Empathify at WASSA 2024 Empathy and Personality Shared Task: Contextualizing Empathy with a BERT-Based Context-Aware Approach for Empathy Detection

Arda Numanoğlu[1], Süleyman Ateş[1], Nihan Kesim Çiçekli[1], Dilek Küçük[1]

[1]Middle East Technical University, Ankara, Turkey

{arda.numanoglu, ates.suleyman, cicekli, kucuk}@metu.edu.tr

Abstract

Empathy detection from textual data is a complex task that requires an understanding of both the content and context of the text. This study presents a BERT-based context-aware approach to enhance empathy detection in conversations and essays. We participated in the WASSA 2024 Shared Task (Giorgi et al., 2024), focusing on two tracks: empathy and emotion prediction in conversations (CONV-turn) and empathy and distress prediction in essays (EMP). Our approach leverages contextual information by incorporating related articles and emotional characteristics as additional inputs, using BERT-based Siamese (parallel) architecture. Our experiments demonstrated that using article summaries as context significantly improves performance, with the parallel BERT approach outperforming the traditional method of concatenating inputs with the '[SEP]' token. These findings highlight the importance of context-awareness in empathy detection and pave the way for future improvements in the sensitivity and accuracy of such systems. Our system officially ranked 8th at both CONV-T and EMP tracks.

1 Introduction

The exploration of empathy detection from text presents a complex yet fascinating challenge that bridges the gap between human emotions and computational analysis. It is an area rich with potential for understanding how we connect and empathize through written communication. Empathy detection involves not only identifying the presence of empathy but also understanding its context, intensity, and the specific emotions it is associated with. The subjective nature of empathy amplifies the complexity of this task, considering the diversity of its expression in language and the contextual sensitivity required to accurately interpret it.

The WASSA 2024 Shared Task 1 focuses on empathy detection in different textual data with four different tracks for participants to compete in. We have participated in two of the four tracks, which are:

Track 2: Empathy and Emotion Prediction in Conversations (CONV-turn).

Track 3: Empathy and Distress Prediction in Essays (EMP).

Section 2 summarizes the related work. In section 3, we detail our system descriptions, including data preprocessing for each of the tracks. In section 4, we present our experimental results for our proposed system architectures, with useful comparisons. In section 5, we discuss our conclusions.

2 Related Work

Research in empathy detection and emotion classification has rapidly evolved with the introduction of sophisticated deep learning models. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) is a pre-trained transformer model that has revolutionized natural language processing tasks. Guda et al. (2021) introduced a demographic-aware BERT-based model for empathy prediction, emphasizing the role of demographic information in enhancing accuracy. However, Wang et al. (2023) showed that demographic data does not always improve performance, as their text-only system excelled in empathy detection, indicating context-dependency. Chavan et al. (2023) improved BERT-based models' performance through ensembles for empathy and distress detection. Lu et al. (2023) highlighted the importance of window size in fine-tuning DeBERTa models for conversation-level empathy prediction, finding that optimizing window size is crucial for capturing empathetic content effectively.

In our work, we prioritize context-awareness by integrating contextual information into our model. This strategic integration allows for a more comprehensive analysis of textual nuances, significantly

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 393–398

enhancing our system's ability to accurately assess empathy across various text forms, including essays and dialogues. By analyzing both the input text and its context, our model achieves a higher Pearson correlation in detecting empathy and distress, especially in essay-level inputs.

3 System Description

In recent years, advancements in empathy detection at the essay level have not matched the progress seen in conversation-level detection (Barriere et al., 2023). We attribute this disparity to a few key challenges:

- the increased complexity associated with the larger number of words in essays.
- the potential reliance of the empathy concept on other aspects of the input.
- the relatively small size of available datasets.

We approach the problem from a different perspective. We begin by rethinking the concept of empathy itself. In our opinion, empathy should not be assessed solely based on the sentences, but also in relation to the context of those sentences. Since each essay-level and conversation-level input is written based on an article, we think it is beneficial to include those articles and the emotional characteristics of the sentences as context for the model. This enables the model to better conceptualize the empathetic nuances.

To develop an effective context-aware model for empathy detection, we initially included related articles along with the emotional characteristics of the input in our model configuration. However, systematic testing revealed that using just the related articles as context was more effective only at the essay level. Conversely, for conversational analysis, incorporating articles as context did not yield the desired results. We observed that most conversational turns consist of everyday dialogue, which does not share similar empathetic features with the articles. This discrepancy led the model to misunderstand the relationship between the context and the conversational turns. Therefore, we decided to use window-based turns as contextual input for the conversational turn analysis. This approach considers the immediate turn before and after each target turn, providing a more relevant context for predicting emotional intensity, emotional polarity, and empathy scores.

Detailed experiment results demonstrating the impact of these contextual elements on model performance are presented in the subsequent sections of this paper. Guided by these observations, we opted for the architectures illustrated in Figures 2 and 1, which are further detailed in the following sections.

3.1 Track 2: Empathy and Emotion Prediction in Conversations (CONV-turn)

The input for this track, which focuses on predicting emotional intensity, emotional polarity, and empathy in conversations, is the individual turn, with the context being the combination of turns determined by the window size. This allows the model to understand the immediate emotional and empathetic context within the conversation.

Inspired by the Siamese BERT Network architecture Reimers and Gurevych (2019), our model processes the input and the context in parallel. The input and the context are fed independently into identical BERT encoders as depicted in Figure 1. After pooling the outputs from the BERT encoder using the CLS token, the embeddings are concatenated and passed through fully connected layers (MLP) to predictions. The final output layer provides emotional intensity, emotional polarity and empathy scores.

3.2 Track 3: Empathy and Distress Prediction in Essays (EMP)

The architecture for empathy and distress prediction in essays integrates contextual information from the related article. This allows the model to capture nuanced expressions of empathy and distress.

Initially, we faced a challenge with articles being too long for the model's maximum input length, averaging 916 tokens. To address this, we used ChatGPT[1] to summarize the articles, reducing their length to 123 tokens while preserving essential context.

Our model includes input layers for both the essay text and the summarized article text, which are independently fed into identical BERT encoders. The outputs are pooled using the CLS token, concatenated, and passed through fully connected layers (MLP) to provide empathy and distress scores. The overall model architecture is detailed in Figure 2.

[1]We used ChatGPT version *gpt-3.5-turbo-1106* in our work.

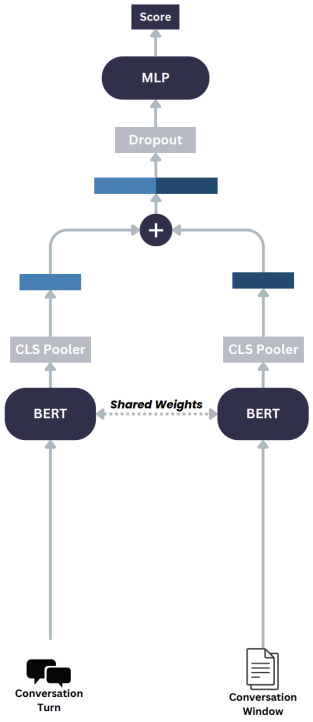

Figure 1: CONV-T Architecture

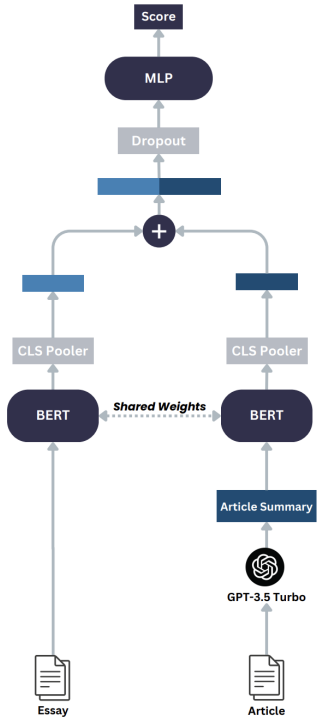

Figure 2: EMP Architecture

4 Experiments and Results

In this section, due to page limitations, we present the experimental setup and results of our study on essay-level empathy and distress detection using context-aware BERT-based models only. We aim to evaluate the effectiveness of our proposed architecture by examining the impact of layer manipulation and the type of context provided on the model's performance. We trained both models using a learning rate of 5e-4, an Adam epsilon of 1e-3, and a batch size of 16. Additionally, for the conversation-level model, we used a window size of 1.

4.1 Experimental Setup

BERT consists of 12 encoder layers. To determine the optimal configuration for empathy and distress detection, we tested three distinct model configurations regarding the BERT layers:

- **No layers frozen**: All layers of BERT model are trainable during the fine-tuning process.
- **Last two layers unfrozen**: Only the last two layers of the BERT model are trainable, while the remaining layers are kept frozen.
- **Last four layers unfrozen**: Similar to the previous configuration, but with the last four layers of the BERT model being trainable.

Additionally, we experimented with various types of context inputs to identify the most effective approach:

- **Emotion only**: The model receives only the emotional characteristics of the input text as context. We used a pretrained model from Hugging Face called roberta-base-go_emotions Lowe (2024) to extract the emotional labels of the sentences.
- **Article only**: The model receives the summarized article text as context, which provides relevant background information for the essay.
- **Combination of emotion and article**: Both emotional characteristics and summarized article text are provided as context.
- **No context**: The model receives no additional context, relying solely on the input text for empathy and distress detection.

These configurations were chosen to explore how different levels of fine-tuning and context types affect the model's ability to accurately detect empathy and distress.

We also conducted an experiment to measure the effectiveness of using two parallel BERT architectures similar to Siamese Networks and providing the input using the '[SEP]' token. This approach

Model Configuration	No Context	Context as Emotion	Context as Article	Context as Article and Emotion
No layers frozen	0.677	0.665	**0.698**	0.677
Last two layers unfrozen	0.607	0.615	**0.663**	0.634
Last four layers unfrozen	0.632	0.635	0.641	**0.652**

Table 1: The table shows the performance of different model configurations and context types in terms of Pearson correlations for empathy and distress detection.

Input Configuration	Pearson Correlation
Parallel BERT (Independent Inputs)	**0.698**
Inputs separated with [SEP] Token	0.671

Table 2: The table shows the Pearson correlation scores for essay-level empathy detection using Independent inputs and BERT with [SEP] token configurations.

CONV-T Results	Pearson Correlation
Emotion Intensity	0.743
Emotional Polarity	0.758
Empathy	0.706

Table 3: The table shows the Pearson correlation scores for CONV-T track.

aims to assess whether concatenating context and input text with '[SEP]' improves performance compared to processing them independently. All experiments use an evaluation dataset derived from WASSA2024's training data.

4.2 Results

Our findings are summarized in Table 1, which shows the performance of each configuration in terms of Pearson correlation for empathy and distress detection.

As depicted in Table 1, using the article alone as context yielded the best performance. This configuration outperformed the others, highlighting the significant impact of relevant textual context on the model's ability to accurately detect empathy and distress in essays. The results demonstrate that providing context as a separate input, rather than concatenating it with the primary text, significantly improves model performance.

Table 2 compares the performance of the parallel BERT architecture (independent input) with BERT using the '[SEP]' token to concatenate context and input text. The results indicate that the parallel BERT approach yields a higher Pearson correlation score, suggesting that processing context and input text independently is more effective than concatenating them with '[SEP]'.

4.3 Discussion

The experiments underscore the critical role of tailored model architecture and context selection in enhancing empathy and distress detection performance. Specifically, the use of article summaries as context proved to be the most effective, as illustrated in Table 1. This approach allows the model to better utilize its attention mechanism, enhancing its understanding of empathy.

Additionally, our results show that the performance improves as we increase the number of unfrozen layers, suggesting that deeper fine-tuning allows the model to better capture the nuances of empathy and distress. For completeness, we presented CONV-T results are in Table 3.

Comparing parallel BERT to BERT with the [SEP] token shows that processing context and input text independently improves performance. This suggests the model benefits from treating context and primary text as distinct inputs rather than concatenating them.

5 Conclusion

In this paper, we detailed the experimental setup and results, providing a comprehensive analysis of the impact of different configurations and context types on empathy and distress detection performance. Our findings illustrate that context-awareness is a key factor in accurately detecting empathy in textual data. By effectively leveraging contextual information through a separate input strategy, our model demonstrated improved performance over concatenating inputs with the '[SEP]' token. This research opens the door for further developments in enhancing the sensitivity and accuracy of empathy detection systems. Future work can explore additional dimensions of context and refine the methods of integrating context to further boost performance in real-world applications.

Limitations

While our study demonstrates the potential of a context-aware BERT-based model for empathy and distress detection, several limitations must be acknowledged. First, the relatively small size and domain-specific nature of the available datasets constrain the generalizability of our findings to real-life applications, necessitating larger and more diverse datasets for validation. Additionally, our approach's reliance on the presence of context, such as article summaries for essay-level inputs, means that the model's performance is influenced by the availability and quality of this contextual information. In scenarios where context is absent, the model may face challenges in achieving similar levels of accuracy. Therefore, it is important to generalize the concept of context to enhance the model's applicability in diverse real-life situations. Developing methods to effectively incorporate and optimize contextual information will be crucial for future improvements and broader applicability.

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, Toronto, Canada. Association for Computational Linguistics.

Tanmay Chavan, Kshitij Deshpande, and Sheetal Sonawane. 2023. Empathy and distress detection using ensembles of transformer models. *Preprint*, arXiv:2312.02578.

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*.

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*.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. EmpathBERT: A BERT-based framework for demographic-aware empathy prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3072–3079, Online. Association for Computational Linguistics.

Sam Lowe. 2024. roberta-base-go_emotions. https://huggingface.co/SamLowe/roberta-base-go_emotions.

Xin Lu, Zhuojun Li, Yanpeng Tong, Yanyan Zhao, and Bing Qin. 2023. HIT-SCIR at WASSA 2023: Empathy and emotion analysis at the utterance-level and the essay-level. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 574–580, Toronto, Canada. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Yukun Wang, Jin Wang, and Xuejie Zhang. 2023. YNU-HPCC at WASSA-2023 shared task 1: Large-scale language model with LoRA fine-tuning for empathy detection and emotion classification. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 526–530, Toronto, Canada. Association for Computational Linguistics.

Table 4: Performance of the Empathify team on the CONV-T and EMP tracks on the WASSA2024 test set. Numbers represent Pearson correlation scores.

Track	Empathy (r)	Emotion Polarity (r)	Emotion Intensity (r)	Empathy (r)	Distress (r)
CONV-T	0.541	0.638	0.584	-	-
EMP	-	-	-	0.290	0.217

A Appendix

We presented our WASSA2024 competition results in Table 4. Due to page limitations, we included our detailed results in the appendix. This table showcases the performance of the Empathify team on both the CONV-T and EMP tracks, based on the WASSA2024 test set. The numbers represent Pearson correlation scores (r values) for different emotion detection metrics. For the CONV-T track, we report the scores for Empathy, Emotion Polarity, and Emotion Intensity. For the EMP track, we provide the scores for Empathy and Distress.