Context-Aware Siamese Networks for Efficient Emotion Recognition in Conversation
Barbara Gendron, Gaël Guibon

To cite this version:
Barbara Gendron, Gaël Guibon. Context-Aware Siamese Networks for Efficient Emotion Recognition in Conversation. 2024. hal-04532408

HAL Id: hal-04532408
https://hal.science/hal-04532408
Preprint submitted on 16 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License
Context-Aware Siamese Networks for Efficient Emotion Recognition in Conversation

Barbara Gendron¹,² Gaël Guibon¹

¹ LORIA, Université de Lorraine, CNRS
² Université du Luxembourg
(firstname.lastname)@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 to 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 regarding the use of metric learning for emotion recognition, yet perfectible compared to the microF1 score obtained.

Keywords: emotion recognition in conversation, metric learning

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 and graph-related approaches 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 variant of emotions that goes 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 both using 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, as far as we know, the 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 requiring emotion recognition of different label granularities.

Our main contribution lies in the development of a metric-learning training strategy for emotion recognition on utterances using 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 Lan-
guage Models (LLMs) such as light versions of Falcon 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 DailyDialog in macro F1 score with 57.71% and satisfactory results on micro F1 with 57.75%.

In the following sections, we first review related work on ERC (Section 2). We then dive in our methodology (Section 3) and describe the experimental setup we use (Section 4). We then evaluate our models compared to a baseline without any conversational context and to SotA models for ERC in Section 5. Finally, we end up with key findings and perspectives for future work in section 6.

We will make our code and models available on github and Hugging Face models hub.

2. Related Work

ERC. Although most of the studies on ERC has been held on multi-modal 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 multi-modal datasets restricted to text such as IEMOCAP (Bussos et al., 2008) or MELD (Poria et al., 2019), or on 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) encountered in Transformer architecture (Vaswani et al., 2017). Graph-based methods also proved 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 micro F1 score excluding the majority neutral label. However, recent work actually skipped this evaluation to instead only focus 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 in artificially human-generated conversations about daily life concerns, with utterance-wise emotion labelling. Liang et al. (2022) propose a model based on Graph Neural Networks (GNN) and 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 a 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 the meta-representation learning 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. 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 Siamese Network architecture. It has the advantage to be conceptually simple, rendering 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 instance, 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 to learn 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.

Figure 1: Illustration of the triplet loss principle.
Given a triplet \((A, P, N)\) corresponding to respectively anchor, positive and negative, the positive sample should be closer to the anchor than the negative sample in order to minimize the triplet loss.

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 triplets 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 \(\text{[SEP]}\) 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 triplet loss. The whole is illustrated in figure 2. 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 labelling. 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 are visible in Table 1.

<table>
<thead>
<tr>
<th>Daily Dialog Stats</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td>Max Msg/Conv</td>
<td>35</td>
</tr>
<tr>
<td>Avg Msg/Conv</td>
<td>8</td>
</tr>
<tr>
<td>Labels</td>
<td>7</td>
</tr>
<tr>
<td>Emotion Labels</td>
<td>6</td>
</tr>
<tr>
<td>Nb. Conv.</td>
<td>13,118</td>
</tr>
</tbody>
</table>

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: 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 highly 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, MCC was originally defined in (Matthews, 1975) as:

$$ MCC = \frac{TP/N - S \times P}{\sqrt{PS(1-S)(1-P)}} \tag{1} $$

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 setting. We developed a prompt asking for prediction on the last utterance of each DailyDialog test set dialog, regarding the conversational context. For both LLM, 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 3.

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...
<table>
<thead>
<tr>
<th>Model name</th>
<th>macroF1*</th>
<th>microF1*</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN+cLSTM (Poria et al., 2017)</td>
<td>50.24</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>KET (Zhong et al., 2019)</td>
<td>53.37</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>COSMIC (Ghosal et al., 2020)</td>
<td>51.05</td>
<td>58.48</td>
<td>–</td>
</tr>
<tr>
<td>RoBERTa (Ghosal et al., 2020)</td>
<td>48.20</td>
<td>55.16</td>
<td>–</td>
</tr>
<tr>
<td>Rpe-RGAT (Ishiwatari et al., 2020)</td>
<td>–</td>
<td>54.31</td>
<td>–</td>
</tr>
<tr>
<td>Glove-DRNN (Ghosal et al., 2021)</td>
<td>51.05</td>
<td>58.48</td>
<td>–</td>
</tr>
<tr>
<td>roBERTa-DRNN (Ghosal et al., 2021)</td>
<td>54.65</td>
<td>57.32</td>
<td>–</td>
</tr>
<tr>
<td>CNN (Ghosal et al., 2021)</td>
<td>36.87</td>
<td>50.32</td>
<td>–</td>
</tr>
<tr>
<td>DAG-ERC (Shen et al., 2021)</td>
<td>–</td>
<td>59.33</td>
<td>–</td>
</tr>
<tr>
<td>TODKAT (Zhu et al., 2021)</td>
<td>52.56</td>
<td>58.47</td>
<td>–</td>
</tr>
<tr>
<td>SKAIG (Li et al., 2021)</td>
<td>51.95</td>
<td>59.75</td>
<td>–</td>
</tr>
<tr>
<td>Sentic GAT (Tu et al., 2022)</td>
<td>–</td>
<td>54.45</td>
<td>–</td>
</tr>
<tr>
<td>CauAIN (Zhao et al., 2022)</td>
<td>–</td>
<td>58.21</td>
<td>–</td>
</tr>
<tr>
<td>DialogueRole (Ong et al., 2022)</td>
<td>–</td>
<td>60.95</td>
<td>–</td>
</tr>
<tr>
<td>S+PAGE (Liang et al., 2022)</td>
<td>–</td>
<td>64.07</td>
<td>–</td>
</tr>
<tr>
<td>DualGAT (Zhang et al., 2023)</td>
<td>–</td>
<td>61.84</td>
<td>–</td>
</tr>
<tr>
<td>CD-ERC (Pereira et al., 2023)</td>
<td>51.23</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Llama2-7b (Touvron et al., 2023)</td>
<td>9.70</td>
<td>24.92</td>
<td>0.08</td>
</tr>
<tr>
<td>Llama2-13b (Touvron et al., 2023)</td>
<td>22.26</td>
<td>43.37</td>
<td>0.15</td>
</tr>
<tr>
<td>Falcon-7b (Penedo et al., 2023)</td>
<td>07.54</td>
<td>42.75</td>
<td>0.01</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>57.71</strong></td>
<td><strong>57.75</strong></td>
<td><strong>0.49</strong></td>
</tr>
</tbody>
</table>

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.

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. On the other hand, the current SotA 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 this paper we think it is important to also consider the macro F1 score, excluding the majority class, as it shows the overall performance on all emotions. Some work already decided to do so since 2020 (Ghosal et al., 2020), leading to an improvement of 2.5 points in 3 years. This reinforces the claim that the ERC task is indeed challenging.

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 SotA 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 in order to ensure the model is not deciding randomly. Given a MCC score ranges from -1 to 1, and 0 indicating randomness, the 0.49 MCC score of SentEmoContext model indicates our approach is both balanced and accurate in terms of predictions (Chicco and Jurman, 2020). Of course, we cannot compare to 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 we only need 20 minutes per epoch, and train it using only 5 epochs. This makes a striking difference with existing approaches using multiple streams per speaker (Pereira et al., 2023), graph modeling for context and knowledge representation (Zhong et al., 2019; Li et al., 2021), or other heavy representation 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.
Here is a dialog:
- Hello, Mike 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.

(a) Prompt for llama

Here is a dialog:
- Hello, Mike 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.

(b) Prompt for Falcon

Figure 3: Prompts for llama and falcon

5.1. Comparison with Emotion Classifiers on Utterance Level

Table 3 shows the results of 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 serve as an ablation study for our SentEmoContext model since this step is part of its training. With Table 3, we can see 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 macroF1 score, 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 shows it is not enough to deal with an extreme imbalanced data such as conversations.

5.2. LLM-related Limitations

LLMs results on a zero-shot setting are visible in Table 4. These serves as an indication on the performance of such models, albeit in their lightweight version, in the ERC task. Even though these generative models are not designed for this quite peculiar task, they still manage to outperform utterance emotion classifiers from Table 3, which can be considered as a display of emergent capacities from LLMs (Srivastava et al., 2022).

5.3. Imbalance Factor

While Table 1 shows the characteristics from the dataset, it omits to present the main characteristic from conversational data in terms of emotion labels: the extreme imbalance. Most of the difficulty from ERC comes from the label definition, the context but also from the imbalance factor that prevents the model from easily learning emotion representation 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. We actually stem from Guibon et al. (2023) to tackle the imbalance characteristic in two-steps. First, we balance the data loader to yield somewhat 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, in this paper 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. Discussions & Limitations

The work we present in this paper still possesses some limitations. We hereby draw some conclusion from them.

#### 6.1. LLMs Limitations

The first limitation we faced with LLMs is the requirement of high memory GPUs to test them. This explains why in Table 4 we only consider the lightweight version of these two open source LLM. 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.

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.

#### 6.2. Model Size and Efficiency

Our SentEmoContext is efficient. It yields state-of-the-art results on macro F1 score and good results on microF1. But 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 a limited cost in memory.

Moreover, Table 5 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.

#### 6.3. Relative Label Representation

Our approach actually learn twice from the data, first by using a supervised setting, and then by actually considering the relative distances between encoded element, 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 by ignoring the 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 mixing utterance level representation, metric learning and Siamese Networks. This model efficiently represent the conversational context, which makes it achieves state-of-the-art macroF1 score with 57.71, and satisfactory microF1 scores with 57.75.
<table>
<thead>
<tr>
<th>Model name</th>
<th>Seq. Length</th>
<th>Tokens</th>
<th>Dimensions</th>
<th>Size</th>
<th>Parameters</th>
<th>Tr. Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>all-MiniLM-L6-v2</td>
<td>256</td>
<td>1bn+</td>
<td>384</td>
<td>80 MB</td>
<td>22M</td>
<td>6</td>
</tr>
<tr>
<td>all-mpnet-base-v2</td>
<td>384</td>
<td>1bn+</td>
<td>768</td>
<td>420 MB</td>
<td>110M</td>
<td>12</td>
</tr>
<tr>
<td>State-of-the-art LLMs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Llama-2-7b-chat-hf</td>
<td>4096</td>
<td>2T</td>
<td>11008</td>
<td>13 GB</td>
<td>7B</td>
<td>32</td>
</tr>
<tr>
<td>Llama-2-13b-chat-hf</td>
<td>4096</td>
<td>2T</td>
<td>11008</td>
<td>25 GB</td>
<td>13B</td>
<td>32</td>
</tr>
<tr>
<td>falcon-7b-instruct</td>
<td>2048</td>
<td>1.5T</td>
<td>4544</td>
<td>15 GB</td>
<td>7B</td>
<td>32</td>
</tr>
<tr>
<td>Ours</td>
<td>256</td>
<td>4M</td>
<td>384</td>
<td>604.8 MB</td>
<td>159M</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 5: Insights about model sizes, comparing the pretrained sentence Transformers used in our approach to state-of-the-art LLMs.

on the Emotion Recognition 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 overcome 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 lead to a more robust and efficient training method that does not require a lot of epoches to obtain satisfactory results. We also show small to average size open source LLMs are still behind on emotion recognition in conversations as it requires a lot 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 on other dataset, with added modalities in order 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 more meta-learning setting.

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).

8. Bibliographical References


Guilherme Penedo, Quentin Malartic, Daniel Hessel, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Aloibeidi, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refined-web dataset for falcon llm: Outperforming curated corpora with web data, and web data only.


Weixiang Zhao, Yan Yan Zhao, and Xin Lu. 2022. 


9. Language Resource References

