HIT-SCIR at WASSA 2023: Empathy and Emotion Analysis at the Utterance-Level and the Essay-Level

Xin Lu*, **Zhuojun Li***, **Yanpeng Tong***, **Yanyan Zhao**†, **Bing Qin**
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{xlu, zjli, yptong, yyzhao, qinb}@ir.hit.edu.cn

Abstract

This paper introduces the participation of team HIT-SCIR to the WASSA 2023 Shared Task on Empathy Detection and Emotion Classification and Personality Detection in Interactions. We focus on three tracks: Track 1 (Empathy and Emotion Prediction in Conversations, CONV), Track 2 (Empathy Prediction, EMP) and Track 3 (Emotion Classification, EMO), and designed three different models to address them separately. For Track 1, we designed a direct fine-tuning DeBERTa model for three regression tasks at the utterance-level. For Track 2, we designed a multi-task learning RoBERTa model for two regression tasks at the essay-level. For Track 3, we designed a RoBERTa model with data augmentation for the classification task at the essay-level. Finally, our team ranked 1st in the Track 1 (CONV), 5th in the Track 2 (EMP) and 3rd in the Track 3 (EMO) in the evaluation phase.

1 Introduction

In the field of human-computer interaction systems, a discernible trend is the increased focus on the emotion and empathy status of users and the facilitation of emotional exchanges with them. This approach significantly contributes to enhancing service quality and boosting user satisfaction.

However, analyzing the emotion and empathy status of users is still a challenging problem, which requires researchers to conduct thorough exploration and in-depth study. The WASSA 2023 Shared Task 1 (Barriere et al., 2023) provides a unified evaluation benchmark, on the basis of which we have conducted corresponding work.

We have participated in three of five tracks, which are:

Track 1: Empathy and Emotion Prediction in Conversations (CONV), which consists in predicting the perceived empathy, emotion polarity and emotion intensity at the utterance-level in a dialog.

Track 2: Empathy Prediction (EMP), which consists in predicting both the empathy concern and the personal distress at the essay-level.

Track 3: Emotion Classification (EMO), which consists in predicting the emotion at the essay-level.

We conducted analyses and experiments on these three tracks concurrently. In Section 2, we present the methodologies designed for different tasks, the dataset features used in our design and our ensembling method. In Section 3, we introduce the experimental results of our proposed methods, along with corresponding result analyses. In Section 4, we provide our conclusions and summarize our methodologies. The implementation details can be found in Appendix A.

2 System Description

2.1 Track 1: Empathy and Emotion Prediction in Conversations (CONV)

The training set is initially analyzed, revealing an average dialogue length of 23 turns, with each utterance averaging 18 tokens. More details of how this dataset was designed can be found in Omitaomu et al. (2022). To encode the context information of each utterance effectively, we employ a concatenation approach to encode the dialogue information without significant loss. Given the strong contextual relevance of emotion polarity, emotion intensity, and empathy in dialogues, each turn is assigned a context window, and through comprehensive experimentation, we determine the optimal window size for each metric. Our approach involves direct fine-tuning of the DeBERTa (He et al., 2020) model for regression tasks (more details can be found on Appendix A), resulting in a

* Equal Contribution.
† Email Corresponding.

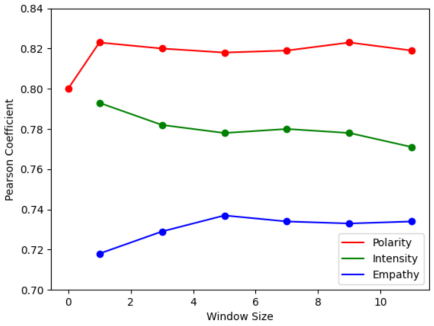


Figure 1: Pearson Correlation Coefficient of different window sizes using the deberta-xl model in the official development dataset.

collection of models that exhibited favorable performance. Then employing a model selection method, unstable models are filtered out, and the remaining models in the collection are ensembled for the final results.

2.1.1 Model Architecture

The model architecture is illustrated in Figure 3a. For a given utterance, denoted as u_k, the corresponding input is constructed as follows.

$$< s > u_{k-w}...u_{k-1} < s > u_k < /s > u_{k+1}...u_{k+w}$$

Here, $u_k = w_1, w_2, ..., w_n$, where n represents the number of tokens in the k-th utterance. The input is fed into the encoder, and the output corresponding to the first <s> token is taken as the contextual representation for the k-th utterance. It is then passed through an MLP to obtain the corresponding output regression value. We apply the same data processing to the validation set. The method of processing the input text without altering the model architecture is quite simple and effective . We adopt DeBERTa as the contextual encoders.

2.1.2 Contextual Window

We conduct extensive experiments on different models and different context window sizes, and the results are shown in Figure 1. It can be observed that for emotion intensity, the trend indicates that the metric decreases as the window size increases. This may be because the expression of emotion intensity is often highly correlated with the expression of the current sentence and does not depend on context too far away. Therefore, we abandon windows larger than 3 for this task. For emotion polarity, we find that the performance is relatively similar for window sizes larger than 0, so we discard the windows with relatively worse results, such as 5 and 11. For empathy, we discard models with window sizes less than 5. All of the results above are achieved by deberta-xl on the official development set. We also train deberta-xxl and select some models based on similar criteria to form a model set χ, which participate in the final model ensemble. We also find that for emotion polarity prediction tasks, models with window sizes greater than 0 are significantly better than those with a window size of 0 (single-sentence prediction). However, when the window size is too large (9 or 11), the metric decrease (as experiment results on xxl proved). This is consistent with the intuition that emotion polarity depends on context but not on irrelevant context.

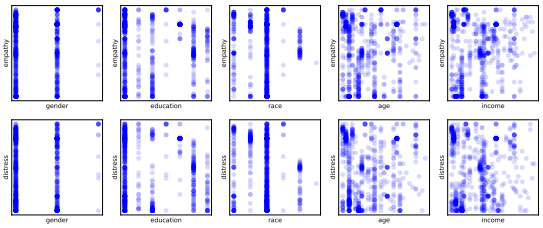


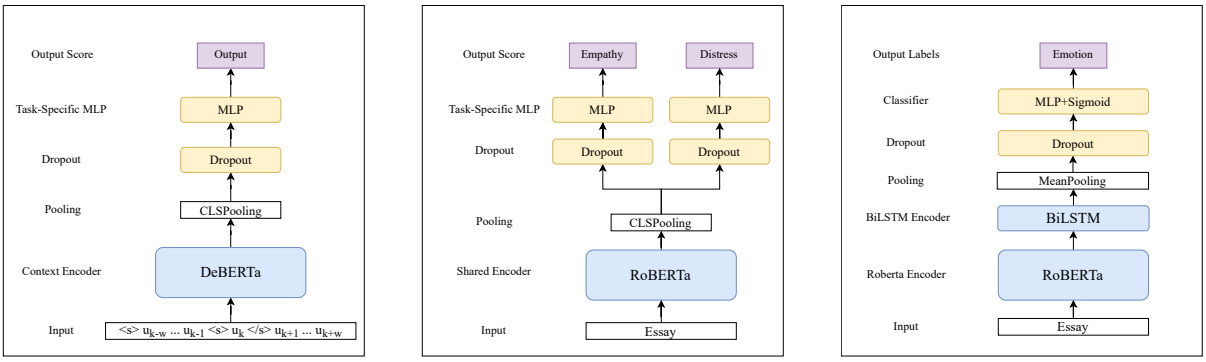
Figure 2: Scatter plot of the bivariate distribution of empathy/distress ratings and demographic features.

2.2 Track 2: Empathy Prediction (EMP)

Initially, we perform a correlation analysis on the train set, examining the relationship between empathy/distress ratings and demographic features. Our findings indicate no significant correlation between the demographic features and empathy/distress ratings. Additionally, building upon Batson's Empathy Theory (Batson et al., 1987) and considering the high Pearson correlation score observed between empathy and distress in a previous study (Buechel et al., 2018), we proceed to investigate the correlation between empathy and distress within the train set. This subsequent analysis reveals a strong correlation between these two variables. Consequently, we employ a multi-task learning approach to effectively model both the empathy and distress subtask.

2.2.1 Data Analysis

In Figure 2, we display the bivariate distribution of empathy/distress ratings and demographic features, indicating a lack of significant correlation between them. Additionally, based on previous researches (Lahnala et al., 2022; Chen et al., 2022; Ghosh et al., 2022), most models have achieved good re-



(a) CONV track model architecture. (b) EMP track model architecture. (c) EMO track model architecture.

Figure 3: Track 1 (CONV), Track 2 (EMP) and Track 3 (EMO) model architectures.

sults without incorporating these features. To avoid introducing more noise into the model, we choose to follow approach by Chen et al. (2022) and construct a model by fine-tuning of the RoBERTa (Liu et al., 2019) model. Inspired by Buechel et al. (2018), we compute the Pearson correlation between empathy and distress in the training set, which results in a high score of 0.63. This finding suggests that a multi-task learning approach, which simultaneously models both empathy and distress, is a suitable choice.

2.2.2 Multi-task Learning Model

In Figure 3b, we select RoBERTa for encoding the essays in the EMP task. To represent the entire sentence, we use the CLS token and applied a single-layer MLP with dropout to predict the empathy and distress outputs. For the multi-task model, we share the RoBERTa encoding layer and equally weight the losses of both subtasks during fine-tuning.

2.3 Track 3: Emotion Classification (EMO)

An initial analysis of the distribution of data labels in the training set reveals a small dataset size and an uneven distribution. To address this issue, we employ data augmentation techniques, attempting various methods, including EDA (Wei and Zou, 2019), GoEmotions (Demszky et al., 2020), and ChatGPT rephrasing. Our experiments ultimately show that ChatGPT rephrasing produce the best results. Additionally, after testing different model architectures, we select a structure that is both effective and robust.

2.3.1 Data Analysis

We analyze the distribution of emotion labels in the training set. The number of instances for the "Fear", "Hope", "Joy", and "Surprise" is significantly lower compared to the other labels. On the other hand, the "Sadness" and "Neutral" labels have a relatively larger number of instances. Additionally, we have computed the distribution of single-label instances for each category, and it is found that the proportion of single-label instances for "Fear", "Hope", "Joy", and "Surprise" is consistently lower than 50%. This indicates that the classification of these labels is prone to be influenced by other labels, posing a significant challenge for modeling them. More detailed statistics are shown in Appendix B.

2.3.2 Data Augmentation with ChatGPT

Developed by OpenAI and released in November 2022, ChatGPT is an artificial intelligence chatbot that achieves strong instruction-following abilities through fine-tuning and reinforcement learning on large language models such as GPT 3.5 and GPT4. Leveraging ChatGPT's powerful language modeling capabilities, our objective is to perform data augmentation on imbalanced samples in order to mitigate the potential biases. We use it to rephrase the original essay for data augmentation, and more details are shown in Appendix C.

To tackle any potential data imbalance and improve our model's performance, we generate over 200 additional instances for each of the categories, except for "Sadness" and "Neutral" due to their relative abundance of data. To ensure that the expanded data did not introduce excessive noise, we apply a sorting process based on ROUGE-L scores and prioritized sentences with higher scores.

2.3.3 Emotion Classification Model

In Figure 3c, we use the RoBERTa model as the essay encoder. The encoded vectors are then processed through a BiLSTM layer to capture long-

Team	Avg	Polarity	Intensity	Empathy
HIT-SCIR	0.758	0.852	0.714	0.708
YNU-HPCC	0.730	0.824	0.693	0.674
Team Hawk	0.725	0.809	0.701	0.665

Table 1: Test dataset results (Pearson correlations) for Track 1 (CONV) in the evaluation phase.

distance word dependencies within the essay. Afterwards, the BiLSTM outputs are averaged, followed by a dropout operation. Finally, an 8-dimensional vector is obtained through a single-layer MLP, using the sigmoid function as the activation function for multi-label classification.

2.4 Ensembling Method

On the official essay-level development set, speaker information such as gender, education level, race, and age are available. Using these attributes, we divide the speakers into 21 groups and then partition the samples in development datasets according to the speaker groups. This results in 21 datasets with different distributions. We consider that a model with strong generalization ability should not have too much variation in performance across these 21 different development subsets. Therefore, we further filter the initial model set χ based on variance and obtained the final model set $\chi\prime$. Using these models, we can further achieve model ensemble. Especially, for regression models, we directly average the regression values output by each model in the set to obtain the ensembled regression value. The division details can be found in Appendix D.

3 Results and Discussions

3.1 Results for Track 1 (CONV)

The results presented in Table 1 indicate that our final ensembled model achieved the top rank on the official test set. Specifically, our model outperforms the second-ranked model by almost 3 points in predicting emotional polarity and empathy, and by 2 points in predicting intensity. This remarkable performance demonstrates the superior generalization ability of our final ensembled model on the test set, which can be attributed to our effective model ensemble strategy and our context window selection.

3.2 Results for Track 2 (EMP)

Table 2 presents the results of our systems on the test set of the EMP task. We also provide our results on the dev set, where it outperformed all known results. However, we observe a significant drop in performance on the test set for both our systems and note that other teams experience similar performance drops. This suggests the data distribution between the dev and test sets may differ significantly, leading to overfitting to the dev set and poor generalization performance on the test set.

Team	Average	Empathy	Distress
NCUEE-NLP	0.4178	0.4150	0.4206
CAISA	0.3838	0.3478	0.4197
earendil	0.3462	0.3585	0.3339
zex	0.3420	0.2933	0.3906
HIT-SCIR	0.3416	0.3287	0.3545
HIT-SCIR (Dev set)	0.6571	0.6662	0.6480

Table 2: Test dataset results (Pearson correlations) for Track 2 (EMP) in the evaluation phase.

Team	Macro F1	Micro F1
adityapatkar	0.701	0.750
Bias Busters	0.647	0.700
HIT-SCIR	0.644	0.720

Table 3: Test dataset results (Macro F1 & Micro F1) for Track 3 (EMO) in the evaluation phase.

3.3 Results for Track 3 (EMO)

Table 3 presents the test results of Top-3 systems in this Task, and our system ranks 3rd. Upon analyzing the error logs in Codalab, we find that none of the instances in the test set were labeled as "Hope", "Joy", or "Surprise", which are precisely the three least represented labels in the training set. We hypothesize that our model introduce a trade-off between the underrepresented and overrepresented categories, which may have led to the slight decrease in performance on the test set. More ablation studies can be found in Appendix E.

4 Conclusion

Our team HIT-SCIR participated in the WASSA 2023 Shared Task on Empathy Detection and Emotion Classification and Personality Detection in Interactions. We focused on empathy and emotion analysis and participated in three of five tracks. We analyzed the features of each task and designed different methodologies for them. Finally, our team ranked 1st in the Track 1 (CONV), 5th in the Track 2 (EMP) and 3rd in the Track 3 (EMO) in the evaluation phase.

References

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Salvatore Giorgi. 2023. Wassa 2023 shared task: Predicting empathy, emotion and personality in interactions and reaction to news stories. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*.

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 *EMNLP 2018*.

Yue Chen, Yingnan Ju, and Sandra Kübler. 2022. IUCL at WASSA 2022 shared task: A text-only approach to empathy and emotion detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 228–232, Dublin, Ireland. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Soumitra Ghosh, Dhirendra Maurya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Team IITP-AINLPML at WASSA 2022: Empathy detection, emotion classification and personality detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 255–260, Dublin, Ireland. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Allison Lahnala, Charles Welch, and Lucie Flek. 2022. CAISA at WASSA 2022: Adapter-tuning for empathy prediction. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 280–285, Dublin, Ireland. Association for Computational Linguistics.

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*.

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. *arXiv preprint arXiv:2205.12698*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

A Implementation Details

We train the model using the Pytorch (Paszke et al., 2019) on the NVIDIA A100 GPU and use the hugging-face (Wolf et al., 2020) framework. The optimizer used for model training is AdamW (Loshchilov and Hutter, 2017) optimizer which is a fixed version of Adam (Kingma and Ba, 2014) with weight decay, and set β_1 to 0.9, β_2 to 0.99 for the optimizer. All experiments select the best parameters in the valid set and then use the ensembling

method in 2.4 . Below is the details to the custom parameter settings for different tracks.

For CONV track, we use both the DeBERTa-v2-xl and DeBERTa-v2-xxl as our pre-trained models. For each metric of this track (polarity, intensity and empathy), we fine-tune on both models. Specifically, during the process of choosing the best window size, for each window size, we fine-tune on both models. The DeBERTa model comes with 24(48) layers and a hidden size of 1536. The total parameters are 900M(1.5B), and it is trained with 160GB raw data. DeBERTa improves the BERT and RoBERTa models using two novel techniques, the disentangled attention mechanism and the enhanced mask decoder. We use the learning rate {3e-6, 4e-6}, total training batch size 32, training epoch 6 for DeBERTa-xl and DeBERTa-xxl separately. We conduct distributed training on 4 NVIDIA A100-80GB GPUs and use fp16 training. We set the maximum length of 1024, and delete the excess.

For EMP track, we use the roberta-base as our pre-trained model, and fine-tune the model. RoBERTa (Liu et al., 2019) extends BERT (Devlin et al., 2019) by incorporating techniques like dynamic masking and removing the next sentence prediction pre-training objective. We conduct grid search with the learning rate varying in {1e-5, 2e-5, 3e-5}, batch size varying in {8, 16, 32}, and dropout rate varying in {0, 0.3}. We set the maximum length of 512, and delete the excess.

For EMO track, we use the roberta-large as our pre-trained model, and fine-tune the model. We conduct grid search with the same parameter search range as that for the EMP track. Additionally, the hidden dimension of the BiLSTM layer is 256.

B Data Statistics for Track 3 (EMO)

Table 4 presents the distribution of emotion labels in the training set.

Emotion	All Instances	Single-Label Instances
Anger	124	67
Disgust	100	44
Fear	33	10
Hope	32	10
Joy	10	5
Neutral	240	202
Sadness	383	297
Surprise	19	9

Table 4: Data distribution over emotion classes in the origin train dataset. "Single label instances" refers to the number of data instances that contain only one label.

C Our Prompt for ChatGPT

We use the following text as a prompt to provide to ChatGPT in order to rephrase the original essay for data augmentation.

> Prompt: You are a helpful assistant that rephrase text and make sentence smooth. Besides, you should keep the emotion in the text unchanged. Please rephrase the following text, it's written by participants after reading news articles where there is harm to a person, group or other. Pay attention to retain emotion of `{emotions}` in the source text, keep the word count in 300-800 characters. Text: `{content}`

The `{content}` field pertains to the essay in the training set that requires rephrasing, and `{emotions}` represents the emotion labels associated with that essay.

D Division Details

The division rules are shown in Table 5.

Attribution	Set1	Set2	Set3	Set4
age	<30	other	-	-
education	<5	other	-	-
income	<35000	other	-	-
race	1	2	3	5
gender	1	2	-	-

Table 5: Rules used to divide speakers into 21 groups.

E Ablation Study for Track 3 (EMO)

Table 6 shows the results of the ablation study on the dev set of the EMO Task. We use several different settings to demonstrate the effectiveness of our proposed methods. The ensembling strategy significantly improve the performance of our system, and adding BiLSTM and data augmentation methods also contributed to the improvement of the model's performance. Moreover, our final results exceed all known results on the dev set, but show a slight decrease on the test set.

Methods	Macro F1
RoBEERTa-large finetune	0.5798
+ BiLSTM	0.6117
+ BiLSTM + data augmentation	0.6178
+ BiLSTM + data augmentation + ensemble	0.6630

Table 6: Ablation study on dev set of Track 3 (EMO).