hyy33 at WASSA 2024 Empathy and Personality Shared Task: Using the CombinedLoss and FGM for Enhancing BERT-based Models in Emotion and Empathy Prediction from Conversation Turns

Huiyu Yang, Liting Huang, Tian Li, Nicolay Rusnachenko, Huizhi Liang*
Newcastle University, Newcastle Upon Tyne, England
{huiyu.yang33, huangliting2019, litianricardolee, rusnicolay}@gmail.com,
huizhi.liang@newcastle.ac.uk

Abstract

This paper presents our participation to the WASSA 2024 Shared Task on Empathy Detection and Emotion Classification and Personality Detection in Interactions. We focus on Track 2: Empathy and Emotion Prediction in Conversations Turns (CONV-turn), which consists of predicting the perceived empathy, emotion polarity and emotion intensity at turn level in a conversation. In the method, we conduct BERT and DeBERTa based finetuning, implement the CombinedLoss which consists of a structured contrastive loss and Pearson loss, adopt adversarial training using Fast Gradient Method (FGM). This method achieved Pearson correlation of 0.581 for *Emotion*, 0.644 for *Emotional Polarity* and 0.544 for *Empathy* on the test set, with the average value of 0.590 which ranked 4th among all teams. After submission to WASSA 2024 competition, we further introduced the segmented mix-up for data augmentation, boosting for ensemble and regression experiments, which yield even better results: 0.6521 for *Emotion*, 0.7376 for *Emotional Polarity*, 0.6326 for *Empathy* in Pearson correlation on the development set. The implementation and fine-tuned models are publicly-available at https://github.com/hyy-33/hyy33-WASSA-2024-Track-2.

1 Introduction

Emotion detection and empathy analysis are important and inevitable topics in the processing of human interactions, which show great potential in various application scenarios (Nandwani and Verma, 2021; Sharma et al., 2020). To provide more insights into this topic, WASSA organizes workshop on related topics each year (Barriere et al., 2023). This year, WASSA 2024 focuses on Shared Task on Empathy Detection and Emotion Classification and Personality Detection in Interactions (Giorgi et al., 2024).

*The corresponding author.

In this paper, we propose a solution towards Track 2: Empathy and Emotion Prediction in Conversations Turns (CONV-turn). In this task, participants are given conversations between two users that read the same essay, which contains reaction to news articles where there is harm to a person or group (Omitaomu et al., 2022). Each of their conversation turn (text content) has been annotated in perceived empathy, emotion polarity, and emotion intensity. Other meta information such as *article_id*, *conversation_id*, *turn_id* and *speaker_id* are also provided. A sample from the dataset is demonstrated in Figure 1.

A Training Sample from Track 2

Text:
I can't imagine just living in an area that is constantly being ravaged by hurricanes or earthquakes. I take my location for granted.

Label:
Emotion: 3
EmotionalPolarity: 2
Empathy: 4.6667
SelfDisclosure: 3.3333

Other meta information:
id: 3, article_id: 35, conversation_id: 1,
turn_id: 3, speaker: "Person 2",
person_id_1: "p019", person_id_2: "p012"

Figure 1: A Data Sample from Track 2

This task aims at developing appropriate methods to predict the perceived empathy, emotion polarity, and emotion intensity at the speech-turn-level during human conversation. In previous works, BERT (Devlin et al., 2019) is frequently used for emotion classification (Luo and Wang, 2019; Kannan and Kothamasu, 2022), its variations such as RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020) also play important roles in empathy prediction and sentiment analysis (Vasava et al., 2022; Lu et al., 2023). Based on fine-tuned encoders, different strategies are further introduced

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 430–434

to build more robust and reliable models, including adversarial training (Chen and Ji, 2022; Chang et al., 2023), data augmentation (Kwon and Lee, 2023) and ensemble strategy (Plaza-del Arco et al., 2022).

Our goal is to predict the interior emotion and empathy state of the user according to turn-level information from human-to-human conversations. To achieve this goal, we adopt BERT-based models including BERT (Devlin et al., 2019) and its variation of DeBERTa (He et al., 2020). Then, they are fine-tuned using task-oriented data from Track 2 with adversarial training using Fast Gradient Method (FGM). Also, we design a novel CombinedLoss, which consist of a structured contrastive loss and a Pearson loss. Then, after the submission to WASSA 2024 competition, data augmentation using the segmented mix-up strategy, ensemble with boosting method and regression experiments are further conducted.

2 Methodology

This section introduces the methodolgy of our proposed system for Track 2 in WASSA 2024. As in Figure 2, the proposed model includes: the fine-tuned BERT or DeBERTa, the CombinedLoss and the downstream head for classification (or regression). Also, augmentation and ensemble are implemented in corresponding parts.

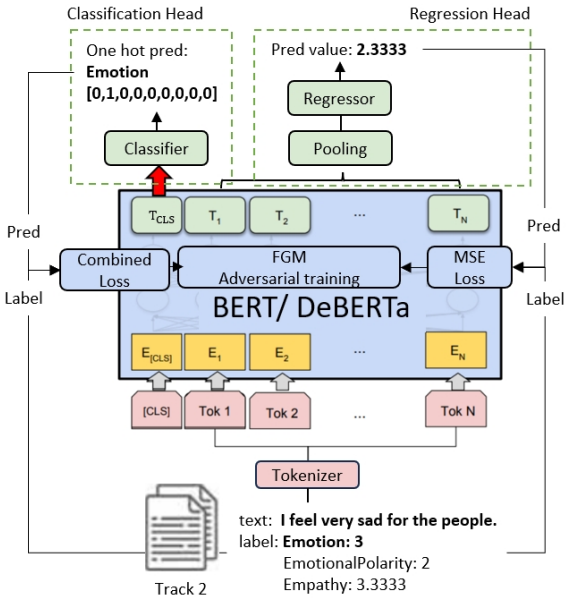

Figure 2: The proposed model

2.1 Fine-tuned BERT and DeBERTa

In this paper, we conduct task-oriented fine-tuning for Track2 based on BERT (Devlin et al., 2019) and its variation of DeBERTa (He et al., 2020). For the base models, *bert-base-uncased* and *deberta-base* are introduced as the pretrained language models. Then, fine-tuning is conducted on the training set of Track 2, so that the encoders could adapt from general language modelling to our specific prediction task by adjusting the parameters.

2.2 The CombinedLoss

Different from commonly-used loss functions, our work proposes the CombinedLoss, which could be expressed as:

$$L_{\text{total}} = L_{\text{loss}} + \lambda(1 - Corr_{Pear}(\hat{\mathbf{y}}, \mathbf{y})), \quad (1)$$

where L_{loss} is the structured contrastive loss for classification, λ is the regularization coefficient, and $Corr_{Pear}(\hat{\mathbf{y}}, \mathbf{y})$ is the Pearson correlation coefficient (Cohen et al., 2009) between the prediction of $\hat{\mathbf{y}}$ and the ground truth label of $\mathbf{y}$.

Since emotions are classified into multiple levels in the dataset (Omitaomu et al., 2022), the Pearson correlation coefficient is used as a regularization term in the loss function. By using the negative Pearson coefficient, this loss function aims to capture the subtle scale between emotion levels.

2.3 Adversarial Training with FGM

To improve the robustness and the generalization ability of the proposed model, adversarial training is introduced as follows:

$$Obj = \min_{\theta} E(x, y) \left[\max L(f_\theta(x + \delta), y)\right], \quad (2)$$

in which x is the input sample, δ is the added perturbation for adversarial training, f_θ is neural network function with θ as parameters. By maximizing $L(f_\theta(x + \delta))$, the most disturbing perturbation are introduced to the model, then the model is optimized to minimize the training error, which helps it to be robust to potential perturbations.

In this work, Fast Gradient Method (FGM) (Andriushchenko and Flammarion, 2020) is implemented as adversarial training strategy, which computes the most disturbing perturbation through scaling the gradient as below.

$$\delta = \epsilon \cdot \frac{g}{||g||_2} \quad (3)$$

$$g = \nabla_x L(x, y, \theta) \quad (4)$$

2.4 Augmentation: the Segmented Mix-up

To improve the generalization ability of models, mix-up is often used as a method for data augmentation. In this work, a segmented mix-up is proposed, which mixes inputs and labels within specific label ranges. This segmentation is essential because simple mix-up (Gong et al., 2022) between highly negative and highly positive samples could not generate meaningful intermediate samples.

For each dimension, e.g. *Emotion*, samples are divided into two segments: the lower segment with labels smaller or equal to the middle label, and the upper one with labels larger than the middle label. Each sample (x_i, y_i) is paired with a partner sample (x_j, y_j) from the same label segment, with x_i and x_j denote the tokenized sentences, and y_i and y_j represent their labels. The mix-up coefficient μ is sampled from a Beta distribution: $\mu \sim \text{Beta}(\alpha, \alpha)$, where α controls the mix-up strength. The generated inputs and labels are computed as:

$$\tilde{x}_i = \mu x_i + (1-\mu)x_j, \tag{5}$$

$$\tilde{y}_i = \mu y_i + (1-\mu)y_j, \tag{6}$$

2.5 Ensemble with Boosting

To build a more accurate and robust system, boosting is implemented as an ensemble strategy (Bühlmann, 2012), which combines fine-tuned BERT and DeBERTa models. In order to enhance the overall performance, weights are assigned according to the accuracy of each model on the development set. Through this, it is ensured that the model with the most reliable prediction has the greatest impact on the final output.

3 Experiments and Results

In this section, extensive experiments were conducted on the fine-tuned BERT and DeBERTa. Also, ablation study is performed to test the performance of different parts in the proposed model.

3.1 Datasets

The dataset of Track 2 includes a training set of 11,166 samples, a development set of 990 samples and a test set of 2,061 valid samples (Omitaomu et al., 2022). Each sample consists of the text content of a single dialogue turn and the corresponding label of *Emotion*, *Emotional Polarity* and *Empathy*, as well as some meta information of the speakers and the conversation. A data sample is shown in Figure 1.

Model	Loss	FGM	Emo	EmoP	Emp	Avg
BERT	Cross-entropy	No	0.5867	0.6824	0.5703	0.6131
BERT	CombinedLoss	No	0.5921	0.6836	0.5803	0.6187
BERT	CombinedLoss	Yes	**0.6142**	**0.6899**	**0.5852**	**0.6298**
DeBERTa	Cross-entropy	No	0.6255	0.7281	0.5918	0.6485
DeBERTa	CombinedLoss	No	0.6348	0.7364	0.6042	0.6585
DeBERTa	CombinedLoss	Yes	**0.6399**	**0.7366**	**0.6064**	**0.6610**

Table 1: Pearson correlation of fine-tuned models with CombinedLoss and FGM on the development set

3.2 Evaluation Metrics

To test the performance of the proposed solution, the official evaluation metric for Track 2 is the Pearson correlation (Cohen et al., 2009). Given sequences of prediction $\hat{y}$ and ground truth y, their Pearson correlation coefficient can be calculated as:

$$Corr_P(\hat{y}, y) = \frac{\sum_{i=1}^{n}\left(\frac{(\hat{y}_i-\bar{\hat{y}})}{\sigma_{\hat{y}}}\frac{(y_i-\bar{y})}{\sigma_y}\right)}{n}, \tag{7}$$

in which $E(\hat{y})$ and $E(y)$ stand for the expectations of $\hat{y}$ and y, $\sigma_{\hat{y}}$ and σ_y stand for the standard deviations of $\hat{y}$ and y.

3.3 Implementation Details

Baselines. To compare the performance of proposed models, BERT (Devlin et al., 2019) and its variation DeBERTa (He et al., 2021) are introduced. For BERT, *bert-base-uncased* is used, with 12 encoder layers and 110M parameters. For DeBERTa, *deberta-base* is adopted with 390M parameters.

Hyper-parameters. For tokenization, input sentences are tokenized with *BertTokenizer* and *DebertaTokenizer* with the maximum length of 128. For optimization, AdamW optimizer is adopted with learning rate of 1×10^{-6} and exponential decay with $\gamma = 0.99$ after grid search, the batch size is 400 for fine-tuning BERT and 200 for fine-tuning DeBERTa. For the segmented mix-up, $\alpha = 0.2$ is used. Other details could be found in our implementation.

Labels and Categories. The experiments are conducted on different downstream tasks of classification and regression. Because the original labels not only contain integer values but also include float values, such as 0.3333, 0.6667 in the training set and 0.5, 1.5 in the development set, we manually divided 9 categories for *Emotion*, 5 categories for *Emotional Polarity* and 11 categories for *Empathy* in classification (details could be found in our code of implementation). In regression experiments, we directly use original labels as target values.

Model	Ensemble	Augment	Emo	EmoP	Emp	Avg
BERT	Boosting	No	0.6521	0.7045	0.6069	0.6545
DeBERTa	Boosting	No	0.6470	0.7215	0.6112	0.6599
BERT, DeBERTa	Boosting	No	0.6485	0.7253	0.6140	0.6626
BERT, DeBERTa	Boosting	Mix-up	**0.6521**	**0.7334**	**0.6326**	**0.6727**

Table 2: Pearson correlation of fine-tuned models with ensemble and augmentation on the development set

Model	Task	Emo	EmoP	Emp	Avg
DeBERTa	Classification	0.6399	0.7366	0.6064	0.6610
DeBERTa	Regression	**0.6409**	**0.7376**	**0.6105**	**0.6630**

Table 3: Pearson correlation of fine-tuned DeBERTa (with CombinedLoss and FGM) in different downstream tasks on the development set

3.4 Results and Analysis

This section presents the results of Pearson correlation based on the experiments of the proposed models on the development set, and conducts analysis for the results.

Fine-tuned BERT and DeBERTa. It can be observed from Table 1 that the average results of fine-tuned DeBERTa is better than fine-tuned BERT, which shows the stronger ability of DeBERTa-based solution. And by implementing the CombinedLoss, both models demonstrate performance gain in *Emotion*, *Emotional Polarity* and *Empathy* prediction. Also, adding adversarial training using Fast Gradient Method (FGM) brings better overall performance, proving its contribution to the robustness and generalization ability of models. Our submission to WASSA 2024 competition is based on this fine-tuned DeBERTa with CombinedLoss and Fast Gradient Method (FGM).

Ensemble and Augmentation. The results of Table 2 show the combined boosting yields the best overall result, which confirms the effectiveness of our boosting strategy by assigning weights to models according to their accuracy. An interesting finding is that ensembling fine-tuned DeBERTas not always achieves the highest score in single dimension, this may due to the reason that single DeBERTa already achieves its upper limit, combining them only decreases the possible lower bound, while on the other hand, single BERT may has unstable scoring performance, thus the ensemble of BERTs leads to high reliablity and better results. Also, augmentation brings further improvement, indicating our segmented mix-up strategy successfully generates meaningful intermediate samples, which contribute to the fine-tuning process.

Classification and Regression. Table 3 presents the results of the fine-tuned DeBERTa (with CombinedLoss and FGM) in different downsteam tasks. The labelling details for classification and regression could be found in Section 3.3. From the results, it is shown that the fined-tuned DeBERTa achieved slightly better performance in regression task, which provides future research direction for us.

4 Conclusions

This paper presents our solution for Track 2 in WASSA 2024, which focused on the prediction of *Emotion*, *Emotional Polarity* and *Empathy* using turn-level information from user conversations. The submitted solution is built using fine-tuned DeBERTa with our proposed CombinedLoss and adversarial training startegy using Fast Gradient Method (FGM), which achieved Pearson correlation of 0.581 for *Emotion*, 0.644 for *Emotional Polarity* and 0.544 for *Empathy* on the test set, with the average value of 0.590 which ranked 4th among all teams. After the submission to WASSA 2024 competition, ensemble strategy using boosting method and data augmentation with the segmented mix-up are implemented, which further improve the performance of our model and yield better results: 0.6521 for *Emotion*, 0.7376 for *Emotional Polarity*, 0.6326 for *Empathy* in Pearson correlation on the development set. In the future, we plan to introduce larger datasets for model retraining at earlier stage (e.g. the Masked Language Model) for better domain adaptation, and consider introducing conversational context and speaker personality for better model construction. Also, the performance of such models in downstream regression tasks will be further investigated.

Limitations

The limitations of the proposed work included: 1) The training set was relatively small with less than 12000 samples. Fine-tuning the models on larger datasets might improve the performance. 2) The labels in the training set and the development set was mis-matched. For instance, the development set contained *Emotion* labels of 0.5, 1.5 and 2.5, which were not presented in the training set. If the test set had similar patterns, then, the inconsistent labels between training and testing could cause degradation of the fine-tuned models.

References

Maksym Andriushchenko and Nicolas Flammarion. 2020. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059.

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.

Peter Bühlmann. 2012. Bagging, boosting and ensemble methods. *Handbook of computational statistics: Concepts and methods*, pages 985–1022.

Yu Chang, Yuxi Chen, and Yanru Zhang. 2023. nicenlp at semeval-2023 task 10: Dual model alternate pseudo-labeling improves your predictions. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 307–311.

Hanjie Chen and Yangfeng Ji. 2022. Adversarial training for improving model robustness? look at both prediction and interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10463–10472.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

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.

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*.

Xiaokang Gong, Wenhao Ying, Shan Zhong, and Shengrong Gong. 2022. Text sentiment analysis based on transformer and augmentation. *Frontiers in Psychology*, 13:906061.

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Eswariah Kannan and Lakshmi Anusha Kothamasu. 2022. Fine-tuning bert based approach for multi-class sentiment analysis on twitter emotion data. *Ingénierie des Systèmes d'Information*, 27(1).

Soonki Kwon and Younghoon Lee. 2023. Explainability-based mix-up approach for text data augmentation. *ACM transactions on knowledge discovery from data*, 17(1):1–14.

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*.

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.

Linkai Luo and Yue Wang. 2019. Emotionx-hsu: Adopting pre-trained bert for emotion classification. *arXiv preprint arXiv:1907.09669*.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81.

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.

Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural language inference prompts for zero-shot emotion classification in text across corpora. *arXiv preprint arXiv:2209.06701*.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Himil Vasava, Pramegh Uikey, Gaurav Wasnik, and Raksha Sharma. 2022. Transformer-based architecture for empathy prediction and emotion classification. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 261–264.