ACL 2024
Bangkok, Thailand

WASSA 2024 Shared Task

Paper ID: WASSA-EMP-10

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 Turn

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

Contents

1. Introduction

Introduction

Emotion detection and empathy analysis are important and inevitable topics with great application potentials. To provide more insights, **WASSA 2024 Shared Task** focuses on Empathy Detection and Emotion Classification and Personality Detection.

We propose a solution towards **Track 2: Empathy and Emotion Prediction in Conversations Turns (CONV-turn)**, predicting the **Emotion, Emotion Polarity and Empathy** according to turn-level information during conversations

- To achieve this goal:
  BERT and DeBERTa are fine-tuned.
  Fast Gradient Method is used as **adversarial training**.
  The **CombinedLoss** is designed.

- After submitting to the competition:
  Data augmentation is adopted with the Segmented Mix-up.
  Boosting is used as **ensemble** method.
  Regression experiments are also conducted.

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

2. Methodology

2.1 Methodology: Fine-tuned BERT and DeBERTa

- The proposed model includes:

  Fine-tuned BERT or DeBERTa.
  The **CombinedLoss**.
  Downstream head for **classification or regression**.
  Augmentation and **ensemble**.

- The pretrained language models

  BERT: **bert-base-uncased**.
  DeBERTa: **deberta-base**.

- Task-oriented fine-tuning on Track 2

  Conducted on the training set of Track 2.
  Adapt from general language model to **specific prediction task**.

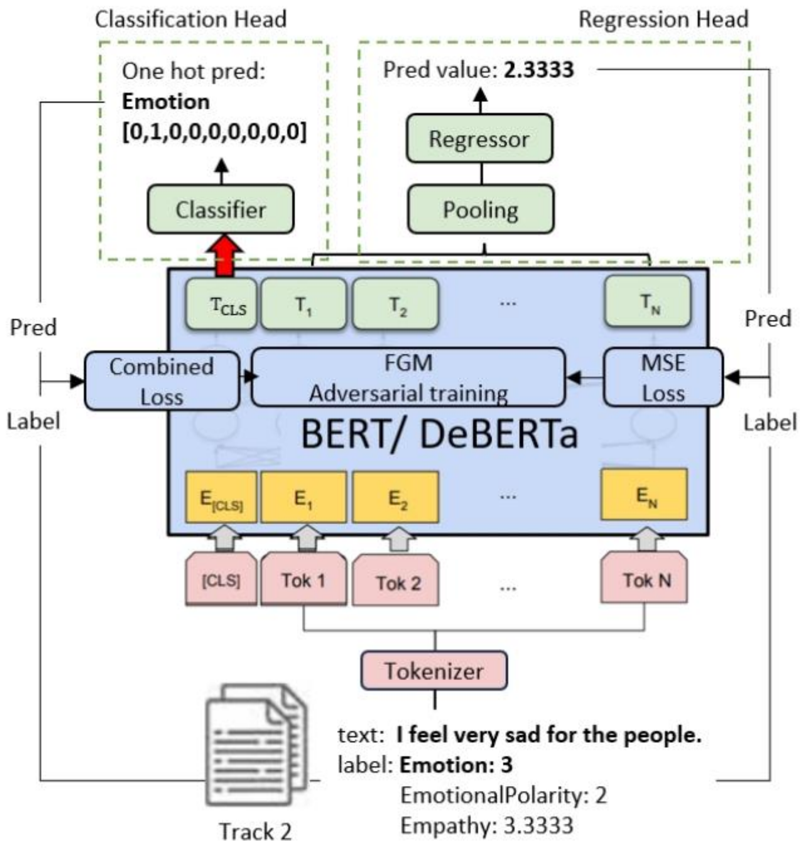

Figure 2: The proposed model

2.2 Methodology: The CombinedLoss

- Different from commonly-used loss functions, we proposed the **CombinedLoss**

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

- L_{loss}: the structured contrastive loss for classification
- λ: the regularization coefficient
- $Corr_{Pear}(\hat{y},\ y)$: Pearson correlation coefficient between prediction and the ground truth

- The Pearson correlation coefficient is used as a **regularization term**

2.3 Methodology: Adversarial Training with FGM

- To improve its robustness and generalization, **adversarial training** is introduced.

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

- x: the input sample
- δ: the added perturbation for adversarial training
- f_θ: neural network function with θ as parameters

- By maximizing $L(f_\theta(x + \delta))$, the **most disturbing perturbation** are introduced
- The model is trained to minimize the error, which helps it to be **more robust**

- **Fast Gradient Method** is used as adversarial training strategy
- Computes the most disturbing perturbation through **scaling the gradient**

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

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

2.4 Methodology: Augmentation with the Segmented Mix-up

- Mix-up is often used as a **data augmentation** method.
- Mix-up without constraint can't generate meaningful samples -> We proposed **Segmented Mix-up**

- Samples are divided into two segments: the lower one and the upper one (according to their labels)
- Sample (x_i, y_i) is paired with a (x_j, y_j) from the **same label segment**

- The **generated samples** are computed as:

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

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

- μ: the mix-up coefficient, sampled from a $Beta(\alpha, \alpha)$, with α controls the mix-up strength.

2.5 Methodology: Ensemble with Boosting

- To build more accurate and robust system, **boosting** is used to **ensemble** fine-tuned models
- Base models:
  Fine-tuned BERT and DeBERTa
- Weights:
  Weights are assigned according to the accuracy of each model on the development set
- The model with the **most reliable prediction** has the **greatest impact** on the final output

3. Experiments and Results

3.1 Experiments and Results: the Dataset and the Metric

- The **dataset** of Track 2 includes:
  - Training set: 11,166 samples
  - Development set: 990 samples
  - Test set: 2,061 valid samples
- Each **sample** consists:
  - **Text** content of a single dialogue turn
  - The labels of **Emotion, Emotional Polarity and Empathy**
  - Meta information of the speakers and the conversation
- Evaluation **Metric**:
- **Pearson correlation** of the prediction sequence $\hat{y}$ and the ground truth sequence y

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

3.2 Implementation Details

– Baselines:

BERT: **bert-base-uncased**, with 12 encoder layers and 110M parameters

DeBERTa: **deberta-base**, with 390M parameters

– Hyper-parameters:

Tokenization: **BertTokenizer** and **DebertaTokenizer**, with $max\_length = 128$

Optimization: **AdamW** optimizer, $learning\_rate = 1e - 6$, with exponential decay

The Segmented Mix-up: $\alpha = 0.2$ is used

– Labels and Categories:

The original labels also include float values: 0.3333, 0.6667 (training set) and 0.5, 1.5 (development set)

In classification, samples are **manually divided into categories** according to the label range

In regression, **original labels** are directly used as target values

3.3 Results and Analysis

- **Fine-tuned BERT and DeBERTa**
- The average results of fine-tuned **DeBERTa** is better than fine-tuned BERT
- By implementing the **CombinedLoss**, both models demonstrate performance gain
- Adding **adversarial training using FGM** brings better overall performance

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

3.3 Results and Analysis

- **Ensemble and Augmentation**
- The **combined boosting** yields the best avg. result among non-augment models
- Ensembling fine-tuned DeBERTas not always achieves the highest score
- **Augmentation** with our **Segmented Mix-up** brings further improvement

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

3.3 Results and Analysis

- **Classification and Regression**
- 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
- The fined-tuned DeBERTa achieved slightly better performance in **regression task**

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

4. Conclusions

Conclusion

- This is our solution to **WASSA 2024 Track 2**, which predicts **Emotion, Emotional Polarity and Empathy** using turn-level information.

- **BERT and DeBERTa** is fine-tuned with the **CombinedLoss** and **adversarial training with FGM**.
- 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 (**ranked 4th** among all teams).

- After the submission, **ensemble using boosting** method and **data augmentation with Segmented Mix-up** are adopted, 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.

- In the future, we plan to introduce larger datasets for **model re-training** at earlier stage (e.g. the Masked Language Model), and consider introducing **conversational context and speaker personality** for better model construction.

THANKS

Q & A