SEC: Context-Aware Metric Learning for Efficient Emotion Recognition in Conversation

Barbara Gendron[1,2], Gaël Guibon[1],
[1]LORIA, Université de Lorraine, CNRS, [2]Université du Luxembourg,
Correspondence: barbara.gendron@loria.fr ; gael.guibon@loria.fr

Abstract

The advent of deep learning models has made a considerable contribution to the achievement of Emotion Recognition in Conversation (ERC). However, this task still remains an important challenge due to the plurality and subjectivity of human emotions. Previous work on ERC provides predictive models using mostly graph-based conversation representations. In this work, we propose a way to model the conversational context that we incorporate into a metric learning training strategy, with a two-step process. This allows us to perform ERC in a flexible classification scenario and end up with a lightweight yet efficient model. Using metric learning through a Siamese Network architecture, we achieve 57.71 in macro F1 score for emotion classification in conversation on DailyDialog dataset, which outperforms the related work. This state-of-the-art result is promising in terms of the use of metric learning for emotion recognition, yet perfectible compared to the micro F1 score obtained.

1 Introduction

Computer Mediated Communication (CMC) is constantly evolving and new means of communicating are emerging. With the advent of conversational agents, there is a need to detect emotions within a conversation. Although many modalities are now considered in the communication process, the textual modality still remains essential for fast and easy everyday communication, through messaging applications, social media, and other networking platforms. Textual modality, however, is ambiguous, it does not preserve the extra-linguistic context, especially for dyadic human-to-human conversations. One main ambiguity that arises in CMC is the emotional state of the speaker, often misinterpreted by humans through short, and unpolished messages. This motivates Emotion Recognition in Conversation (ERC), a trending research topic dedicated not only to identifying emotion in messages, but also on taking into account the conversational context to recognize emotions. ERC has been shown to be challenging, especially with respect to the way to represent the context (Ghosal et al., 2021). Lately, it has seen a surge of multimodal models (Wen et al., 2023; Liang et al., 2023; Fan et al., 2024) and graph-related approaches (Zhang et al., 2023; Wang et al., 2023; Li et al., 2023) which often try to map the pattern of each speaker and better represent the conversational context, often resulting in good performance at the cost of efficiency. One additional issue ERC models are facing is their dependency on labels, models are mainly supervised and face the issue of extreme label imbalance due to emotional utterances being so scarce.

In this paper, we tackle these two challenges by incorporating the conversational context into metric learning while heavily controlling the data imbalance by multiple means. Considering that we want to tackle information across emotions to make our model usable for variants of emotions that go beyond the scope of the 6 basic emotions, we do not use supervised contrastive learning (Khosla et al., 2020) in our method. Instead, we focus on a two-step process to update the model using both direct label predictions through a cross-entropy loss and relative label assignment through the contrastive loss. This two-step process is quite straightforward, while using isolated elements, such as isolated utterances. However, to the best of our knowledge, contextual representation through contrastive learning for ERC has yet to be used. This represents our main contribution in this paper, as we present a model that can achieve competitive performance compared to the state-of-the-art while rendering the adaptation to other emotion labels feasible. Thus, our model can be applied and adapted in multiple contexts that require recognition of different label granularities.

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 11–22
August 15, 2024

Our main contribution lies in the development of a metric-learning training strategy for emotion recognition in utterances that incorporates the conversational context. The presented model leverages sentence embeddings and Transformer encoder layers (Vaswani et al., 2017; Devlin et al., 2019) to represent dialogue utterances and deploy attention on the conversational context. Our method involves Siamese Networks (Koch et al., 2015) in the setup but can be adapted to any metric-learning model. We further demonstrate that our approach outperforms some of the latest state-of-the-art Large Language Models (LLMs) such as light versions of Falcon (Penedo et al., 2023) or LLaMA 2 (Touvron et al., 2023). In addition, our method is efficient in the sense that it involves lightweight, adaptable and quickly trainable models, which still yield state-of-the-art performance on the DailyDialog dataset in macro F1 score with 57.71% and satisfactory results on micro F1 with 57.75%.

Our code and models are available on GitHub [1] to reproduce training, inference and qualitative experiments.

2 Related Work

ERC. Although most of the studies on ERC have been carried out on multimodal datasets (Song et al., 2022; Li et al., 2022; Hu et al., 2022), thus leveraging multi-modality, there are still some models developed for emotion recognition on textual conversation only, whether it be on multimodal datasets restricted to text such as IEMOCAP (Busso et al., 2008) or MELD (Poria et al., 2019), or on a fully textual dataset such as DailyDialog (Li et al., 2017). The advent of deep learning enables significant progress in ERC on text, starting by the use of Recurrent Neural Networks (RNN) (Rumelhart et al., 1985; Jordan, 1986) by Poria et al. (2017). Further work using recurring structures followed, such as DialogueRNN (Majumder et al., 2019; Ghosal et al., 2020). This model leverages the attention mechanism (Bahdanau et al., 2014) combined with RNN. Graph-based methods also proved to be efficient as shown in (Ghosal et al., 2019), not only as such but also when considering external knowledge, as Lee and Choi (2021) use a Graph Convolutional Network (GCN) to perform ERC by extracting relations between dialogue instances.

Existing work on ERC relies mainly on evaluating their model using a micro F1 score excluding the majority neutral label. However, recent work actually skipped this evaluation to instead focus only on the macro version of this metric (Pereira et al., 2023), while other considered the Matthew Coefficient Correlation as an indication suitable for this task (Guibon et al., 2021).

In this work, we focus on DailyDialog, which consists of artificially human-generated conversations about daily life concerns, with utterance-wise emotion labeling. Liang et al. (2022) propose a model based on Graph Neural Networks (GNN) and Conditional Random Fields (Lafferty et al., 2001) (CRF) that achieves 64.01% in micro F1.

Although it is known not to provide the best performance compared to few-shot learning approaches (Dumoulin et al., 2021), meta-learning allows better generalization through more robust training (Finn et al., 2017; Antoniou et al., 2019), which is particularly adapted in the case of emotion detection due to both variability and complexity of human feelings (Plutchik, 2001).

Metric learning. As reviewed by (Hospedales et al., 2022), a meta-learning approach consists in a *meta-optimizer* that describes meta-learner updates, a *meta-representation* that stores the acquired knowledge and the *meta-objective* oriented towards the desired task. This optimization-based meta-learning setup provides end-to-end algorithms often based on episodic scenarios (Ravi and Larochelle, 2016; Finn et al., 2017; Mishra et al., 2017) that reflect the "learning to learn" strategy. Besides, learning to learn implies second order gradient computations which is costly. Palliative solutions to this problem, such as implicit differentiation (Lorraine et al., 2020), still involve a trade-off between performance and memory cost (Hospedales et al., 2022). Therefore, variants has emerged such as *metric learning*, which meta-objective is to learn the meta-representation itself. Starting with Siamese Networks (Koch et al., 2015), this model structure leverages parameter sharing between identical sub-networks to learn a distance between data samples. Relation Networks (Sung et al., 2018) also consider a distance metric, departing from the traditional Euclidean approach. Matching Networks (Vinyals et al., 2016) leverage training examples to identify weighted nearest neighbors. Prototypical Networks (Snell et al., 2017) compute average class representations and utilize cosine distance for element comparison.

[1] https://github.com/B-Gendron/sentEmoContext

This model has been adapted to perform ERC in a few-shot setting by Guibon et al. (2021) in a way that outperformed few-shot learning baselines.

In this work, we focus on the Siamese Networks architecture. It has the advantage of being conceptually simple, which makes it easily controllable and scalable. Nevertheless, the model structure proposed in this paper is easily adaptable to more complex meta-learning setups. Siamese Networks have been used, for example, in NLP for intention detection on text (Ren and Xue, 2020), in computer vision for facial recognition (Hayale et al., 2023), and in complex representation learning (Jin et al., 2021).

3 Methodology

In this work, we use a metric-learning architecture based on learning emotions as they relate to each other, thus extracting meta-information from the data. The model is a Siamese network (Koch et al., 2015) with three identical sub-networks, whose outputs are compared using the triplet loss (Schultz and Joachims, 2003). Initially applied to computer vision problems (Chechik et al., 2010; Schroff et al., 2015), triplet loss is defined on a triplet of data samples (a, p, n) so that if a and p belong to the same class and n belongs to a different class, then:

$$\mathcal{L}(a, p, n) = \max\{d(a, p) - d(a, n) + \text{margin}, 0\}$$

where the margin parameter is a strictly positive number.

While the triplet loss could be used in several strategies, ranging from only retrieving the most difficult triplets (when the positive is far from the anchor, meanwhile the anchor is close to the negative) to skipping the most easy ones (i.e. when the positive is closer to the anchor), we only tackle the overall strategy by considering each triplet in our data, due to the limited size of the data.

Isolated representations. As the aim of our experiments is to characterize the contribution of conversational context to emotion prediction, we first developed a baseline model on isolated utterances. This formally refers to computing emotion predictions for utterances independently of their context. To do this, we first consider a mapping for each utterance word to its associated FastText embedding (Bojanowski et al., 2017). From such embeddings, aforementioned (a, p, n) triplets are randomly sampled and given as input for the Siamese Network, whose sub-network gradually improves in emotion prediction as triplet loss backpropagates.

Contextual representations. Regarding the contextual case, we build contextual utterance representations upon a BERT-like encoding. Sentence embeddings are preferred to word-piece embeddings (like BERT produces) as they provide lighter utterance representations. After the dialog is mapped to its associated series of pretrained embeddings, these outputs are concatenated forming a dialog representation, and contextual information is considered by deploying attention over it. Concretely, a Transformer encoder layer is stacked to the gathered frozen pre-trained embeddings. This newly conversation-aware dialog representation is then split at `[SEP]` tokens to end up with contextual representations at the utterance level, on which the emotion prediction is performed. In order to fit contextual utterance representations to the emotion prediction objective, we add an emotion classifier that is pre-trained on DailyDialog training set. The classifier is not frozen to ensure a complete backpropagation. Meanwhile, contextual representations are optimized according to the metric learning objective, using a triplet loss. The whole training procedure is illustrated in Figure 1. This training scenario enables both individual and relative emotion learning, in such a way that each learning phase strengthens the other. Thanks to this meta-learning setting, meta-information about emotions is extracted, and we can expect that this model is able to achieve relevant classification on unseen labels in a few-shot setting.

4 Experimental Protocol

Data. All the experiments have been carried out on DailyDialog dataset (Li et al., 2017) that provides more than 10,000 dialogues about daily concerns along with utterance-wise emotion labeling. In addition to providing utterance-level emotion labeling, an advantage in using DailyDialog is that it is relatively small, therefore it is quite easy to handle the entries and run tests on it. There exist six emotional labels (anger, disgust, fear, happiness, sadness and surprise) and a neutral label. Regarding emotion prediction, the evaluation is carried out only on the emotional labels following previous work procedure (Ghosal et al., 2021; Zhong et al., 2019). We use the original dataset splits (train, validation and test) from Li et al. (2017). The main characteristics from DailyDialog dataset

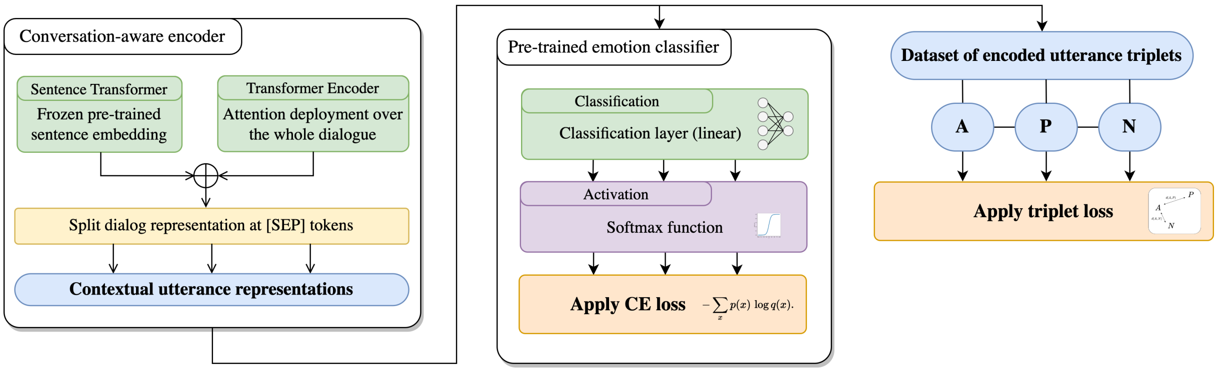

Figure 1: Illustration of the three main steps of the training procedure in the case of conversation-aware emotion predictions. Both losses (CE and triplet) backpropagate in order to gradually improve the encoder.

are visible in Table 1.

Daily Dialog Stats	
Language	English
Max Msg/Conv	35
Avg Msg/Conv	8
Labels	7
Emotion Labels	6
Nb. Conv.	13,118

Table 1: Main statistics for DailyDialog dataset

Model specificities. For the isolated utterance model, we consider two different types of sub-networks being simple linear layers and Long Short-Term Memory layers (LSTM) (Hochreiter and Schmidhuber, 1997). In the contextual case, the sub-network is a Transformer encoder fed with sentence embeddings. We carried out experiments with three different models of pre-trained sentence Transformers available in the Python library `sentence transformers`[2]: MPNet (Song et al., 2020), MiniLM (Wang et al., 2020) and RoBERTa (Liu et al., 2019). In order to ensure a good balance, the (a, p, n) triplets are made at this stage, meaning right before applying the pre-trained emotion classifier, which is composed of a linear layer stacked upon one Transformer encoder layer.

Training specificities. Whether it be for the isolated utterance model or for the contextual one, the emotion prediction is always performed at the utterance level, therefore the triplets are always utterance triplets. This involves balance issues as DailyDialog dataset is very imbalanced regarding emotion labels (Figure 4). Indeed, the class rebalancing induced by sampling triplets according to a uniform distribution does not sufficiently mitigate bias during training and prevents the loss from converging due to excessive oversampling in frequent classes. Thus, we addressed the imbalance problem all along the training pipeline, by implementing a random sampler weighted with inverse label frequencies to account for the rareness of some emotional labels like `fear` or `disgust`.

Evaluation. For quantitative evaluation we needed to account for both performance and relevancy of the training procedure so that generalization abilities enabled by the meta-learning architecture are actually usable. This way, we selected, in addition to usual performance metrics, a more demanding metric: Matthews Correlation Coefficient (MCC) (Cramér, 1946). This measures a Pearson correlation (Pearson, 1895) between the predicted and the actual class, giving more precise information on classification quality than F1 score (Baldi et al., 2000). Using TP. TN. FP and FN as respectively the number of true positives, true negatives, false positives and false negatives, P and R being respectively precision and recall, and N the total number of samples, MCC was originally defined in (Matthews, 1975) as:

$$\text{MCC} = \frac{TP/N - R \times P}{\sqrt{PR(1-R)(1-P)}} \qquad (1)$$

Comparison with LLMs. In order to place the results of our isolated and contextual models into perspective, we compare our models with state-of-the-art LLMs, namely LLaMA (Touvron et al., 2023) and Falcon (Penedo et al., 2023). Both are considered with instruction fine-tuning and evaluated on text generation inference in a zero-shot

[2] https://www.sbert.net/

setting. We developed a prompt asking for prediction on the last utterance of each DailyDialog test set dialog, regarding the conversational context. For both LLMs, we went through an iterative process to find the most adapted prompt in the sense that the model actually generates only one label. The prompt is the same for each model of the same type (either LLaMA or Falcon). We experienced more difficulty on prompt tuning with Falcon as the model generates `happiness` on 86% of DailyDialog test set. Both prompts full texts are provided in Figure 2.

```
Here is a dialog :

-  Hello , Miao Li , Where are you going ?
-   Hello , I am going to the store to buy some fruit .
-   Oh , Would you do me a favor ?
-   Yes ?
-   Please mail this letter for me on your way to the store .
-   Sure . Do you want it to be registered ?
-   Yes , I think so . There are some pictures in it . It would be a great
pity if they were lost .
-   Yes , I will be glad to mail your letter .
-   Thanks .
-   you are welcome .

Regarding its conversational context, give me the appropriate emotion to
describe this utterance : "Yes , I think so . There are some pictures in
it . It would be a great pity if they were lost .", using only one of the
following labels: happiness, sadness, anger, surprise, fear, disgust, no
emotion. Predicted label :
```

(a) Prompt for LLaMA

```
Here is a dialog :

-  Hello , Miao Li , Where are you going ?
-   Hello , I am going to the store to buy some fruit .
-   Oh , Would you do me a favor ?
-   Yes ?
-   Please mail this letter for me on your way to the store .
-   Sure . Do you want it to be registered ?
-   Yes , I think so . There are some pictures in it . It would be a great
pity if they were lost .
-   Yes , I will be glad to mail your letter .
-   Thanks .
-   you are welcome .

Regarding its conversational context, return the appropriate emotion for
the last utterance among: sadness, happiness, anger, surprise, fear and
disgust. If none of them properly correspond, return 'no emotion'.
```

(b) Prompt for Falcon

Figure 2: Prompts for LLaMA and falcon

5 Results

Table 2 gives an overview of the different results obtained by the research community on ERC with DailyDialog. This actually shows a slow progression since 2017 where Poria et al. (2017) proposed to evaluate the model on the micro F1 score excluding the majority class (i.e., the neutral class). This became the first baseline for this task, achieving 50.24 in micro F1 score. However, the current state-of-the-art model now achieves 64.07 in micro F1 score (Liang et al., 2022) which amounts to a 14 points improvement during 6 years. As visible in Table 2, the community mainly followed this pattern and evaluation scheme. However, in

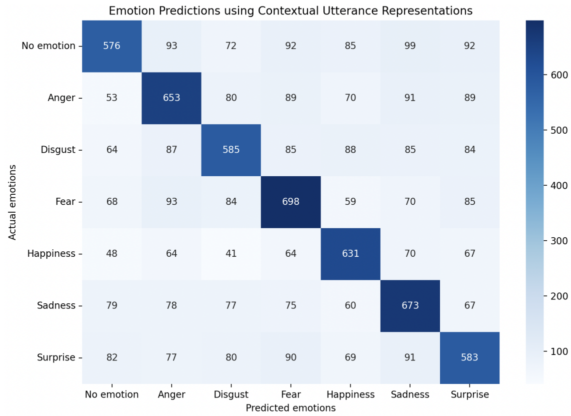

Figure 3: Confusion matrix for emotion predictions using contextual utterance representations

this paper, we think it is important to also consider the macro F1 score, excluding the majority class, as it shows the overall performance in all emotions. Some work has already decided to do so since 2020 (Ghosal et al., 2020), leading to an improvement of ~2.5 points in 3 years. Following this idea, Figure 3 and Table 3 illustrate this adaptability in emotion prediction showing the detailed classification results.

Compared to these results, our SentEmoContext model achieves 57.75 in micro F1 score, which is a decent but somewhat modest result in terms of metric comparison. However, Table 2 also shows the average performance of our model over 10 runs. Our SentEmoContext is state-of-the-art on the macro F1 score with 57.71 points, outperforming CD-ERC (Pereira et al., 2023) by 6.48 points, which is considerable since they only focused on this metric, and TODKAT (Zhu et al., 2021) by 5.15 points. We also evaluate our model using the multiclass MCC (Matthews, 1975; Baldi et al., 2000) score to ensure that the model does not arbitrary decide. Given an MCC score range of -1 to 1, and 0 indicating randomness, the 0.49 MCC score of the SentEmoContext model indicates that our approach is balanced and accurate in terms of predictions (Chicco and Jurman, 2020). Of course, we cannot compare with other ERC works with the MCC metric, as they did not use it. However, we think it is important to consider it as an additional metric to indicate the quality of the classification, minimizing the effect of the highly imbalanced data from conversations.

Given these results, our SentEmoContext performs really well considering that we only need ~20 minutes per epoch on GPU Nvidia A40 (45 GB

Model name	macro F1*	micro F1*	MCC
State-of-the-art models on ERC			
CNN+cLSTM (Poria et al., 2017)	–	50.24	–
KET (Zhong et al., 2019)	–	53.37	–
COSMIC (Ghosal et al., 2020)	51.05	58.48	–
RoBERTa (Ghosal et al., 2020)	48.20	55.16	–
Rpe-RGAT (Ishiwatari et al., 2020)	–	54.31	–
Glove-DRNN (Ghosal et al., 2021)	41.8	55.95	–
roBERTa-DRNN (Ghosal et al., 2021)	49.65	57.32	–
CNN (Ghosal et al., 2021)	36.87	50.32	–
DAG-ERC (Shen et al., 2021)	–	59.33	–
TODKAT (Zhu et al., 2021)	<u>52.56</u>	58.47	–
SKAIG (Li et al., 2021)	51.95	59.75	–
Sentic GAT (Tu et al., 2022)	–	54.45	–
CauAIN (Zhao et al., 2022)	–	58.21	–
DialogueRole (Ong et al., 2022)	–	60.95	–
S+PAGE (Liang et al., 2022)	–	**64.07**	–
DualGAT (Zhang et al., 2023)	–	<u>61.84</u>	–
CD-ERC (Pereira et al., 2023)	51.23	–	–
Llama2-7b (Touvron et al., 2023)	9.70	24.92	0.08
Llama2-13b (Touvron et al., 2023)	22.26	43.37	0.15
Falcon-7b (Penedo et al., 2023)	07.54	42.75	0.01
MCM-CSD (Xu and Yang, 2024)	–	60.70	–
Ours			
SentEmoContext	**57.71**	57.75	**0.49**

Table 2: All results for ERC on DailyDialog. Metrics are all computed on the official test set. DRNN stands for DialogueRNN as it is called in the original paper. MCC = Matthew Coefficient Correlation. The * indicates metrics that do not include the neutral label.

Emotion	P	R	F1	Supp.
No emotion	0.594	0.519	0.554	1109
Anger	0.570	0.580	0.575	1125
Disgust	0.574	0.543	0.558	1078
Fear	0.585	0.603	0.594	1157
Happiness	0.594	0.641	0.617	985
Sadness	0.571	0.607	0.588	1109
Surprise	0.546	0.544	0.545	1072

Table 3: Emotion prediction details using contextual utterances. F1 is the F1-score for each class, and Supp. is the support. P is precision and R is recall.

RAM) and train it using only 5 epochs. This makes a striking difference from existing approaches that use multiple streams per speaker (Pereira et al., 2023), graph modeling for the representation of context and knowledge (Zhong et al., 2019; Li et al., 2021), or other heavy representations in their model (Liang et al., 2022). In addition to this, our model is stable with a standard deviation of only 0.01 on average across the three metrics, which reinforces the quality of such an efficient approach.

5.1 Comparison with Emotion Classifiers on Utterance Level

Table 4 shows the results of the direct emotion classification on utterances. For this task, we only considered the 6 emotion labels, excluding the neutral one not only from the evaluation but also from the training. By doing so, we want to determine the difference between our approach and a dedicated emotion classifier. This also serves as an ablation study for our SentEmoContext model, since this step is part of its training. With Table 4, we can see that our model leverages both the embedded conversational context and the metric learning scheme to increase all metrics. We can especially note the difference in terms of macro F1 scores, which shows the importance of the triplet loss representation in our model. Indeed, the emotion utterance classifiers are trained using batches balanced on the whole training set distribution and a weighted cross-entropy loss. Results show that it is not enough to deal with extreme imbalanced data such as conversations.

5.2 LLMs Results

The LLM results in a zero shot setting are visible in Table 5. These serve as an indication on the performance of such models, albeit in their lightweight version, in the ERC task. Although these generative models are not designed for this quite peculiar task, they still manage to outperform the utterance emotion classifiers of Table 4, which can be considered as a display of emergent capacities of LLMs (Srivastava et al., 2022).

5.3 Imbalance Factor

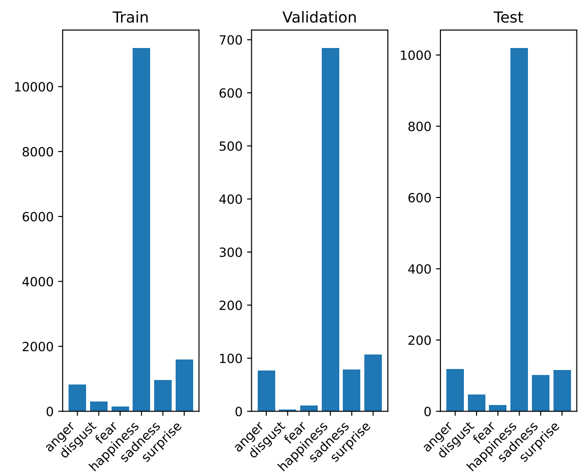

Figure 4: Histograms of only the emotion label distribution in DailyDialog subsets.

Although Table 1 shows the characteristics of the dataset, it omits to present the main characteristic of the conversational data in terms of emotion labels: the extreme imbalance. Most of the difficulty in ERC comes from the label definition, the context, but also from the imbalance factor that prevents the model from easily learning the representation of emotions in the context. Figure 4 shows the distribution of the labels in DailyDialog, without the neutral one. Considering the latter is the majority label and is excluded from the evaluation metrics by all the ERC community. The fact that even in the emotion labels the data is that imbalanced proves to be challenging and needs to be addressed. In fact, we are derived from Guibon et al. (2023) to tackle the imbalance in two steps. First, we balance the data loader to produce more balanced batches given the training set weights. Second, we weight the cross-entropy loss from the emotion classifier considering the remaining imbalance on each batch.

In addition to this, we add another way to address the imbalance. By considering triplets, we remove the imbalance factor while using hidden states that come from balanced representation. We think this partly explains the effectiveness and the efficiency of our model, considering its limited size compared to the related work.

6 Discussion

6.1 Model Size and Efficiency

Our SentEmoContext is efficient. It produces state-of-the-art results on macro F1 score and good results on micro F1. However, our model trains relatively fast and does not require a lot of epochs to converge. We think this efficiency, along with the limited memory needed to train, is due to both our two-step backpropagation and to the fact that we are using utterance-embedded representations with sentence transformers. Thus, our model can efficiently tackle long conversational contexts with limited memory cost.

In addition, Table 6 shows the difference between the models we used in terms of size, parameters, and number of layers. Our model is relatively small considering the recent advances and related work in ERC, but also compared to LLMs.

6.2 Relative Label Representation

Our approach actually learns twice from the data, first by using a supervised setting, and then by actually considering the relative distances between encoded elements, updating through the triplet loss. This enables the use of our model to different conversation datasets with different labels. The only requirement to extend the scope of this model would be to consider another triplet sampling strategy ignoring labels, such as the batch-hard strategy (Do et al., 2019).

7 Conclusion

In this paper, we present our SentEmoContext model, which comes from an approach that mixes utterance level representation, metric learning, and Siamese Networks. This model efficiently represents the conversational context, which makes

Model name	**macro F1**	**micro F1**	**MCC**
	Pre-trained emotion utterance classifier		
all-MiniLM-L6-v2	20.22	33.11	0.40
	Ours		
SentEmoContext	**57.71**	**57.75**	**0.49**

Table 4: Comparison with a direct emotion classification at the utterance level. The `all-MiniLM-L6-v2` fine-tuning is also part of the whole SentEmoContext approach.

Model name	**P**	**R**	**macro F1***	**micro F1***	**MCC**
llama2-7b-chat-hf	26.77	24.77	9.70	24.92	0.08
llama2-13b-chat-hf	32.63	83.49	22.26	43.37	0.15
falcon-7b-instruct	–	–	07.54	42.75	0.01

Table 5: Results using two open-source LLMs with specific prompts. An example of the prompt is shown in Figure 2. * indicates metrics that do not include the neutral label.

Model name	**Seq. Length**	**Tokens**	**Dimensions**	**Size**	**Parameters**	**Tr. Layers**
			Pre-trained sentence transformers			
all-MiniLM-L6-v2	256	1bn+	384	80 MB	22M	6
all-mpnet-base-v2	384	1bn+	768	420 MB	110M	12
			State-of-the-art LLMs			
Llama-2-7b-chat-hf	4096	2T	11008	13 GB	7B	32
Llama-2-13b-chat-hf	4096	2T	11008	25 GB	13B	32
falcon-7b-instruct	2048	1.5T	4544	15 GB	7B	32
			Ours			
SentEmoContext	256	4M	384	604.8 MB	159M	6

Table 6: Insights about model sizes, comparing the pretrained sentence Transformers used in our approach to state-of-the-art LLMs. These insights demonstrate that SentEmoContext provides a lightweight yet efficient way to perform ERC on DailyDialog.

it achieve state-of-the-art macro F1 score with 57.71, and satisfactory micro F1 scores with 57.75 on the Emotion Recognition in Conversation on DailyDialog. We also propose to use the Matthew Correlation Coefficient to better evaluate this task.

With SentEmoContext we use contrastive learning with balanced samplers to minimize the imbalance factor, which is inherent to conversational data. We also leverage sentence BERT to both minimize the memory required for training considering the whole conversational context and to actually represent the conversational context by considering utterances as the minimal unit. This led to a more robust and efficient training method that does not require a lot of epochs to obtain satisfactory results. We also show that small- to average-size open-source LLMs are still behind on emotion recognition in conversation, as it requires a lot of context to be incorporated in the prompt and is not specifically relevant to generative models.

In our future work, we want to consider applying this approach to other datasets, with added modalities, to stress-test our model. We also plan to use it on slightly different labels, as our model learns relative positions toward labels. Thus, we plan to adapt it to a setting leaning towards meta-learning.

8 Limitations

The first limitation we faced with LLMs is the requirement of high-memory GPUs to test them.

This explains why in Table 5 we only consider the lightweight version of these two open source LLMs. While LLaMA 7b and 13b gave answers in a good format, i.e. with only one label chosen, Falcon did not behave the way we wanted. In order to solve this, we look for the first mentioned emotion in the output to consider it as a label.

Also, it is important to note that we did not want to tackle OpenAI's ChatGPT due to the fact that we do not have a clear control on the model version, size and approach used behind its API, but also because we wanted to consider open source models, and open source data as we will release both our models and source code to the community. Moreover, we limited ourselves to LLaMA 2 as experiments were performed prior to the release of LLaMA 3.

An additional possible limitation on LLMs is the context size. In ERC, context size is key, but with LLMs adding examples in the prompt to do few-shot learning would take a lot of space in the overall context, the prompt being part of the context. This explains our decision to only consider zero-shot in this paper for LLMs, even though we should also consider prompt tuning to enhance them on this specific task.

Acknowledgments

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS. RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

References

Antreas Antoniou, Harri Edwards, and Amos Storkey. 2019. How to train your maml. In *Seventh International Conference on Learning Representations, ICLR*.

Dzmitry Bahdanau, Kyunghyun Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015*, 1409.

Pierre Baldi, Søren Brunak, Yves Chauvin, Claus Andersen, and Henrik Nielsen. 2000. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics (Oxford, England)*, 16:412–24.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. 2010. Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.*, 11:1109–1135.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6.

Harald Cramér. 1946. *Mathematical Methods of Statistics (PMS-9), Volume 9*. Princeton University Press, Princeton.

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.

Thanh-Toan Do, Toan Tran, Ian Reid, Vijay Kumar, Tuan Hoang, and Gustavo Carneiro. 2019. A theoretically sound upper bound on the triplet loss for improving the efficiency of deep distance metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10404–10413.

Vincent Dumoulin, Neil Houlsby, Utku Evci, Xiaohua Zhai, Ross Goroshin, Sylvain Gelly, and Hugo Larochelle. 2021. A unified few-shot classification benchmark to compare transfer and meta learning approaches. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Chunxiao Fan, Jie Lin, Rui Mao, and Erik Cambria. 2024. Fusing pairwise modalities for emotion recognition in conversations. *Information Fusion*, page 102306.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1126–1135. JMLR.org.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP*

2020, pages 2470–2481, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1435–1449, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Gaël Guibon, Matthieu Labeau, Hélène Flamein, Luce Lefeuvre, and Chloé Clavel. 2021. Few-shot emotion recognition in conversation with sequential prototypical networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic.

Gaël Guibon, Matthieu Labeau, Luce Lefeuvre, and Chloé Clavel. 2023. An adaptive layer to leverage both domain and task specific information from scarce data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7757–7765.

Wassan Hayale, Pooran Singh Negi, and Mohammad H. Mahoor. 2023. Deep siamese neural networks for facial expression recognition in the wild. *IEEE Transactions on Affective Computing*, 14(2):1148–1158.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2022. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5149–5169.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370, Online. Association for Computational Linguistics.

Ming Jin, Yizhen Zheng, Yuan-Fang Li, Chen Gong, Chuan Zhou, and Shirui Pan. 2021. Multi-scale contrastive siamese networks for self-supervised graph representation learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1477–1483. International Joint Conferences on Artificial Intelligence Organization. Main Track.

M I Jordan. 1986. Serial order: a parallel distributed processing approach. technical report, june 1985-march 1986.

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.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Bongseok Lee and Yong Suk Choi. 2021. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023. Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing*, 550:126427.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1204–1214, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1610–1618, Dublin, Ireland. Association for Computational Linguistics.

Chen Liang, Jing Xu, Yangkun Lin, Chong Yang, and Yongliang Wang. 2022. S+PAGE: A speaker and position-aware graph neural network model for emotion recognition in conversation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 148–157, Online only. Association for Computational Linguistics.

Xingwei Liang, You Zou, Xinnan Zhuang, Jie Yang, Taiyu Niu, and Ruifeng Xu. 2023. Mmateric: Multi-task learning and multi-fusion for audiotext emotion recognition in conversation. *Electronics*, 12(7):1534.

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. *Preprint*, arXiv:1907.11692.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. 2020. Optimizing millions of hyperparameters by implicit differentiation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1540–1552. PMLR.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825.

Brian W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et biophysica acta*, 405 2:442–51.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and P. Abbeel. 2017. A simple neural attentive meta-learner. In *International Conference on Learning Representations*.

Donovan Ong, Jian Su, Bin Chen, Anh Tuan Luu, Ashok Narendranath, Yue Li, Shuqi Sun, Yingzhan Lin, and Haifeng Wang. 2022. Is discourse role important for emotion recognition in conversation? *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11121–11129.

Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *Preprint*, arXiv:2306.01116.

Patrícia Pereira, Helena Moniz, Isabel Dias, and Joao Paulo Carvalho. 2023. Context-dependent embedding utterance representations for emotion recognition in conversations. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 228–236, Toronto, Canada. Association for Computational Linguistics.

Robert Plutchik. 2001. The Nature of Emotions. *American Scientist*, 89(4):344.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*.

Fuji Ren and Siyuan Xue. 2020. Intention detection based on siamese neural network with triplet loss. *IEEE Access*, 8:82242–82254.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1985. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Matthew Schultz and Thorsten Joachims. 2003. Learning a distance metric from relative comparisons. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

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.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4080–4090, Red Hook, NY, USA. Curran Associates Inc.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aarohi Srivastava et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3637–3645, Red Hook, NY, USA. Curran Associates Inc.

Binqiang Wang, Gang Dong, Yaqian Zhao, Rengang Li, Qichun Cao, Kekun Hu, and Dongdong Jiang. 2023. Hierarchically stacked graph convolution for emotion recognition in conversation. *Knowledge-Based Systems*, 263:110285.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

Jintao Wen, Dazhi Jiang, Geng Tu, Cheng Liu, and Erik Cambria. 2023. Dynamic interactive multiview memory network for emotion recognition in conversation. *Information Fusion*, 91:123–133.

Yuan Xu and Meng Yang. 2024. Mcm-csd: Multi-granularity context modeling with contrastive speaker detection for emotion recognition in real-time conversation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11956–11960. IEEE.

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.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4524–4530. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. 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 1571–1582, Online. Association for Computational Linguistics.