10] H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

[11] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2999-3007, doi: 10.1109/ICCV.2017.324.

[12] Usuga-Cadavid, J., Gil, D., & Ruiz, E. (2021). Exploring the influence of focal loss on transformer models. Procedia Computer Science, 184, 193-200. https://doi.org/10.1016/j.procs.2021.03.055