Zhenmei at WASSA-2024 Empathy and Personality Shared Track: Incorporating Pearson Correlation Coefficient as a Regularization Term for Enhanced Empathy and Emotion Prediction in Conversational Turns

Liting Huang
Guangzhou Xinhua College, China
huangliting2019@gmail.com

Huizhi Liang
University of Newcastle, UK
Huizhi.Liang@newcastle.ac.uk

Abstract

In the realm of conversational empathy and emotion prediction, emotions are frequently categorized into multiple levels. This study seeks to enhance the performance of emotion prediction models by incorporating the Pearson correlation coefficient as a regularization term within the loss function. This regularization approach ensures closer alignment between predicted and actual emotion levels, mitigating extreme predictions and resulting in smoother and more consistent outputs. Such outputs are essential for capturing the subtle transitions between continuous emotion levels. Through experimental comparisons between models with and without Pearson regularization, our findings demonstrate that integrating the Pearson correlation coefficient significantly boosts model performance, yielding higher correlation scores and more accurate predictions. Our system officially ranked 9th at the Track 2: CONV-turn. The code for our model can be found at Link [1].

1 Introduction

Accurately predicting emotions is crucial for creating responsive and empathetic conversational systems. Emotions are typically classified into multiple levels, reflecting their nuanced and continuous nature. Many existing approaches focus on minimizing prediction errors but often overlook the linear relationship between predicted and actual emotion levels, resulting in extreme and unstable predictions (Acheampong et al., 2021; Zhou et al., 2024; Creanga and Dinu, 2024).

To address this, we propose enhancing emotion prediction models by incorporating the Pearson correlation coefficient as a regularization term in the loss function. The Pearson correlation measures the linear correlation between predicted and actual emotion levels. By including this metric, our approach aims to generate predictions that minimize absolute error while maintaining high correlation with actual emotions. Theoretical analysis confirms the differentiability and convergence of the proposed loss function, ensuring a stable and reliable optimization process.

Additionally, we introduce Consistent-Mixup data augmentation(CMDA) and boosting techniques to further improve model performance. CMDA creates new training samples by combining data from different classes, enhancing the model's ability to generalize. Boosting leverages the strengths of multiple models, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020), by combining their predictions based on individual accuracies, thereby improving overall accuracy.

We conducted experiments to validate our approach, comparing models trained with and without Pearson regularization, and those enhanced with CMDA and boosting. Results show that incorporating the Pearson correlation coefficient significantly improves performance, yielding higher correlation scores and more accurate emotion predictions. Furthermore, combining CMDA and boosting techniques leads to even greater improvements in model effectiveness.

2 Related Work

Recent research has explored various fine-tuning strategies for Transformer-based models like BERT, RoBERTa, and DeBERTa to enhance downstream performance. Sun et al. (2019) demonstrated significant improvements with techniques such as layer-wise learning rate decay and data augmentation. Mosbach et al. (2020) provided insights into stable fine-tuning through learning rate schedules and early stopping. Additionally, Dong et al. (2019) proposed a unified pre-training framework for lan-

[1] https://github.com/gongziruo/Empathy-and-Emotion-Prediction-in-Conversations-Turns-CONV-turn

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 399–403
August 15, 2024 ©2024 Association for Computational Linguistics

guage understanding and generation, while Gao et al. (2023) introduced progressive module training to incrementally fine-tune models, enhancing performance and stability.

In dialogue systems, Transformer models have been effectively applied to emotion prediction (Acheampong et al., 2021; Vazquez-Rodriguez et al., 2022). Tu et al. (2022) improved emotion recognition by leveraging context-aware embeddings and fine-tuning on emotion-labeled dialogue datasets. The WASSA 2023 shared task further explored empathy, emotion, and personality detection in conversations and reactions to news articles, highlighting the challenges and advancements in this domain (Barriere et al., 2023; Giorgi et al., 2024).

While Pearson correlation regularization remains underexplored, other methods like adversarial training (Liu et al., 2020) have been examined to enhance model robustness by adding input perturbations. These studies underscore the evolving fine-tuning methodologies for Transformer models, showcasing strategies such as layer-wise learning rate decay, context-aware embeddings, adversarial training, and progressive module training to enhance performance and stability in NLP tasks.

3 Methodology

3.1 Pearson Coefficient as Regularization Term

To incorporate the negative Pearson coefficient as a regularization term in the loss function, the total loss can be expressed as:

$$L_{\text{total}} = L_{\text{com}} + \lambda(1 - \rho(\hat{\mathbf{y}}, \mathbf{y})), \tag{1}$$

where λ is the regularization coefficient, and $\rho(\hat{\mathbf{y}}, \mathbf{y})$ represents the Pearson correlation between predictions $\hat{\mathbf{y}}$ and true labels $\mathbf{y}$.

The combined loss L_{com} is defined as:

$$L_{\text{com}} = aL_{\text{CE}} + \beta \left(-\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} P_{y_i j} \log(p_{ij}) \right). \tag{2}$$

In this equation, L_{CE} stands for Cross-Entropy Loss, a is the weight of L_{CE}, $P_{y_i j}$ indicates the penalty for misclassifying class y_i as class j, p_{ij} is the predicted probability for class j, N is the number of samples, and C represents the number of classes. Each value in the penalty matrix $\mathbf{P}$ is non-negative, with higher penalties assigned for misclassifications between labels that are numerically farther apart.

3.1.1 Differentiability

The Pearson correlation coefficient between two variables $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ is defined as:

$$\rho(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}},$$

where $\bar{\hat{y}}$ and $\bar{y}$ are the sample means of $\hat{\mathbf{y}}$ and $\mathbf{y}$, respectively.

To derive the gradient of the Pearson correlation coefficient, we apply the quotient rule. Let $u = \sum_{i=1}^{n} (\hat{y}_i - \bar{x})(y_i - \bar{y})$ and $v = \sqrt{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}$. Then, the partial derivative of $\rho(\hat{\mathbf{y}}, \mathbf{y})$ with respect to $\hat{y}_i$ is given by:

$$\frac{\partial \rho(\hat{\mathbf{y}}, \mathbf{y})}{\partial \hat{y}_i} = \frac{v \frac{\partial u}{\partial \hat{y}_i} - u \frac{\partial v}{\partial \hat{y}_i}}{v^2}. \tag{3}$$

The partial derivatives of u and v with respect to $\hat{y}_i$ are:

$$\frac{\partial u}{\partial \hat{y}_i} = \frac{1}{n}(y_i - \bar{y}), \tag{4}$$

$$\frac{\partial v}{\partial \hat{y}_i} = \frac{\sigma_{\mathbf{y}}}{n\sigma_{\hat{\mathbf{y}}}}(\hat{y}_i - \bar{\hat{y}}), \tag{5}$$

where $\sigma_{\hat{\mathbf{y}}}$ and $\sigma_{\mathbf{y}}$ are the sample standard deviations of $\hat{\mathbf{y}}$ and $\mathbf{y}$, respectively.

Substituting these partial derivatives into the quotient rule and simplifying, we obtain the final expression for the partial derivative of the Pearson correlation coefficient with respect to $\hat{y}_i$:

$$\frac{\partial \rho(\hat{\mathbf{y}}, \mathbf{y})}{\partial \hat{y}_i} = \frac{1}{n\sigma_{\hat{\mathbf{y}}}\sigma_{\mathbf{y}}} \left((y_i - \bar{y}) - \rho(\hat{\mathbf{y}}, \mathbf{y}) \cdot (\hat{y}_i - \bar{\hat{y}}) \right).$$

Similarly, the partial derivative with respect to y_i is given by:

$$\frac{\partial \rho(\hat{\mathbf{y}}, \mathbf{y})}{\partial y_i} = \frac{1}{n\sigma_{\hat{\mathbf{y}}}\sigma_{\mathbf{y}}} \left((\hat{y}_i - \bar{\hat{y}}) - \rho(\hat{\mathbf{y}}, \mathbf{y}) \cdot (y_i - \bar{y}) \right).$$

3.1.2 Convergence Analysis

Consider the total loss function L_{total} which includes the Pearson correlation coefficient as a regularization term, as shown in Equation (1).

The Pearson correlation coefficient $\rho(\hat{\mathbf{y}}, \mathbf{y})$ is a smooth function, composed of means, covariances, and standard deviations. Consequently, the regularization term $\lambda(1 - \rho(\hat{\mathbf{y}}, \mathbf{y}))$ is also smooth.

Since $\rho(\hat{\mathbf{y}}, \mathbf{y})$ is bounded within $[-1, 1]$, the regularization term is bounded as well:

$$0 \leq \lambda(1 - \rho(\hat{\mathbf{y}}, \mathbf{y})) \leq 2\lambda. \tag{6}$$

This boundedness ensures the regularization term does not overshadow the combined loss term L_{com}, allowing it to guide the optimization process effectively. Both L_{com} and the regularization term are smooth and differentiable, making L_{total} smooth and differentiable.

In gradient descent, a smooth and differentiable loss function typically ensures convergence to a local optimum with an appropriate learning rate.

The gradients of L_{total} with respect to $\hat{\mathbf{y}}$ and $\mathbf{y}$ are:

$$\begin{aligned}\frac{\partial L_{\text{total}}}{\partial \hat{y}_i} &= \frac{\partial L_{\text{com}}}{\partial \hat{y}_i} - \frac{\lambda}{n\sigma_{\hat{\mathbf{y}}}\sigma_{\mathbf{y}}}\left((y_i - \bar{y})\right.\\ &\quad \left.-\rho(\hat{\mathbf{y}}, \mathbf{y}) \cdot (\hat{y}_i - \bar{\hat{y}})\right), && (7)\\ \frac{\partial L_{\text{total}}}{\partial y_i} &= \frac{\partial L_{\text{com}}}{\partial y_i} - \frac{\lambda}{n\sigma_{\hat{\mathbf{y}}}\sigma_{\mathbf{y}}}\left((\hat{y}_i - \bar{\hat{y}})\right.\\ &\quad \left.-\rho(\hat{\mathbf{y}}, \mathbf{y}) \cdot (y_i - \bar{y})\right). && (8)\end{aligned}$$

The gradient descent update rules for $\hat{\mathbf{y}}$ and $\mathbf{y}$ are:

$$\hat{y}_i^{(t+1)} = \hat{y}_i^{(t)} - \eta\frac{\partial L_{\text{total}}}{\partial \hat{y}_i}, \tag{9}$$

$$y_i^{(t+1)} = y_i^{(t)} - \eta\frac{\partial L_{\text{total}}}{\partial y_i}, \tag{10}$$

where η is the learning rate and t denotes the iteration.

Given the smoothness and differentiability of the total loss function, and with an appropriate learning rate, the gradient descent algorithm is expected to converge to a local optimum, minimizing the total loss L_{total}.

In conclusion, the inclusion of the Pearson correlation coefficient as a regularization term maintains the smoothness and differentiability of L_{total}. This ensures the gradient descent algorithm converges to a local optimum, minimizing L_{total}.

3.2 Consistent-Mixup Data Augmentation

To enhance the robustness of emotion and empathy prediction models, we employed a Consistent-Mixup data augmentation (CMDA) technique. Traditional Mixup methods, which interpolate between pairs of inputs and their labels, can lead to inaccuracies in multi-label tasks like emotion and empathy prediction due to label complexity.

Our approach ensures that only samples with the same label are mixed, maintaining label consistency and preventing erroneous data. Given input sequences x with labels y, the data augmentation process is:

$$\tilde{x} = \gamma x_i + (1 - \gamma)x_j, \tag{11}$$

where $y_i = y_j$ and $\gamma \sim \text{Beta}(\alpha, \alpha)$. The Beta distribution, denoted as $\text{Beta}(\alpha, \alpha)$, is a continuous probability distribution defined on the interval $[0, 1]$ and parameterized by two positive shape parameters α. Setting both parameters to α ensures a symmetric distribution, which results in a balanced interpolation between inputs. This preserves the integrity of the labels, enhancing the quality of the augmented data and improving model performance and generalization.

Data set	Model Type	Emotion	Emotional Polarity	Empathy
Dev	BERT (S)	0.620	0.697	0.567
	BERT (P)	0.625	0.705	0.597
	RoBERTa (S)	0.637	0.705	0.592
	RoBERTa (P)	0.648	0.724	0.595
	DeBERTa (S)	0.620	0.720	0.599
	DeBERTa (P)	0.648	0.725	0.607
	Boosting (S)	0.653	0.744	0.616
	Boosting (P)	**0.667**	0.757	0.625
	Boosting (PC)	0.659	**0.765**	**0.658**
Test	Boosting (PC)	**0.581**	**0.644**	**0.561**

Table 1: Performance comparison of various models on Emotion, Emotional Polarity, and Empathy tasks on the development dataset (Dev) and test dataset (Test). (S) indicates the standard model, (P) indicates the model with added Pearson loss, and (PC) represents the model with Pearson loss and CMDA. The test results are reported for the Boosting (PC) model.

3.3 Experimental Setup

To validate the effectiveness of incorporating the Pearson correlation coefficient as a regularization term in emotion prediction models, we conducted a series of experiments using several state-of-the-art language models: BERT, RoBERTa, and DeBERTa. These models were chosen for their robust performance in various natural language understanding tasks. Additionally, we applied CMDA and boosting techniques to combine these models, aiming to enhance performance by leveraging their complementary strengths.

3.4 Datasets and Data Preprocessing

We used the Track 2 (CONV-turn) dataset, which contains 11,166 training items, 990 develop items, and around 2,300 test items, each with a dialogue text and three corresponding labels: emotional intensity, emotional polarity, and empathy. The length of each dialogue text ranges from 2 characters to 128 characters. The values of emotional intensity and empathy range from 0-5, and the values of emotional polarity range from 0-3. The data is preprocessed by identifying redundant columns and merging the redundant column contents into the correct columns to ensure that the final data is aligned with the corresponding features, and no data is lost in this process(Omitaomu et al., 2022).

3.5 Models and Training

Baseline Models: We implemented baseline versions of BERT, RoBERTa, and DeBERTa without Pearson regularization. These models were trained using the loss function (L_{com}).

Enhanced Models: For the enhanced versions, we added the Pearson correlation coefficient as a regularization term to the loss function. For a given predicted sentiment level $\hat{y}$ and actual sentiment level y is defined as:

$$\text{Loss} = L_{\text{com}}(\hat{y}, y) - \lambda \cdot (1 - \text{Pearson}(\hat{y}, y)),$$

where λ is a hyperparameter that controls the weight of the Pearson regularization term.

Training process: All models were trained using the Adam optimizer with a learning rate of $1e^{-5}$. To enhance the models, we incorporated data augmentation. Specifically, we employed CMDA ensuring the mixed labels remained consistent. Additionally, we adopted a boosting strategy by training three different models(BERT, RoBERTa, DeBERTa) and aggregating their outputs using a weighted average method to form the final prediction. This ensemble approach aimed to leverage the strengths of each individual model and improve overall performance.

3.6 Experimental Results

The evaluation metric used in this study is the Pearson Correlation Coefficient, which evaluates the linear correlation between the predicted and actual sentiment levels, reflecting the consistency of the predictions.

The study evaluates sentiment prediction models using the Pearson Correlation Coefficient to measure the linear correlation between predicted and actual sentiment levels. Table 1 shows that using Pearson correlation as a regularizer significantly enhances performance across all tested configurations. Enhanced models (BERT, RoBERTa, and DeBERTa with Pearson regularization) consistently outperform their baselines in Emotion, Emotional Polarity, and Empathy tasks.

Furthermore, Boosting models demonstrate additional improvements. The Boosting (Standard) model, which combines the results of the individual standard models using weighted averages, shows better performance than the individual models. The Boosting (Pearson) model, which similarly combines the Pearson-regularized models, achieves even higher scores. The best performance is from the Boosting (Pearson, CMDA) model, with top scores in Emotional Polarity (0.765) and Empathy (0.658). The Boosting (Pearson) model excels in Emotion (0.667), underscoring the benefits of Pearson correlation regularization.

The test set results also highlight the robustness of the models. The Boosting (Pearson, CMDA) model achieved scores of 0.581, 0.644, and 0.561 in Emotion, Emotional Polarity, and Empathy respectively. It is important to note that these scores are significantly higher than the official results of -0.027, -0.020, and -0.043 respectively. The discrepancy arose because an early version of the model was submitted by mistake, leading to the lower scores. The updated results presented here reflect the true performance of the final, optimized models.

4 Conclusion

We proposed an enhanced approach for emotion prediction by incorporating the Pearson correlation coefficient as a regularization term in the loss function, ensuring closer alignment between predicted and actual emotion levels. This method, along with CMDA and boosting techniques, significantly improved model performance, yielding higher correlation scores and more accurate predictions. Our findings underscore the potential of correlation-based regularization and advanced training techniques in enhancing Transformer-based models for emotion prediction tasks.

Limitations

Due to time constraints, we submitted an earlier version of our results, leading to a lower score of -0.03 on TRACK CONV-turn. Here, we present the best results to accurately represent our system's performance, as shown in table 1. Relying solely on the Pearson correlation coefficient may not fully demonstrate our approach's effectiveness. A 1-3% increase in the Pearson coefficient, though modest, shows consistent improvement. For a more comprehensive evaluation, we will include other metrics, such as the F1 score, in future work. These additional metrics will further validate our approach.

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.

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.

Claudiu Creanga and Liviu P Dinu. 2024. Transformer based neural networks for emotion recognition in conversations. *arXiv preprint arXiv:2405.11222*.

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*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.

Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. 2023. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11483–11493.

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

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

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*.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

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.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, pages 194–206. Springer.

Geng Tu, Jintao Wen, Cheng Liu, Dazhi Jiang, and Erik Cambria. 2022. Context-and sentiment-aware networks for emotion recognition in conversation. *IEEE Transactions on Artificial Intelligence*, 3(5):699–708.

Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin, and James L Crowley. 2022. Transformer-based self-supervised learning for emotion recognition. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2605–2612. IEEE.

Weiwei Zhou, Jiada Lu, Chenkun Ling, Weifeng Wang, and Shaowei Liu. 2024. Boosting continuous emotion recognition with self-pretraining using masked autoencoders, temporal convolutional networks, and transformers. *arXiv preprint arXiv:2403.11440*.