A Triple-View Framework for Fine-Grained Emotion Classification with Clustering-Guided Contrastive Learning

Junqing Gong Binhan Yang Wei Shen*
DISSec, TMCC, TBI Center, College of Computer Science,
Nankai University, Tianjin, China
{gongjq, yangbinhan}@mail.nankai.edu.cn, shenwei@nankai.edu.cn

Abstract

Fine-grained emotion classification (FEC) aims to analyze speakers' utterances and distinguish dozens of emotions with subtle differences, allowing for a more nuanced understanding of human emotional states. However, compared to traditional coarse-grained emotion classification, two difficulties arise as the granularity of emotions becomes finer, i.e., the presence of closely confusable emotions which are hard to distinguish, and the biased performance caused by long-tailed emotions. Although addressing both difficulties is vital to FEC, previous studies have predominantly focused on dealing with only one of them. In this paper, we propose TACO, a novel triple-view framework that treats FEC as an instance-label (i.e., utterance-emotion) joint embedding learning problem to tackle both difficulties concurrently by considering three complementary views. Specifically, we design a clustering-guided contrastive loss, which incorporates clustering techniques to guide the contrastive learning process and facilitate more discriminative instance embeddings. Additionally, we introduce the emotion label description as a helpful resource to refine label embeddings and mitigate the poor performance towards under-represented (i.e., long-tailed) emotions. Extensive experiments on two widely-used benchmark datasets demonstrate that our proposed TACO achieves substantial and consistent improvements compared to other competitive baseline methods.

1 Introduction

Emotion classification, which aims to recognize emotions conveyed in speakers' utterances, has exhibited strong momentum in motivating various emotion-driven applications such as conversational agent (Mishra et al., 2023), empathetic system (Samad et al., 2022) and affective computing (Mai et al., 2019). Traditionally, the majority of previous works (Zhang et al., 2023; Shen et al., 2021; Yin and Shang, 2022) addresses emotion classification in a coarse-grained way, and focuses on identifying only 6 (Ekman, 1992) or 8 (Plutchik, 1980) emotions with significant differences.

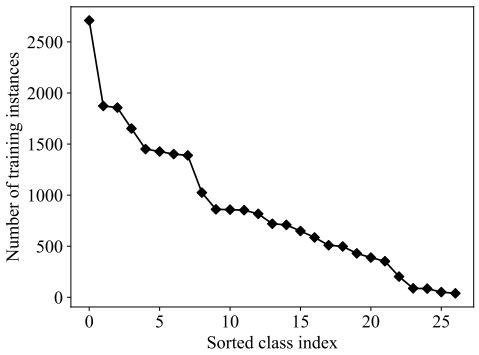


Figure 1: Data distribution of the training set in the GoEmotions dataset (Demszky et al., 2020).

However, humans experience dozens of emotions (Smith, 2015), far exceeding the limited number of classes commonly utilized in coarse-grained emotion classification. Thus, conversational agents that solely rely on coarse-grained approaches may fall short in capturing the diverse spectrum of emotions humans encounter and express in their daily lives. In order to facilitate more empathetic interactions, future conversational agents are expected to be equipped with the capacity of performing fine-grained emotion classification (FEC). Unlike coarse-grained emotion classification, FEC encompasses the recognition of much larger number of emotions, typically 28 (Demszky et al., 2020) or 32 (Rashkin et al., 2019). These fine-grained emotion classes capture subtle differences between emotions, allowing for a more nuanced understanding of human emotional states.

Unfortunately, two difficulties arise as the granularity of emotion classes becomes finer. First, there emerge some closely confusable classes with semantic overlap (e.g., *admiration* and *approval*). Ut-

*Wei Shen is the corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4970–4984

terances from these closely confusable classes may have small inter-class variations that are difficult even for humans to distinguish (Zhao et al., 2017). Second, fine-grained emotion datasets typically exhibit a long-tailed data distribution, as illustrated in Figure 1. This means a few dominant classes account for most of the instances, while most of the other classes are represented by relatively few instances (Cui et al., 2019). Naïve learning approaches on such data are susceptible to poor performance towards the under-represented emotion classes (Menon et al., 2020).

Armed with these insights, the core challenges lied in FEC can be categorized into two-fold: **1)** Simultaneously increasing the inter-class variation and the intra-class coherence to better distinguish closely confusable classes; **2)** Promoting the performance towards the long-tailed emotion classes.

To tackle the above challenges, researchers on FEC have proposed several approaches. Label-aware contrastive loss (LCL) (Suresh and Ong, 2021) enhances the model's capacity of distinguishing closely confusable classes by elevating the weights of negative instances that are more likely to be confused with positive ones in its contrastive loss, which focuses on increasing the inter-class variation. HypEmo (Chen et al., 2023) aims to improve the performance for long-tailed emotion labels by introducing a label hierarchy to provide additional information for long-tailed emotions. While these approaches have made valuable contributions, they only concentrate on addressing a single challenge without considering the two challenges at the same time. Moreover, they exhibit inadequacies in mining various knowledge concealed within FEC.

In this paper, we propose **TACO**, a novel **T**riple-view framework for fine-gr**A**ined emotion classification with **C**lustering-guided c**O**ntrastive learning, which leverages mutli-aspect knowledge to tackle the above two challenges concurrently. Actually, TACO treats FEC as an instance-label joint embedding learning problem through the following three views, where each view characterizes different aspects of knowledge. In the $instance$-$label$ view, we utilize a dual-encoder architecture to encode the utterance instance and the emotion label respectively, where the emotion label description is introduced as a helpful resource to offer rich semantic information, particularly beneficial for encoding long-tailed emotion labels. In the $instance$-$instance$ view, a clustering-guided contrastive loss is proposed to increase the inter-class variation and the intra-class coherence simultaneously. Specifically, we resort to the clustering result of the current instance embeddings to recognize hard instance pairs for contrastive learning, leading to superior instance embeddings. In the $label$-$label$ view, we employ a label-aware disentangled loss to adaptively push apart the label embeddings, resulting in greater differentiation among closely confusable labels and thereby enlarging the inter-class variation. It can be seen that the above three views are complementary with each other, and the unified utilization of them is expected to yield a better classification of fine-grained emotions.

In summary, the main contributions of this paper are as follows:

- We propose a novel triple-view framework TACO, which combines multi-aspect knowledge from three complementary views to jointly resolve the two core challenges of FEC.
- In the instance-instance view, we incorporate clustering techniques to guide the contrastive learning process, resulting in refined instance embeddings that exhibit larger inter-class variation and improved intra-class coherence.
- A thorough experimental study over two widely-used benchmark datasets demonstrates that the proposed TACO significantly outperforms all the competitive baseline methods[1].

2 Methodology

We begin by providing a formal definition of the FEC task. Given a user from a conversational system, we concatenate the utterances spoken by the user as one utterance instance denoted by u. The objective of the FEC task is to recognize an emotion label e for each utterance instance u from a predefined emotion label set E. Compared to coarse-grained emotion classification, the label set of FEC is much more diverse and the difference between labels can be subtle.

The overall framework of our proposed TACO is illustrated in Figure 2, which is built on three complementary views: the instance-label view, the instance-instance view, and the label-label view. In the following, we firstly outline each of these views in detail, and then provide an elaboration of how to fuse them via a view combination module.

[1] https://github.com/Alcyoneus87/TACO

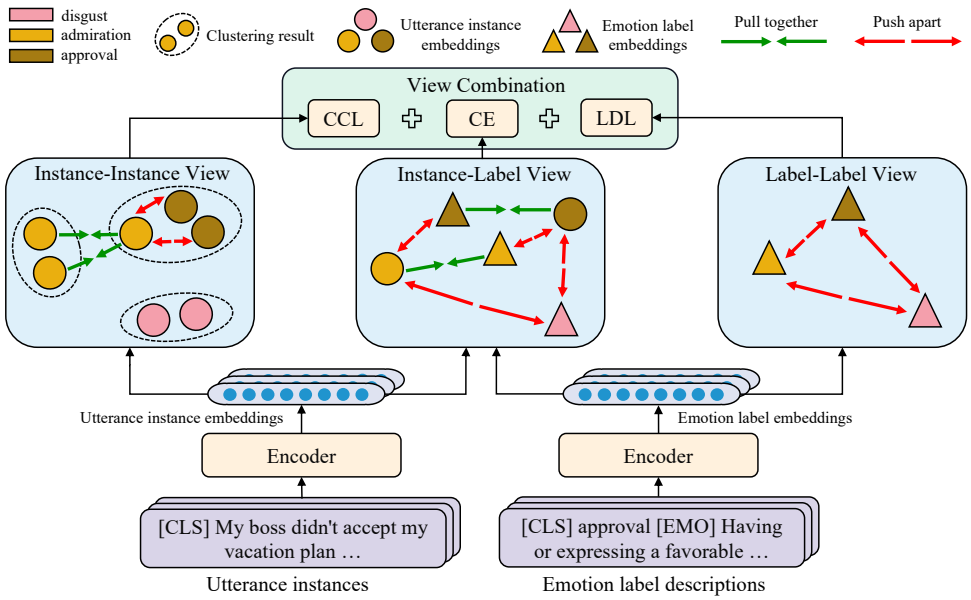


Figure 2: The overall framework of TACO. Here we visualize the case of classifying three emotion labels (i.e., *disgust*, *admiration* and *approval*), in which *admiration* and *approval* are closely confusable. The three emotion labels are painted in different colors in the diagram. Additionally, the instance embedding and the label embedding share the same color with their corresponding emotion label. For the initialization, *utterance instances* and *emotion label descriptions* are encoded independently into a unified embedding space via the same encoder. Subsequently, three complementary views optimize the obtained embeddings with the aim of increasing inter-class variation and intra-class coherence from different perspectives, followed by a view combination module to fuse the knowledge from them. For simplicity, the clustering-guided contrastive loss, cross-entropy loss and label-aware disentangled loss are denoted by CCL, CE and LDL in the diagram, respectively.

2.1 Instance-Label View

The task of FEC is traditionally formalized as a classification problem (Singh et al., 2023; Suresh and Ong, 2021), in which the prediction mainly depends on the learned embeddings of utterance instances, while the emotion labels are treated as meaningless one-hot vectors. However, as an essential element in this task, emotion labels contain rich semantic information that also merits exploration (Zhang et al., 2021). To leverage this valuable information and address the intricate interactions between utterance instances and emotion labels more effectively, we reformulate FEC as an instance-label joint embedding learning problem, where a dual-encoder architecture is utilized to encode the utterance instance and the emotion label, respectively. Specifically, for an utterance instance u and an emotion label e, their embeddings can be derived as:

$$\begin{aligned} \mathbf{h}_u &= FCN_1(PLM(s_u)) \\ \mathbf{h}_e &= FCN_2(PLM(s_e)) \end{aligned} \tag{1}$$

where s_u and s_e correspond to the input sequences of utterance instance u and emotion label e, respectively. PLM denotes a function for extracting the last layer representation of the `[CLS]` token from a pre-trained language model (e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019)). Moreover, FCN_1 and FCN_2 represent two fully-connected networks, each comprising two layers with ReLU activated.

The above input sequences can be constructed in the following ways. For the instance input sequence s_u, we represent the utterance instance itself by adding a prepended token `[CLS]` and an appended token `[SEP]`, i.e, `[CLS] utterance [SEP]`. About the label input sequence s_e, instead of solely utilizing the emotion label name, we incorporate the emotion label description as a helpful resource to provide additional semantic information, which is especially beneficial for modeling long-tailed emotion labels (Gao et al., 2023). Thus s_e can be defined as `[CLS] name [EMO] description [SEP]`, where `[EMO]` is a special token for separation.

Given an utterance instance u and an emotion label e, the prediction score $\hat{s}(u, e)$ between them is calculated via the dot-product of their corresponding embeddings $\mathbf{h}_u$ and $\mathbf{h}_e$ as:

$$\hat{s}(u, e) = \mathbf{h}_u \cdot \mathbf{h}_e \tag{2}$$

To ensure that each utterance instance is close to its corresponding ground truth emotion label and far from other emotion labels in the embedding space, the instance-label view is optimized with a cross-entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{|U|}\sum_{u\in U}\log\frac{exp(\hat{s}(u,e_u))}{\sum_{e\in E}exp(\hat{s}(u,e))} \quad (3)$$

where U represents the set of utterance instances with size $|U|$, and e_u denotes the ground truth emotion label of utterance instance u.

2.2 Instance-Instance View

In order to simultaneously increase the inter-class variation and the intra-class coherence, it is crucial to establish distinct classification boundaries among closely confusable classes, which requires identifying and correcting misclassified fuzzy instances that lie near those ambiguous boundaries. As we know, clustering is a simple yet effective approach for analyzing data distribution (Xie et al., 2016; Caron et al., 2018), while contrastive learning exhibits strong capabilities in aggregating positive instances and disentangling negative ones (Gao et al., 2021; Gunel et al., 2020). To this end, we propose a clustering-guided contrastive loss (CCL), in which clustering techniques are incorporated to guide the contrastive learning process, leading to more discriminative instance embeddings.

Specifically, given a mini-batch B that consists of $|B|$ utterance instances with indices $I \equiv \{1, 2, ..., i, ..., |B|\}$, we firstly employ a clustering algorithm (e.g., k-means) over the utterance instance embeddings $\{\mathbf{h}_{u_i}\}_{i=1}^{|B|}$ to partition instances into K clusters, where K is the known number of unique labels in the mini-batch. The clustering result can be denoted by a cluster index vector $\{c_i\}_{i=1}^{|B|}$, where $c_i \in \{1, 2, ..., K\}$ represents the cluster index of the i^{th} instance.

As shown in Figure 2, the obtained clusters can be divided into two categories: **1)** homogeneous clusters in which instances belong to the same class (e.g., *disgust*) that is likely to be easily distinguishable; **2)** heterogeneous clusters in which instances belong to different classes (e.g., *admiration* and *approval*) that are closely confusable. To establish distinct classification boundaries, CCL mainly focuses on handling these heterogeneous clusters and aims to convert them into homogeneous ones.

Informed by this insight, with the clustering result $\{c_i\}_{i=1}^{|B|}$ and the mini-batch labels $\{e_i\}_{i=1}^{|B|}$, we construct hard positive sets and hard negative sets based on the misclassified fuzzy instances that lie near ambiguous boundaries. Concretely, for the i^{th} instance, its hard positive set can be denoted by $P(i) \equiv \{p \in I : e_i = e_p \wedge c_i \neq c_p\}$, while $N(i) \equiv \{n \in I : e_i \neq e_n \wedge c_i = c_n\}$ represents its hard negative set. Given the hard postive and negative sets, an InfoNCE loss (Oord et al., 2018) could be utilized to correct the misclassified fuzzy instances, which is defined as:

$$\begin{aligned} l_{i,p} &= -\log\frac{exp(\mathbf{h}_{u_i}\cdot\mathbf{h}_{u_p}/\tau)}{\sum_{j\in P(i)\cup N(i)}exp(\mathbf{h}_{u_i}\cdot\mathbf{h}_{u_j}/\tau)} \\ \mathcal{L}_{CCL} &= \frac{1}{|B|}\sum_{i=1}^{|B|}\frac{1}{|P(i)|}\sum_{p\in P(i)}l_{i,p} \end{aligned} \quad (4)$$

where $\tau > 0$ is an adjustable scalar temperature parameter that controls the contrastive strength, and $|P(i)|$ is the cardinality of $P(i)$.

It can be seen that hard positive sets encourage instances belonging to different clusters but having the same emotion label to get closer, resulting in improved intra-class coherence, while hard negative sets enforce instances belonging to the same cluster but having different emotion labels to be farther away from each other, leading to larger inter-class variation. Hence, our proposed clustering-guided contrastive loss is expected to learn more discriminative embeddings for the instances from closely confusable classes and eliminate those ambiguous classification boundaries.

2.3 Label-Label View

The critical problem in FEC is that instances with closely confusable labels are hard to distinguish. While the aforementioned CCL partially addresses this issue by fine-tuning ambiguous boundaries on the instance side, it fails to account for the considerable semantic overlap among these closely confusable labels, resulting in tiny distances between their corresponding label embeddings. In light of this, we design a label-aware disentangled loss (LDL) on the label side to adaptively push apart the embeddings of closely confusable labels.

To quantify how confusable two emotion labels are for targeted optimization, we introduce a pairwise *confusability* score between emotion labels based on the predictions of utterance instances. Inspired by the classical collaborative filtering algorithm (Sarwar et al., 2001), we define the confusability score between two labels by treating utter-

ance instances and emotion labels as analogous to users and items in recommendation systems, respectively. Concretely, given a mini-batch of $|B|$ utterance instances and a label set E with $|E|$ emotion labels, the model's predictions $M \in R^{|B|\times|E|}$ can be viewed as a user-item rating matrix, where $M_{i,j}$ denotes the prediction score $\hat{s}(u_i, e_j)$ between instance u_i and label e_j calculated via Equation 2. Similar to the collaborative filtering algorithm in which the similarity between two items is measured through their ratings derived from all users, we calculate the confusability score between two emotion labels e_j and $e_{j'}$ based on the prediction distribution of instances over them as follows:

$$f(e_j, e_{j'}) = \frac{cos(M_{:,j}, M_{:,j'})}{\sum_{e_k, e_{k'} \in E} cos(M_{:,k}, M_{:,k'})} \quad (5)$$

Here, $cos(\cdot,\cdot)$ denotes cosine similarity and $M_{:,j}$ is the $|B|$ dimensional vector corresponding to label e_j, obtained by extracting the j^{th} column of M.

Then we push apart the label embeddings via minimizing the cosine similarity of each label pair weighted by its confusability score, ensuring that this label-aware disentangled loss primarily focuses on closely confusable label pairs, expressed as:

$$\mathcal{L}_{LDL} = exp(\sum_{e_j, e_{j'} \in E} f(e_j, e_{j'})\, cos(\mathbf{h}_{e_j}, \mathbf{h}_{e_{j'}})) \quad (6)$$

In summary, the proposed LDL effectively enlarges the inter-class variation, thus facilitating better discrimination between instances with closely confusable labels.

2.4 View Combination

The above three views provide complementary information all of which is vital to the FEC task. To fully exploit and combine knowledge from the three views, the overall loss of TACO can be calculated with a weighted sum operation as follows:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{CCL} + \beta\mathcal{L}_{LDL} \quad (7)$$

where α and β control the strengths of the CCL and the LDL, respectively.

At the inference stage, given an utterance instance u, the emotion label with the highest prediction score is chosen as the predicted label e^* for it based on the following equation:

$$e^* = \arg\max_{e \in E} \hat{s}(u, e) \quad (8)$$

Dataset	Train	Val	Test	# Labels
ED	19,533	2,770	2,547	32
GE	23,485	2,956	2,984	27

Table 1: The statistics of the two FEC datasets.

3 Experimental Setting

3.1 Datasets

We evaluate our proposed TACO on the following two widely-used public benchmark FEC datasets:
Empathetic Dialogues (Rashkin et al., 2019) comprises 24,850 multi-turn conversations annotated with one of 32 emotions, where each conversation circulates between a speaker and a listener. To ensure a fair comparison with previous works (Suresh and Ong, 2021; Chen et al., 2023), we only utilize the first turn of the conversation to construct the corresponding utterance instance, consisting of a situation description about the emotional incident.
GoEmotions (Demszky et al., 2020) is a human-annotated dataset containing 58,000 Reddit comments extracted from popular English subreddits, where each comment (i.e., utterance instance) is annotated with one or multiple labels from a set of 27 emotion labels plus neutral. Following previous works (Suresh and Ong, 2021; Chen et al., 2023), we only utilize the single-labeled instances and exclude instances with the neutral label.

The statistics of these two datasets are shown in Table 1. Empathetic Dialogues and GoEmotions are donated as ED and GE for brevity, respectively.

3.2 Baseline Methods

To conduct a comprehensive evaluation and comparison, we employ the following mainstream methods as our baselines[2]:
Pre-trained language models. $\text{BERT}_{\text{base}}$ (Devlin et al., 2019), $\text{RoBERTa}_{\text{base}}$ (Liu et al., 2019) and $\text{ELECTRA}_{\text{base}}$ (Clark et al., 2020) are utilized to encode the utterance instance, followed by a fully-connected network for classification.
Coarse-grained emotion classification methods. EmoBERTa (Kim and Vossen, 2021) introduces speaker names and separation tokens to classify emotions in a speaker-aware way. SACL (Hu et al., 2023) proposes a supervised adversarial contrastive learning framework to learn class-spread structured representations. TFD (Tu et al., 2023) mitigates biases by generating counterfactual utterances and leveraging subtraction operations.

[2]Refer to Appendix A for more details of baseline methods.

Method	Empathetic Dialogues			GoEmotions		
	Acc↑	Weighted-F1↑	Macro-F1↑	Acc↑	Weighted-F1↑	Macro-F1↑
$\text{BERT}_{\text{base}}$	$55.79_{0.35}$	$55.13_{0.51}$	$54.85_{0.56}$	$64.32_{0.33}$	$63.84_{0.32}$	$54.27_{0.89}$
$\text{RoBERTa}_{\text{base}}$	$57.67_{0.46}$	$57.13_{0.37}$	$56.87_{0.35}$	$64.79_{0.52}$	$64.34_{0.45}$	$54.91_{0.67}$
$\text{ELECTRA}_{\text{base}}$	$57.42_{0.61}$	$56.59_{0.62}$	$56.38_{0.63}$	$64.65_{0.57}$	$63.77_{0.62}$	$52.03_{1.04}$
EmoBERTa	$57.73_{0.44}$	$57.22_{0.45}$	$56.81_{0.43}$	$64.67_{0.51}$	$64.36_{0.49}$	$54.94_{0.72}$
SACL	$58.27_{0.45}$	$57.65_{0.48}$	$57.50_{0.48}$	$64.71_{0.54}$	$64.42_{0.47}$	$54.96_{0.89}$
TFD	$58.47_{0.55}$	$57.96_{0.47}$	$58.12_{0.51}$	$64.83_{0.42}$	$64.41_{0.50}$	$55.59_{1.09}$
$\text{BERT}_{\text{CDP+MLM}}$	$58.51_{0.48}$	$57.94_{0.38}$	$57.74_{0.39}$	$64.98_{0.40}$	$\underline{64.56}_{0.34}$	$\underline{55.84}_{0.93}$
LCL	$\underline{59.52}_{0.43}$	$\underline{58.72}_{0.49}$	$\underline{58.38}_{0.49}$	$\underline{65.22}_{0.39}$	$64.55_{0.47}$	$54.48_{1.27}$
HypEMO	$58.30_{0.50}$	$57.13_{0.42}$	$56.93_{0.48}$	$64.81_{0.46}$	$64.30_{0.39}$	$53.59_{1.14}$
$\text{ChatGPT}_{\text{zero-shot}}$	48.28	48.45	46.34	34.61	35.64	29.45
$\text{ChatGPT}_{\text{eight-shot}}$	$52.21_{0.22}$	$50.69_{0.41}$	$48.68_{0.53}$	$33.90_{0.60}$	$34.88_{0.58}$	$28.87_{0.53}$
Emollama-chat-7b	27.27	28.09	26.59	24.86	24.04	22.22
Emollama-chat-13b	38.37	39.47	38.06	28.35	27.85	23.06
TACO	$\mathbf{60.57}_{0.36}$	$\mathbf{59.94}_{0.42}$	$\mathbf{59.82}_{0.43}$	$\mathbf{65.97}_{0.38}$	$\mathbf{65.42}_{0.40}$	$\mathbf{58.23}_{0.99}$
Δ	+1.05%	+1.22%	+1.44%	+0.75%	+0.86%	+2.39%

Table 2: Overall performance of all baseline methods and our TACO on two FEC datasets. The subscript represents the corresponding standard deviation (e.g., $60.57_{0.36}$ indicates 60.57 ± 0.36). The best and second-best scores are set in **bold** and underlined, respectively, and Δ denotes the relative improvement between them.

Fine-grained emotion classification methods. $\text{BERT}_{\text{CDP+MLM}}$ (Singh et al., 2023) proposes a multi-task learning framework via introducing class definition prediction (CDP) and masked language model (MLM) as two auxiliary tasks. LCL (Suresh and Ong, 2021) trains an additional weighting network to increase the weights of hard negative instances in its contrastive loss. HypEMO (Chen et al., 2023) utilizes the hyperbolic distance between an instance-label pair as the weight of its cross-entropy loss.

Large language models. For each utterance instance, $\text{ChatGPT}_{\text{zero-shot/eight-shot}}$ utilizes the `gpt-3.5-turbo` model via the OpenAI API to generate an emotion label for the corresponding utterance prompt in the zero-shot/eight-shot way. Emollama-chat-7b (Liu et al., 2024) and Emollama-chat-13b (Liu et al., 2024) are two open-source emotional LLMs fine-tuned on emotion-related datasets. Both models operate in a zero-shot manner, utilizing the same utterance prompt as $\text{ChatGPT}_{\text{zero-shot/eight-shot}}$.

3.3 Evaluation Metrics

Following previous works (Suresh and Ong, 2021; Chen et al., 2023), we adopt the same top-1 accuracy (acc) and weighted-F1 as the evaluation metrics. Weighted-F1 assigns greater weights to classes with more instances and diminishes the contribution of long-tailed classes, defined as:

$$\text{weighted-F1} = \sum_{e \in E} \frac{n_e}{N} \times \frac{2 \times P_e \times R_e}{P_e + R_e} \tag{9}$$

where n_e is the number of test instances with label e, N is the total number of test instances, P_e and R_e are the precision and recall with respect to label e, respectively.

In addition, we adopt macro-F1 as a metric for the FEC task. In comparison to weighted-F1, macro-F1 treats all classes equally during testing since emotion labels are equally important in real-world scenarios (Frijda, 1986), which enables a more unbiased evaluation, formulated as:

$$\text{macro-F1} = \sum_{e \in E} \frac{1}{|E|} \times \frac{2 \times P_e \times R_e}{P_e + R_e} \tag{10}$$

3.4 Implementation Details

In TACO, we initialize the pre-trained language model using pre-trained `roberta-base` from HuggingFace's Transformers library (Wolf et al., 2020). For emotion label descriptions, we collect the corresponding label description by asking ChatGPT to define each emotion. The clustering algorithm is instantiated as k-means for simplicity, as the number of clusters K is known in advance. Note that the above three components are flexible and can be easily replaced with any other suitable techniques. For training, we adopt the AdamW (Loshchilov and Hutter, 2019) optimizer with an initial learning rate

of 2e-5 and weight decay of 1e-3. The mini-batch size $|B|$ and temperature parameter τ are set to 64 and 0.5, respectively. The hyperparameters α, β are set to 0.4, 0.4 for ED, and 0.45, 0.5 for GE. All baseline methods have released their source codes, and we used the same hyperparameter settings as specified in their original papers.

4 Experimental Results

4.1 Overall Results

Performance Comparison. The overall performance of all the methods is reported in Table 2. It can be observed that our proposed TACO consistently achieves the state-of-the-art performance on both the ED and GE datasets, indicating its superiority for the FEC task.

From Table 2, we can see that naïve fine-tuning of pre-trained language models alone yields unsatisfactory results, emphasizing the need for more sophisticated models and the utilization of more valuable knowledge in FEC. Coarse-grained emotion classification methods, although effective on various coarse-grained datasets, struggle when applied to fine-grained datasets, possibly due to their limitations in modeling the inter-class variation and the intra-class coherence for closely confusable emotion classes. However, as a coarse-grained method, TFD surprisingly outperforms some fine-grained methods because it tackles the long-tailed data distribution via mitigating label biases.

It is noteworthy that our proposed TACO surpasses all the fine-grained emotion classification baselines across all metrics on both datasets. These experimental results are in line with our intuition that simultaneously increasing the inter-class variation and the intra-class coherence leads to more discriminative representations, while also handling the long-tailed data distribution enables the training of a more generalizable model. By jointly addressing these two core challenges, our proposed TACO could solve the FEC task more effectively compared to previous fine-grained emotion classification baselines that just cope with a single challenge.

In addition, although large language models are capable of generating high-quality responses to prompts based on the knowledge learned from their extensive training corpora, they exhibit inadequacies in classifying fine-grained emotions. This may be attributed to the fact that while large language models store generalized knowledge from various domains, task-specific factors like closely confusable classes and long-tailed emotions require dedicated designs in resolving the FEC task.

Method	Acc↑	Weighted-F1↑	Macro-F1↑
TACO	**60.57**	**59.94**	**59.82**
- w/o LDL	59.85	59.27	59.18
- w/o CCL	59.61	58.93	58.80
- w/o CG	59.92	59.24	59.09
- w/o DES	59.74	59.12	59.04
TACO	**65.97**	**65.42**	**58.23**
- w/o LDL	65.59	65.12	57.99
- w/o CCL	65.16	64.84	57.63
- w/o CG	65.47	64.91	57.86
- w/o DES	65.23	64.80	57.62

Table 3: Performance of different variants of TACO. The *upper* and *lower* part list the results on ED and GE, respectively.

Analysis of macro-F1. In contrast to weighted-F1 that assigns lower weights to long-tailed emotion labels, macro-F1 treats all emotion labels equally. It can be seen from Table 2 that in terms of macro-F1, our TACO achieves large relative improvements (i.e., 1.44% on ED and 2.39% on GE) compared to the second-best baseline, which showcases its superiority in modeling long-tailed emotion labels, making it well-suited for real-world scenarios where various uncommon emotions exist. Moreover, it is noted that TACO has a larger improvement on GE than on ED. This is possibly due to the fact that GE has a greater imbalance rate for the long-tailed data distribution compared with ED.

4.2 Ablation Study

To investigate the contribution of each key module in our TACO, we conduct an ablation study considering the following variants: (1) TACO w/o LDL in which the label-aware disentangled loss in the label-label view is eliminated; (2) TACO w/o CCL in which the clustering-guided contrastive loss corresponding to the instance-instance view is deleted; (3) TACO w/o CG in which the clustering-guided component is omitted so that the instance-instance view is degraded into a naïve supervised contrastive loss (Khosla et al., 2020) without specially considering hard positive and negative instance pairs; (4) TACO w/o DES in which the emotion label description is removed and only the emotion label name is utilized to construct the label input sequence. We present the performance of these four variants as well as the whole framework TACO in Table 3.

From the experimental results, we can observe that TACO outperforms the first two variants on both datasets, which validates that the LDL and CCL modules in TACO are beneficial for the FEC

task and contribute positively to the performance of the overall framework. With regard to the third variant, the removal of the clustering-guided component results in a performance decline, implying that the incorporation of clustering techniques could recognize high-quality hard positive and negative instance pairs, thereby boosting the capability of the contrastive loss. Besides, it can be seen that the performance decreases when the emotion label description is removed, showcasing that the label description could indeed offer rich semantic information and therefore improve the performance towards long-tailed emotions.

5 Quantifying Model Confidence

In this section, we conduct a post-hoc analysis to quantify the ability of the the proposed TACO to distinguish closely confusable emotions. Specifically, beyond solely considering the highest value among prediction scores (i.e., top-1 accuracy), we turn to the distribution of prediction scores based on the intuition that a more discriminative model would generate a steeper score distribution. Following the previous work (Suresh and Ong, 2021), we quantify this by measuring the averaged entropy of the test set. For a given utterance instance u, we obtain its top-k highest prediction scores (denoted by $S_k \in \mathbb{R}^k$) from the overall prediction scores $S \in \mathbb{R}^{|E|}$, which are calculated with respect to every emotion label e via Equation 2. We then normalize S_k and calculate its entropy with an information-theoretic entropy loss:

$$\text{entropy}_k = -\sum_{s \in S_k} s \times log\ s \qquad (11)$$

A more discriminative model is expected to have a lower averaged entropy, indicating higher certainty in distinguishing closely confusable emotions. As the experimental results shown in Figure 3, TACO consistently produces prediction score distributions with lower entropies compared to other fine-grained emotion classification methods, showcasing that TACO could increase both the inter-class variation and the intra-class coherence, resulting in an advanced capacity of distinguishing closely confusable emotions.

6 Related Work

6.1 Fine-Grained Emotion Classification

Fine-grained emotion classification (FEC) has emerged as a popular research topic in recent years.

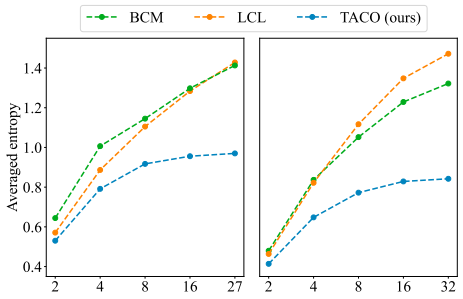


Figure 3: The averaged entropy of the prediction score distribution for different k values from 2 to 27/32. BCM denotes $\text{BERT}_{\text{CDP+MLM}}$. The *left* and *right* part list the results on ED and GE, respectively.

The core challenges in FEC involve accurately distinguishing closely confusable emotions and improving the performance towards long-tailed emotions, while previous works have predominantly concentrated on addressing only one of them. Specifically, $\text{BERT}_{\text{CDP+MLM}}$ (Singh et al., 2023) proposes a multi-task learning framework in which class definition prediction (CDP) is introduced as an auxiliary task for better understanding of various emotions. LCL (Suresh and Ong, 2021) increases the weights of hard negative instances in its contrastive loss to train a more discriminative model. HypEMO (Chen et al., 2023) provides additional information for long-tailed emotions by introducing a label hierarchy to revise the cross-entropy loss. In contrast, our proposed TACO takes into account both core challenges concurrently.

With the rise of large language models (LLMs), there have been notable efforts to exploit them for emotion classification. E-ICL (Yang et al., 2024) utilizes emotionally similar examples with dynamic labels and exclusionary emotion prediction, effectively addressing the limitations of standard in-context learning. EmoLLMs (Liu et al., 2024) presents a series of fine-tuned, open-source emotional LLMs, developed alongside a multi-task affective analysis instruction dataset (AAID) and a comprehensive evaluation benchmark (AEB).

6.2 Contrastive Learning

Recently, contrastive learning exhibits strong capabilities in aggregating positive samples and disentangling negative ones in both self-supervised and supervised settings (Wang et al., 2022; Xu et al., 2023; Yang et al., 2025). Self-supervised contrastive learning techniques, such as SimCLR (Chen et al., 2020) and SimCSE (Gao et al., 2021), leverage unlabeled data to improve representation

learning using only the augmented versions of original samples. Supervised contrastive learning (Khosla et al., 2020; Gunel et al., 2020) takes labels into account and therefore has a more diverse way to generate positive and negative samples. Contrastive learning has also been increasingly applied to multi-modal emotion-related tasks, demonstrating significant improvements in effectiveness via inter-modal alignment (Li et al., 2023; Wang et al., 2023). Yang et al. (2023) proposed SCCL, which conducts contrastive learning in the cluster-level via heuristically constructing clusters to obtain cluster-level instance and label embeddings, then adjusting the distance between them. Unlike previous works, we incorporate clustering techniques to recognize high-quality hard positive and negative instance pairs for the supervised contrastive learning process, leading to more discriminative instance embeddings.

7 Conclusion

In this paper, we propose a novel triple-view framework TACO, which simultaneously resolves the two difficulties existing in FEC. By treating FEC as an instance-label joint embedding learning problem, TACO could exploit multi-aspect knowledge via three complementary views. Concretely, we have devised a clustering-guided contrastive loss to facilitate more discriminative utterance instance embeddings, and the emotion label description is introduced to promote the performance towards long-tailed emotion labels. Empirical experiments indicate the effectiveness of TACO, which consistently surpasses all the baseline methods over two widely-used public benchmark datasets.

Limitations

Although our proposed TACO performs effectively by incorporating clustering techniques to guide the contrastive learning process, the inclusion of clustering techniques increases the computational complexity and lengthens the training time, and the direct correlation between the number of classes and the number of clusters may produce suboptimal results. Moreover, TACO primarily focuses on solving the case where there is only one utterance from a single speaker. However, in real-world scenarios, dialogues often contain multiple speakers with multiple utterances. Furthermore, our proposed TACO is assessed solely on small-scale English-language datasets, and its performance is not evaluated in large-scale or other languages scenarios, since to the best of our knowledge, there is currently a lack of such datasets. Finally, while TACO has made some efforts to address the long-tailed problem lied in fine-grained emotion classification, it is still somewhat biased toward common emotions. The aforementioned limitations will be left for our future research.

Ethics Statement

Our method, which analyzes speakers' utterances and distinguishes fine-grained emotions, has potential applications in psychopathological fields like depression detection. Understanding and identifying chronic expression of negative emotions such as guilty and anger may provide valuable insights as precursors to depression (O'Connor et al., 2002). However, it is important to acknowledge that our model may unintentionally exhibit biases towards uncommon emotions, which could potentially affect fairness and accuracy when the model is utilized in real-world scenarios. It is imperative to be aware of and address any potential ethical concerns to ensure the responsible and ethical use of our method in real-world applications.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62272247, CCF-Huawei Populus Grove Fund and the Fundamental Research Funds for the Central Universities under Grants 079-63243153 and 070-63253228.

References

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149.

Chih Yao Chen, Tun Min Hung, Yi-Li Hsu, and Lun-Wei Ku. 2023. Label-aware hyperbolic embeddings for fine-grained emotion classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10947–10958, Toronto, Canada. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

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.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Nico H Frijda. 1986. *The emotions*. Cambridge University Press.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2023. The benefits of label-description training for zero-shot text classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13823–13844, Singapore. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10835–10852, Toronto, Canada. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. 2023. Joyful: Joint modality fusion and graph contrastive learning for multimoda emotion recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16051–16069.

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*.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 481–492, Florence, Italy. Association for Computational Linguistics.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023. PAL to lend a helping hand: Towards building an emotion adaptive polite and empathetic counseling conversational agent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12254–12271, Toronto, Canada. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Lynn E O'Connor, Jack W Berry, Joseph Weiss, and Paul Gilbert. 2002. Guilt, fear, submission, and empathy in depression. *Journal of affective disorders*, 71(1-3):19–27.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856, Seattle, United States. Association for Computational Linguistics.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560, Online. Association for Computational Linguistics.

Gargi Singh, Dhanajit Brahma, Piyush Rai, and Ashutosh Modi. 2023. Text-based fine-grained emotion prediction. *IEEE Transactions on Affective Computing*.

Tiffany Watt Smith. 2015. *The book of human emotions: An encyclopedia of feeling from anger to wanderlust*. Profile books.

Varsha Suresh and Desmond Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Geng Tu, Ran Jing, Bin Liang, Min Yang, Kam-Fai Wong, and Ruifeng Xu. 2023. A training-free debiasing framework with counterfactual reasoning for conversational emotion detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15639–15650, Singapore. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022. SimKGC: Simple contrastive knowledge graph completion with pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4281–4294, Dublin, Ireland. Association for Computational Linguistics.

Yusong Wang, Dongyuan Li, Kotaro Funakoshi, and Manabu Okumura. 2023. Emp: Emotion-guided multi-modal fusion and contrastive learning for personality traits recognition. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 243–252.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic. 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.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 478–487. JMLR.org.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023. SimCSE++: Improving contrastive learning for sentence embeddings from two perspectives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12028–12040, Singapore. Association for Computational Linguistics.

Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Transactions on Affective Computing*.

Yang Yang, Wei Shen, Junfeng Shu, Yinan Liu, Edward Curry, and Guoliang Li. 2025. Cmvc+: a multi-view clustering framework for open knowledge base canonicalization via contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2296–2310.

Zhou Yang, Zhaochun Ren, Chenglong Ye, Yufeng Wang, Haizhou Sun, Chao Chen, Xiaofei Zhu, Yunbing Wu, and Xiangwen Liao. 2024. E-icl: Enhancing fine-grained emotion recognition through the lens of prototype theory. *arXiv preprint arXiv:2406.02642*.

Wenbiao Yin and Lin Shang. 2022. Efficient nearest neighbor emotion classification with BERT-whitening. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4738–4745, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. 2023. DualGATs: Dual graph attention networks for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7395–7408, Toronto, Canada. Association for Computational Linguistics.

Kun Zhang, Le Wu, Guangyi Lv, Meng Wang, Enhong Chen, and Shulan Ruan. 2021. Making the relation matters: Relation of relation learning network for sentence semantic matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14411–14419.

Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. 2017. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2):119–135.

A Setting Details of Baseline Methods

Coarse-grained emotion classification methods. Since some coarse-grained emotion classification methods (Kim and Vossen, 2021; Hu et al., 2023; Tu et al., 2023) were proposed to classify in a dialog environment with past and future utterances, which are not available in FEC, we have to remove the corresponding modules of such methods when applying them to FEC. Moreover, to facilitate a fair comparison, we change their pre-trained language models from `roberta-large` to `roberta-base`.

Fine-grained emotion classification methods. $\text{BERT}_{\text{CDP+MLM}}$ (Singh et al., 2023) was originally proposed for multi-label emotion classification, so we have to modify its binary cross-entropy loss to a cross-entropy loss since the FEC task focuses on single-label emotion classification. Moreover, it has been found that the `sklearn.metrics.f1_score` was misused in HypEMO (Chen et al., 2023), resulting in the unusual high performance on the weighted-F1 score. To amend this, we obtain its updated performance via running the open-source solution from its official Github repository[3].

Large language models. For $\text{ChatGPT}_{\text{zero-shot/eight-shot}}$, we set the temperature to 0 for replication. In zero-shot setting, we directly ask the LLM to generate an emotion label through instruction. In one-shot setting, we sample a random utterance instance from the training set as an in-context demonstration before asking. Taking the GoEmotions dataset (Demszky et al., 2020) as an example, the designed utterance prompt templates are shown in Table 8. For Emollama-chat-7b (Liu et al., 2024) and Emollama-chat-13b (Liu et al., 2024), we load their checkpoints posted on huggingface directly based on the VLLM engine, and use the same utterance prompt templates as $\text{ChatGPT}_{\text{zero-shot/eight-shot}}$ to ensure consistency.

B Qualitative Analysis and Error Analysis

We conduct qualitative analysis and error analysis on TACO and its two variants: (1) TACO w/o CCL in which the clustering-guided contrastive loss corresponding to the instance-instance view is deleted; (2) TACO w/o LDL in which the label-aware disentangled loss in the label-label view is eliminated. Table 4 shows the results of these variants as well as the whole framework TACO on two test utterance instances sampled from the GoEmotions dataset.

For the first case, our whole framework TACO outputs the correct label admiration. However, without the CCL loss, the model is misled by the word "proud" to the wrong label pride, demonstrating the effectiveness of CCL in refining instance embedding. The exclusion of the LDL loss results in a wrong prediction of approval, suggesting that LDL can indeed help distinguish closely confusable classes. For the second case, even with the help of CCL and LDL, the model still fails to predict correctly, indicating that the FEC task is non-trivial and our proposed TACO cannot fully understand the emotion expressed by the utterance in some difficult cases.

C Experimental Results on Additional Datasets

In addition to the existing Empathetic Dialogues (Rashkin et al., 2019) and GoEmotions (Demszky et al., 2020) datasets, EmpatheticIntent (Welivita and Pu, 2020) and EDOS (Welivita et al., 2021)

[3] https://github.com/dinobby/HypEmo

Utterance Instance	w/o CCL	w/o LDL	TACO	Golden Label
you did so good for yourself though! if you had the mother you deserve she would have the sense to be **proud** of you.	pride	approval	admiration	admiration
i'm not sure why this is a nonononoyes, we didn't see any of the beforehand i really only see the "yes".	curiosity	curiosity	curiosity	confusion

Table 4: Examples of utterance instance from the GoEmotions dataset for the qualitative analysis and error analysis.

Method	EmpatheticIntent			EDOS		
	Acc↑	Weighted-F1↑	Macro-F1↑	Acc↑	Weighted-F1↑	Macro-F1↑
$\text{BERT}_{\text{CDP+MLM}}$	58.53	57.99	58.01	64.24	63.93	51.88
LCL	<u>59.26</u>	<u>58.71</u>	<u>58.47</u>	<u>64.72</u>	<u>64.38</u>	<u>52.01</u>
Emollama-chat-7b	23.76	22.91	21.20	11.40	8.02	7.12
Emollama-chat-13b	31.66	31.61	29.26	11.88	9.87	9.70
TACO	**60.74**	**60.31**	**59.92**	**65.00**	**64.84**	**52.47**
Δ	+1.48%	+1.60%	+1.45%	+0.28%	+0.46%	+0.46%

Table 5: Experimental results on two additional datasets. The best and second-best scores are set in **bold** and <u>underlined</u>, respectively, and Δ denotes the relative improvement between them.

are two other widely recognized datasets for the fine-grained emotion classification task. EmpatheticIntent consists of 6,770 open-domain, human-to-human conversations, where each conversation is associated with one of 41 emotions. The training/validation/test split of his dataset is 4,061 / 1,354 / 1,355. EDOS (**E**motional **D**ialogues in **O**pen**S**ubtitles) is a large-scale emotion dataset containing 50K emotional dialogues from movie subtitles, in which each dialogue turn is automatically annotated with 41 fine-grained emotions. The training/validation/test split of the dataset is 30,000 / 10,000 / 10,000.

In order to further evaluate the effectiveness of our proposed method on these two new datasets, this section presents a comparison between our method and the state-of-the-art baselines, including the most competitive SLM-based approaches (i.e., $\text{BERT}_{\text{CDP+MLM}}$ and LCL) as well as the LLM-based method (i.e., EmoLLMs). The overall performance of these methods is reported in Table 5. It can be seen that our method consistently outperforms both traditional and LLM-based baseline methods on the EmpatheticIntent and EDOS datasets, indicating its superiority for the FEC task.

D Exploration of Cluster Number Determination Methods

In the instance-instance view, we propose a clustering-guided contrastive loss (CCL), in which clustering techniques are incorporated to guide the contrastive process. During clustering, the number of clusters is an important hyper-parameter that directly affects the clustering quality. In our proposed TACO, the cluster number is driven by our task setting, specifically as the number of unique labels in each mini-batch. While this is a direct and task-aligned approach, using automatic algorithms may offer potential improvements in certain cases.

Method	Acc↑	Weighted-F1↑	Macro-F1↑
TACO	**60.57**	**59.94**	**59.82**
TACO w. elbow	60.21	59.54	59.44
TACO	**65.97**	**65.42**	**58.23**
TACO w. elbow	65.85	65.34	58.13

Table 6: Performance of TACO and TACO w. elbow. The *upper* and *lower* part list the results on ED and GE, respectively.

To explore this, we conduct experiments using the elbow algorithm to automatically determine the cluster number. The search range for this number is $[\max(K-5, 0), \min(K+5, B)]$, where K is the number of unique labels in the mini-batch, B represents the batch size. We use the SSE metric to find inflection points. From the experimental results shown in Table 6, it can be seen that the automatic algorithm elbow achieves slightly lower results than our method.

E Exploration of Foundation Model

To ensure a fair comparison with the baselines, we choose the encoder-only model roberta-base as our foundation model for the main experiments. To fur-

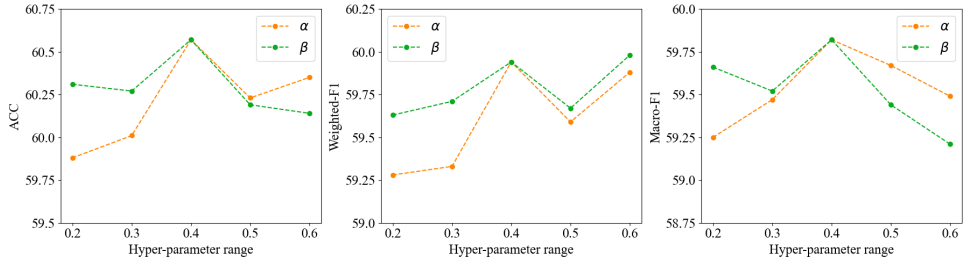


Figure 4: Hyper-parameter sensitivity analysis results on ED dataset. Here we fix one of α and β while adjusting the other to observe its influence.

Method	Acc↑	Weighted-F1↑	Macro-F1↑
TACO	60.57	59.94	59.82
TACO w. Llama	**62.43**	**61.33**	**61.07**
TACO	65.97	65.42	58.23
TACO w. Llama	**67.13**	**66.58**	**59.44**

Table 7: Performance of TACO and TACO w. Llama. The *upper* and *lower* part list the results on ED and GE, respectively.

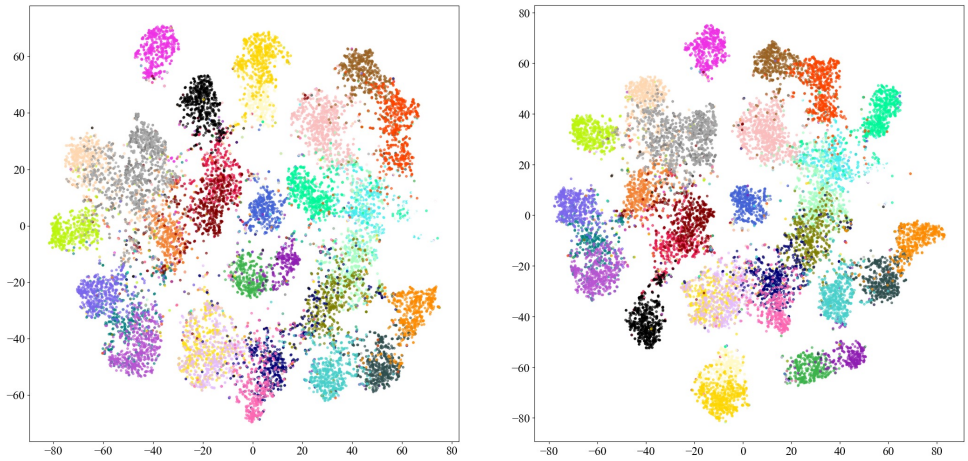
ther verify the generalizability of our method to the decoder-only models, we replace the foundation model with Llama 3.2 1B, and set the LoRA parameters to q_proj, k_proj, v_proj, and r to 16. Additionally, we use mean pooling to get the instance embedding. As shown in Table 7, the decoder-only large language model Llama achieves better results compared to the encoder-only model BERT, likely due to its larger number of parameters. These results confirm that our method is not limited to encoder-only models and can also be effectively applied to decoder-only models.

F Hyper-parameter Sensitivity Analysis

In this section, we conduct the hyper-parameter sensitivity analysis on the ED dataset to investigate how our method responds to changes in key hyper-parameters. Specifically, we analyze the hyper-parameters α and β, which control the strengths of the CCL and LDL loss functions, respectively. Since we set both parameters α and β to 0.4 in the main experiments, here we fix one of them while adjusting the other to observe its influence. The experimental results are summarized in Figure 4. It can be seen that α has a greater impact on the final results and the model is relatively less sensitive to changes in β.

G Visualization of Learned Representations

To further evaluate whether our proposed clustering-guided contrastive loss (CCL) improves the model's ability to distinguish closely confusable classes, we use the t-SNE (Van der Maaten and Hinton, 2008) method for visualization. As can be seen in Figure 5, when the CCL loss is applied, the classification boundaries become more distinct, and the learned representations are more discriminative. This suggests that CCL indeed helps the model differentiate between closely confusable classes.



(a) TACO w/o CCL (b) TACO

Figure 5: t-SNE visualization of learned representations on the Empathetic Dialogues dataset.

Zero-shot Prompt	You are an AI assistant who specializes psychology. You will complete an emotion recognition task. The task is as follows: according to a text written by an author, predicting the author's emotion from the following 27 emotions: 'admiration', 'amusement', 'anger', 'annoyance', 'approval', 'caring', 'confusion', 'curiosity', 'desire', 'disappointment', 'disapproval', 'disgust', 'embarrassment', 'excitement', 'fear', 'gratitude', 'grief', 'joy', 'love', 'nervousness', 'optimism', 'pride', 'realization', 'relief', 'remorse', 'sadness', 'surprise'. Only provide one emotion from above emotions and do not give the explanation. AURTHOR'S TEXT: {breaking news, husband borrows wife's car and should lose their job because of this.}, EMOTION:
One-shot Prompt	You are an AI assistant who specializes psychology. You will complete an emotion recognition task. The task is as follows: according to a text written by an author, predicting the author's emotion from the following 27 emotions: 'admiration', 'amusement', 'anger', 'annoyance', 'approval', 'caring', 'confusion', 'curiosity', 'desire', 'disappointment', 'disapproval', 'disgust', 'embarrassment', 'excitement', 'fear', 'gratitude', 'grief', 'joy', 'love', 'nervousness', 'optimism', 'pride', 'realization', 'relief', 'remorse', 'sadness', 'surprise'. Only provide one emotion from above emotions and do not give the explanation. AURTHOR'S TEXT: {omg! i can only imagine. i've gotten it into a hang nail before and that was not fun.}, EMOTION: surprise; AURTHOR'S TEXT: {breaking news, husband borrows wife's car and should lose their job because of this.}, EMOTION:

Table 8: Utterance prompt templates for ChatGPT of the GoEmotions dataset.