SEC : contexte émotionnel phrastique intégré pour la reconnaissance émotionnelle efficiente dans la conversation

Barbara Gendron, Gaël Guibon

▶ **To cite this version:**

Barbara Gendron, Gaël Guibon. SEC : contexte émotionnel phrastique intégré pour la reconnaissance émotionnelle efficiente dans la conversation. 35èmes Journées d'Études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024), Jul 2024, Toulouse, France. pp.219-233. hal-04623019

HAL Id: hal-04623019

https://inria.hal.science/hal-04623019v1

Submitted on 1 Jul 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 4.0 International License

SEC : contexte émotionnel phrastique intégré pour la reconnaissance émotionnelle efficiente dans la conversation

Barbara Gendron[1, 2] Gaël Guibon[1]
(1) LORIA, Université de Lorraine, CNRS
(2) Université du Luxembourg
prénom.nom@loria.fr

RÉSUMÉ

L'essor des modèles d'apprentissage profond a apporté une contribution significative à la reconnaissance des émotions dans les conversations (ERC). Cependant, cette tâche reste un défi important en raison de la pluralité et de la subjectivité des émotions humaines. Les travaux antérieurs sur l'ERC fournissent des modèles prédictifs utilisant principalement des représentations de la conversation basées sur des graphes. Dans ce travail, nous proposons une façon de modéliser le contexte conversationnel que nous incorporons à une stratégie d'apprentissage de métrique, avec un processus en deux étapes. Cela permet d'effectuer l'ERC dans un scénario de classification flexible et d'obtenir un modèle léger et efficace. En utilisant l'apprentissage de métrique à travers une architecture de réseau siamois, nous obtenons un score de macroF1 de 57,71% pour la classification des émotions dans les conversations sur le jeu de données DailyDialog, ce qui surpasse les travaux connexes. Ce résultat état-de-l'art est prometteur en ce qui concerne l'utilisation de l'apprentissage de métrique pour la reconnaissance des émotions, mais est perfectible au regard du microF1 obtenu.

ABSTRACT

Context-Aware Metric Learning for Efficient Emotion Recognition in Conversation

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 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 macroF1 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.

MOTS-CLÉS : apprentissage profond, reconnaissance d'émotions en conversation, apprentissage de métrique.

KEYWORDS: deep learning, emotion recognition in conversation, metric learning.

1 Introduction

La communication médiée par ordinateur (CMO) est en constante évolution et de nouveaux moyens de communication apparaissent régulièrement. Avec l'avènement des agents conversationnels, il devient nécessaire de détecter les émotions au sein d'une conversation. Bien que plusieurs modalités soient désormais prises en compte dans le processus de communication, la modalité textuelle reste essentielle pour une communication quotidienne rapide et facile, par le biais d'outils comme les applications de messagerie ou les médias sociaux. La modalité textuelle est toutefois ambiguë, car elle ne préserve pas le contexte extra-linguistique présent par exemple dans les conversations dyadiques. L'une des principales ambiguïtés est l'état émotionnel de l'orateur, souvent mal interprété par les humains à travers des messages courts et non polis. Cela motive la reconnaissance d'émotions en conversation (*Emotion Recognition in Conversation*, ERC) qui vise non seulement à l'identification des émotions dans les messages, mais aussi à la prise en compte du contexte conversationnel pour reconnaître les émotions. L'ERC s'est révélée être un défi, notamment en ce qui concerne la représentation du contexte (Ghosal *et al.*, 2021). Récemment, les modèles multimodaux et les approches basées sur les graphes se sont multipliés. Ils représentent souvent le contexte conversationnel à travers un profil des locuteurs, ce qui est performant mais moins efficace, en plus de dépendre des étiquettes émotionnelles. Les approches existantes sont principalement supervisées et confrontées à un fort déséquilibre des étiquettes en raison de la rareté de certaines émotions.

Dans cet article, nous adressons ces deux défis en incorporant le contexte conversationnel dans un scénario d'apprentissage de métrique (que l'on désignera par *metric learning*), tout en contrôlant le déséquilibre des données de plusieurs façons. Dans notre cas, afin de rendre notre modèle utilisable pour d'autres émotions que les 6 émotions primaires (Ekman *et al.*, 1969), nous n'utilisons pas l'apprentissage contrastif supervisé (Khosla *et al.*, 2020) dans notre méthode. L'apprentissage de métrique permet justement de s'abstraire d'une dépendance aux définitions strictes des émotions, ce qui se révèle indispensable pour des émotions fines. Pour cela, nous mettons à jour le modèle en utilisant à la fois les prédictions d'étiquettes isolées (fonction de perte d'entropie croisée), et l'attribution d'étiquettes contextuelles relatives (fonction de perte contrastive). Ce processus en deux étapes est assez simple pour des énoncés isolés. Cependant, à notre connaissance, la représentation contextuelle par apprentissage contrastif pour l'ERC n'a pas encore été utilisée. Ceci représente notre principale contribution puisque nous présentons un modèle qui peut atteindre des performances compétitives par rapport à l'état-de-l'art tout en étant utilisable avec de nouvelles étiquettes d'émotions. Ainsi, notre modèle peut être appliqué et adapté dans de multiples contextes nécessitant la reconnaissance d'émotions à différents niveaux de granularité.

Notre principale contribution réside dans le développement d'une stratégie de *metric learning* pour la reconnaissance d'émotions en utilisant le contexte conversationnel. Le modèle présenté exploite des plongements lexicaux à l'échelle de la phrase et déploie de l'attention à l'aide d'un Transformer (Vaswani *et al.*, 2017; Devlin *et al.*, 2019) pour obtenir une représentation contextuelle de chaque tour de parole (que nous désignerons également par "énoncés" dans ce qui suit). Nous utilisons ici des réseaux siamois (Koch *et al.*, 2015) mais l'approche peut être adaptée à n'importe quel modèle de *metric learning*. Nous démontrons en outre que notre approche est plus performante que certains des derniers grands modèles de langage (*Large Language Models*, LLMs) tels que les versions allégées de Falcon (Penedo *et al.*, 2023) ou LLaMA 2 (Touvron *et al.*, 2023). En outre, notre méthode est efficace dans le sens où elle implique des modèles légers, adaptables et rapidement entraînables, qui donnent des scores état-de-l'art sur DailyDialog en macroF1 avec 57.71% et des résultats satisfaisants en microF1 avec 57.75%.

Dans les sections suivantes, nous passons d'abord en revue les travaux relatifs à l'ERC (Section 2). Nous présentons ensuite notre méthodologie (Section 3) et décrivons le dispositif expérimental que nous utilisons (Section 4). Nous évaluons ensuite nos modèles par rapport à une référence sans contexte conversationnel et aux modèles état-de-l'art pour l'ERC dans la section 5. Enfin, nous exposons nos principales conclusions ainsi que des perspectives pour les travaux futurs dans la section 6. Nous mettrons à disposition notre code et nos modèles sur *GitHub* et *HuggingFace models*.

2 État de l'art

ERC. Bien que la plupart des travaux en ERC tirent parti de la multimodalité (Song *et al.*, 2022; Li *et al.*, 2022; Hu *et al.*, 2022), certains modèles ont été développés pour l'ERC sur des conversations textuelles uniquement, que ce soit des données multimodales limités au texte tels que IEMOCAP (Busso *et al.*, 2008) ou MELD (Poria *et al.*, 2019), ou sur des données uniquement textuelles comme dans DailyDialog (Li *et al.*, 2017). L'apprentissage profond permet des progrès significatifs en ERC sur le texte, en commençant par l'utilisation de réseaux récurrents (RNN) (Rumelhart *et al.*, 1985; Jordan, 1986) par Poria *et al.* (2017). D'autres travaux utilisant des structures récurrentes ont suivi, comme DialogueRNN (Majumder *et al.*, 2019; Ghosal *et al.*, 2020). Ce modèle tire parti du mécanisme d'attention (Bahdanau *et al.*, 2014) implémenté dans le Transformer (Vaswani *et al.*, 2017). Les méthodes basées sur les graphes sont également efficaces comme le montre (Ghosal *et al.*, 2019), non seulement en tant que telles, mais aussi en incluant des connaissances externes (Lee & Choi, 2021).

Les travaux existants en ERC s'appuient principalement sur l'évaluation de leur modèle en microF1 en excluant l'étiquette neutre (pas d'émotion), souvent majoritaire. Cependant, des travaux récents se passent de cette évaluation pour se concentrer uniquement sur la macroF1 (Pereira *et al.*, 2023), tandis que d'autres ont considéré le coefficient de corrélation de Matthews comme une métrique adaptée à cette tâche (Guibon *et al.*, 2021). Dans ce travail, nous nous concentrons sur DailyDialog, qui consiste en des conversations générées artificiellement par l'Homme sur les préoccupations de la vie quotidienne, avec un étiquetage des émotions au niveau du tour de parole. Liang *et al.* (2022) proposent un modèle basé sur un réseau neuronal de graphes (*Graph Neural Network*, GNN) et un champ aléatoire conditionnel (*Conditional Random Field*, CRF) qui atteint 64,01% en microF1.

Bien qu'il soit connu pour ne pas fournir les meilleures performances par rapport aux approches d'apprentissage frugal (Dumoulin *et al.*, 2021), le *metric learning* permet une meilleure généralisation grâce à un entraînement plus robuste (Finn *et al.*, 2017; Antoniou *et al.*, 2019), ce qui est particulièrement adapté à la détection d'émotions humaines complexes et variées (Plutchik, 2001).

Metric learning. Hospedales *et al.* (2022) expliquent que le méta-apprentissage consiste en un *méta-optimiseur* qui décrit les mises à jour du méta-apprenant, une *méta-représentation* qui stocke les connaissances acquises et un *méta-objectif* orienté vers la tâche souhaitée. Cette configuration basée sur l'optimisation fournit des algorithmes complets souvent basés sur des scénarios épisodiques (Ravi & Larochelle, 2016; Finn *et al.*, 2017; Mishra *et al.*, 2017) qui reflètent l'idée d'"apprendre à apprendre". Mais cela implique des calculs de gradient au second ordre, ce qui est coûteux. Des solutions palliatives comme la différenciation implicite (Lorraine *et al.*, 2020) impliquent toujours un compromis entre performance et coût mémoire (Hospedales *et al.*, 2022). C'est pourquoi des variantes sont apparues, telles que le *metric learning*, dont le méta-objectif est l'apprentissage de la méta-représentation elle-même. Par exemple, les réseaux siamois (Koch *et al.*, 2015) tirent parti du partage des paramètres entre des sous-réseaux identiques pour apprendre une distance entre les

données. Les réseaux de relations (*Relation Networks*, Sung *et al.* (2018)) considèrent également une métrique de distance, s'écartant de l'approche euclidienne. Les réseaux de correspondance (*Matching Networks*, Vinyals *et al.* (2016)) exploitent des exemples d'apprentissage pour identifier les plus proches voisins pondérés. Les réseaux prototypiques (*Prototypical Networks*, Snell *et al.* (2017)) calculent les représentations moyennes des classes et les comparent avec la similarité cosinus. Son adaptation pour l'ERC en apprentissage frugal a commencé à partir de Guibon *et al.* (2021).

Dans ce travail, nous utilisons un réseau siamois. Ce modèle à l'architecture simple est facilement contrôlable et évolutif, mais on pourrait tout à fait adapter l'approche à des structures plus complexes. Les réseaux siamois ont été utilisés en TAL pour la détection d'intentions dans le texte (Ren & Xue, 2020), en vision par ordinateur pour la reconnaissance faciale (Hayale *et al.*, 2023) et dans l'apprentissage de représentations complexes (Jin *et al.*, 2021).

3 Méthodologie

Nous utilisons le *metric learning* pour l'apprentissage relatif des émotions, ce qui permet d'extraire des méta-informations des données. Notre réseau siamois comporte trois sous-réseaux identiques, dont les sorties sont comparées à l'aide de la fonction de coût par triplet (Schultz & Joachims, 2003), dénommée ci-après *triplet loss*. Initialement appliquée aux problèmes de vision par ordinateur (Chechik *et al.*, 2010; Schroff *et al.*, 2015), la *triplet loss* est définie sur un triplet d'échantillons de données (a, p, n) de sorte que si a et p appartiennent à la même classe et n à une classe différente, alors :

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

où le paramètre `marge` est un nombre strictement positif.

En considérant la *triplet loss*, plusieurs stratégies s'offrent à nous : récupérer les triplets les plus difficiles, lorsque le positif est loin de l'ancre, tandis que l'ancre est proche du négatif ; ou encore ignorer les triplets les plus faciles, c'est-à-dire lorsque le positif est plus proche de l'ancre. Compte tenu de la taille limitée de nos données, nous abordons la stratégie globale en considérant chaque triplet. Bien que la *triplet loss* puisse être utilisée dans plusieurs stratégies, nous n'abordons la stratégie globale qu'en considérant chaque triplet dans nos données, en raison de leur taille limitée.

Représentations isolées. L'objectif de nos expériences étant de caractériser la contribution du contexte conversationnel à la prédiction des émotions dans le cadre d'un apprentissage contrastif, nous avons développé, dans un premier temps, un modèle sur des énoncés isolés. Il s'agit formellement de prédire l'émotion des énoncés indépendamment de leur contexte. Pour ce faire, nous considérons d'abord une projection pour chaque mot de l'énoncé vers sa représentation FastText associée (Bojanowski *et al.*, 2017). À partir de ces plongements lexicaux, les triplets (a, p, n) susmentionnés sont échantillonnés aléatoirement et donnés en entrée au réseau siamois, dont le sous-réseau s'améliore dans la prédiction des émotions au fur et à mesure de la rétro-propagation de la *triplet loss*.

Représentations contextuelles. Dans le cas contextuel, nous construisons des représentations d'énoncés contextuels à partir d'un encodage de type BERT (Devlin *et al.*, 2019). Les plongements lexicaux utilisés sont à l'échelle de la phrase et non du mot car ils fournissent des représentations d'énoncés plus légères. Une fois que le dialogue est représenté avec la série de plongements pré-entraînés qui lui est associée, ces sorties sont concaténées pour former une représentation du dialogue, et les informations contextuelles sont prises en compte en déployant l'attention sur elles. Concrètement, une couche

d'encodeur de Transformer est appliquée aux plongements gelés concaténés. Cette représentation contextuelle du dialogue est ensuite divisée au niveau des marqueurs `[SEP]` pour aboutir à des représentations contextuelles au niveau de l'énoncé, sur lesquelles on prédit l'émotion. Afin d'adapter les représentations contextuelles des énoncés à l'objectif de prédiction des émotions, nous ajoutons un classificateur d'émotions pré-entraîné sur les données d'entraînement de DailyDialog, qui participe également à la rétro-propagation. En parallèle, les représentations contextuelles sont optimisées selon l'objectif lié à la *triplet loss*. Ceci est illustré en figure 1.

Ce scénario permet l'apprentissage des émotions individuelles et relatives, de telle sorte que chaque phase d'apprentissage renforce l'autre. Grâce à ce cadre de méta-apprentissage, des méta-informations sur les émotions sont extraites, et nous pouvons nous attendre à ce que ce modèle soit capable de réaliser une classification pertinente sur de nouvelles étiquettes en apprentissage frugal.

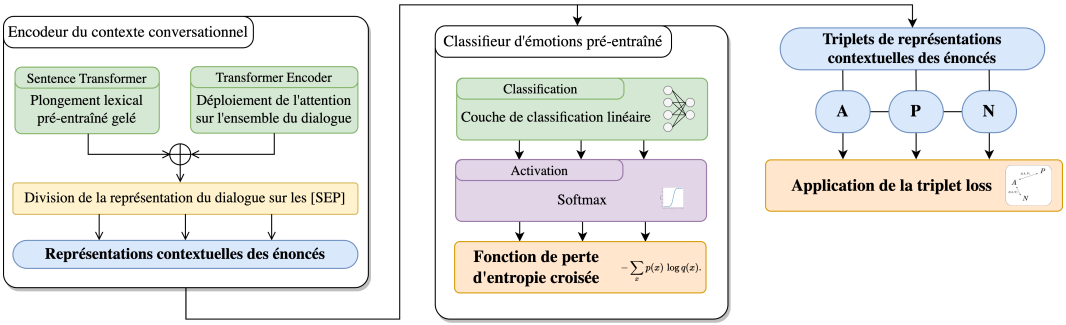

FIGURE 1 – Trois principales étapes de l'entraînement pour la prédiction d'émotions en contexte. Les fonctions de coût sont toutes deux rétro-propagées afin d'améliorer progressivement l'encodeur.

4 Protocole expérimental

Données. Toutes les expériences ont été menées sur DailyDialog (Li *et al.*, 2017) qui fournit 13 118 dialogues sur la vie quotidienne avec un étiquetage des émotions au niveau de l'énoncé. Ce jeu de données est relativement petit, ce qui permet de manipuler les entrées facilement et d'exécuter rapidement des tests. Il existe six étiquettes émotionnelles (colère, dégoût, peur, joie, tristesse et surprise) et une étiquette neutre. Pour la prédiction des émotions, l'évaluation est effectuée uniquement sur les étiquettes émotionnelles conformément aux travaux antérieurs (Ghosal *et al.*, 2021; Zhong *et al.*, 2019). Nous utilisons les sous-ensembles originaux (entraînement, validation et test) de (Li *et al.*, 2017). Les principales caractéristiques de DailyDialog sont données table 1.

Langue	Type	Max Msg/Conv	Moy Msg/Conv	Labels	Labels*	Nb. Conv
Anglais	Artificiel	35	8	7	6	13 118

TABLE 1 – Statistiques de DailyDialog (Li *et al.*, 2017). Labels* exclut l'étiquette neutre.

Spécificités du modèle. Pour le modèle avec énoncés isolés, nous considérons deux types de sous-réseaux : les couches linéaires simples et les couches récurrentes à mémoire court et long terme (*Long*

Short-Term Memory, LSTM, Hochreiter & Schmidhuber (1997)). Dans le cas contextuel, le sous-réseau est un encodeur de Transformer. Nous avons utilisé trois modèles de Transformers pré-entraînés au niveau de la phrase disponibles dans la bibliothèque Python `sentence transformers`[1] : MPNet (Song *et al.*, 2020), MiniLM (Wang *et al.*, 2020) et RoBERTa (Liu *et al.*, 2019).

Spécificités de l'entraînement. Que ce soit pour le modèle avec énoncés isolés ou pour le modèle contextuel, la prédiction de l'émotion est au niveau de l'énoncé. Les triplets sont donc toujours des triplets d'énoncés. Cela occasionne un problème d'équilibre des classes comme le montre la distribution des émotions de DailyDialog figure 3. Ainsi, le rééquilibrage des classes induit par l'échantillonnage des triplets selon une distribution uniforme n'atténue pas suffisamment les biais pendant l'apprentissage et empêche la fonction de coût de converger. Nous avons mis en œuvre un échantillonneur pondéré par l'inverse des fréquences des étiquettes afin de tenir compte de la rareté de certaines étiquettes telles que `fear` (peur) ou `disgust` (dégoût).

Évaluation quantitative. Nous considérons à la fois la performance et la pertinence de l'entraînement afin que bénéficier des capacités de généralisation du méta-apprentissage. Ainsi, nous avons choisi, en plus des mesures de performance habituelles, une métrique très exigeante : le coefficient de corrélation de Matthews (MCC) (Cramér, 1946). Il mesure la corrélation de Pearson (Pearson, 1895) entre classe prédite et classe réelle. Le MCC est défini dans (Matthews, 1975) comme suit :

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

Comparaison avec les LLMs. Afin de mettre en perspective les résultats de nos modèles isolés et contextuels, nous comparons nos modèles avec des LLMs état-de-l'art, à savoir LlaMA 2 (Touvron *et al.*, 2023) et Falcon (Penedo *et al.*, 2023). Les deux sont considérés dans leur version adaptée aux instructions (*instruction fine-tuning*) et évalués en génération sur un seul essai (*zero-shot*). Nous avons développé une requête (*cf.* fig. 2, *prompt* en anglais) demandant une prédiction sur le dernier énoncé de chaque dialogue du jeu de test de DailyDialog. Les requêtes sont conçues de sorte à ce que le modèle ne génère qu'une seule étiquette. La requête est la même pour chaque modèle du même type (LlaMA ou Falcon). Il est difficile de trouver une bonne requête pour Falcon car le modèle génère `happiness` (joie) sur l'ensemble des données, à l'exception d'un dialogue.

```
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) Requête utilisée pour Llama2

```
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) Requête utilisée pour Falcon

FIGURE 2 – Requêtes pour Llama2 et Falcon

1. https://www.sbert.net/

5 Résultats

Nom du modèle	macroF1*	microF1*	MCC
Modèles état-de-l'art en 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)	41.80	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	–	–
LLMs			
Llama2-7b (Touvron *et al.*, 2023)	09.70	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
Notre approche			
SentEmoContext	**57.71**	57.75	**0.49**

TABLE 2 – Résultats en ERC sur DailyDialog, en utilisant le jeu de test de l'article d'origine. DRNN réfère à DialogueRNN. L'astérisque (*) indique l'exclusion de l'étiquette neutre.

Travaux connexes. Le tableau 2 donne les résultats en ERC sur DailyDialog. On observe une lente progression depuis 2017 où Poria *et al.* (2017) propose d'évaluer en microF1 en excluant la classe neutre (majoritaire). Ce modèle est une première référence pour cette tâche, obtenant 50,24% en microF1. En revanche, le modèle état-de-l'art actuel atteint maintenant 64,07% en microF1 (Liang *et al.*, 2022), ce qui représente une amélioration d'environ 14 points en 6 ans. Comme le montre la table 2, la communauté a suivi ce schéma d'évaluation. Cependant, nous pensons qu'il est important de prendre également en compte le macroF1, à l'exclusion de la classe majoritaire, car il montre la performance globale sur toutes les émotions. Certains travaux l'ont proposé à partir de 2020 (Ghosal *et al.*, 2020), conduisant à un gain de 2,5 points en 3 ans. Cela renforce l'affirmation selon laquelle l'ERC est une tâche difficile.

Notre modèle. SentEmoContext obtient 57,75% en microF1, un résultat décent mais quelque peu modeste comparé aux travaux connexes. La table 2 donne la performance moyenne de notre modèle sur 10 exécutions. Notre modèle est état-de-l'art en macroF1 avec 57,71%, surpassant CD-ERC (Pereira *et al.*, 2023) de 6,48 points, ce qui est considérable étant donné qu'ils ne se sont concentrés que sur

cette métrique, et TODKAT (Zhu *et al.*, 2021) de 5,15 points. Nous évaluons également notre modèle en MCC multi-classe (Baldi *et al.*, 2000) pour assurer sa pertinence, sans qu'elle soit comparable aux travaux connexes car ils ne donnent pas cette métrique. Il fournit ici un bon indicateur de la qualité de la classification, en minimisant l'effet des données hautement déséquilibrées des conversations. Étant donné que le MCC varie de -1 à 1, et que 0 indique le caractère aléatoire, un MCC de 0,49 indique que notre approche est à la fois équilibrée et précise en termes de prédictions.

Notre modèle est très performant car nous n'avons besoin que de 20 minutes par époque et nous l'entraînons en utilisant seulement 5 époques. Ceci dénote des approches existantes qui utilisent plusieurs flux par locuteur (Pereira *et al.*, 2023), la modélisation graphique pour la représentation du contexte et des connaissances (Zhong *et al.*, 2019; Li *et al.*, 2021), ou d'autres représentations lourdes dans leur modèle (Liang *et al.*, 2022). Notre modèle est stable avec un écart-type de seulement 0,01 en moyenne pour chaque métrique, ce qui renforce la qualité de cette approche efficace.

Comparaison avec les classifieurs d'émotions au niveau de l'énoncé. La table 3 montre les résultats de la classification directe des émotions sur les énoncés. Pour cette tâche, nous n'avons pris en compte que les 6 étiquettes d'émotion, en excluant complètement l'étiquette neutre. Ainsi, nous voulons déterminer la différence entre notre approche et la prédiction d'émotions isolées. Cela sert également d'étude d'ablation pour notre modèle SentEmoContext puisque cette étape fait partie de son entraînement. En table 3, nous voyons que notre modèle exploite le contexte conversationnel et le *metric learning* pour augmenter toutes les métriques. Notamment, la différence en termes de macroF1 montre l'importance de la *triplet loss* dans notre modèle. En effet, les classifieurs d'émotions sont entraînés en utilisant des lots équilibrés sur la distribution du jeu de données d'entraînement et une fonction de perte d'entropie croisée pondérée. Les résultats montrent que cela n'est pas suffisant pour traiter des données extrêmement déséquilibrées telles que des conversations.

Nom du modèle	**macroF1**	**microF1**	**MCC**
Classifieur d'émotions pré-entrainé sur les tours de parole			
all-MiniLM-L6-v2	20.22	33.11	0.40
all-mpnet-base-v2	14.43	32.90	0.37
Notre approche			
SentEmoContext	**57.71**	**57.75**	**0.49**

TABLE 3 – Comparaison avec un classifieur d'émotions agissant au niveau du tour de parole.

LLMs. Les résultats des LLMs sur un seul essai sont donnés en table 2. Ceux-ci servent d'indication sur la performance de tels modèles (allégés) en ERC. Même si ces modèles génératifs ne sont pas conçus pour cette tâche spécifique, ils réussissent toujours à surpasser les classifieurs d'émotions d'énoncés de la table 3, ce qui peut être considéré comme une manifestation des capacités émergentes des LLMs – *emerging abilities* (Srivastava *et al.*, 2022).

Facteur du déséquilibre des classes. Alors que la table 1 montre les caractéristiques du jeu de données, elle omet la principale caractéristique des étiquettes d'émotions : un fort déséquilibre. En ERC, les principales difficultés sont la définition des étiquettes, du contexte mais aussi le déséquilibre qui limite l'apprentissage des émotions en contexte. La figure 3 montre la distribution des étiquettes dans DailyDialog, sans l'étiquette neutre. Étant donné que cette dernière est l'étiquette majoritaire et qu'elle est exclue des métriques d'évaluation par l'ensemble de la communauté ERC, le fait que même dans les étiquettes d'émotion les données soient aussi déséquilibrées s'avère être un défi et

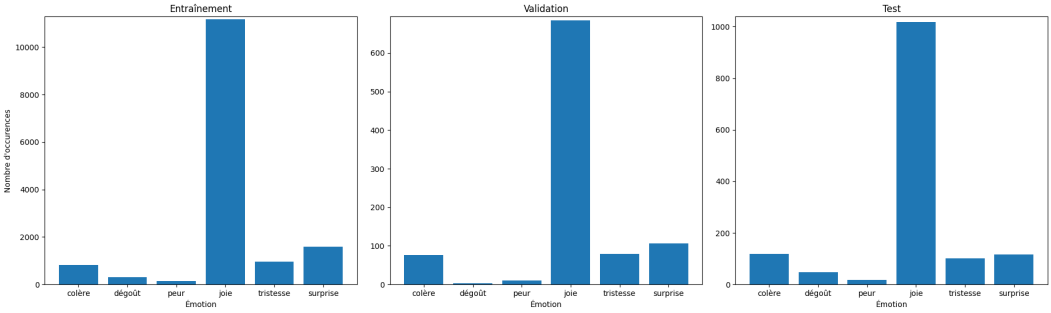

FIGURE 3 – Distributions des étiquettes émotionnelles dans les sous-ensembles de DailyDialog.

doit être abordé. Nous nous appuyons sur Guibon *et al.* (2023) pour traiter le déséquilibre en deux phases : tout d'abord, nous équilibrons les lots en fonction de la fréquence des étiquettes dans le jeu d'apprentissage. Ensuite, nous pondérons la fonction de perte d'entropie croisée du classificateur d'émotions en tenant compte du déséquilibre restant dans chaque lot. Ce travail traite également le déséquilibre en considérant des triplets, car nous éliminons alors le facteur de déséquilibre tout en utilisant les états cachés qui proviennent d'une représentation équilibrée. Nous pensons que cela explique en partie l'efficacité et l'efficience de notre modèle.

6 Discussion

Limitations des LLMs. La première limitation rencontrée avec les LLMs est la nécessité de GPUs à haute mémoire pour les tester. Ceci explique pourquoi en table 2 nous considérons seulement leurs versions légères. Alors que Llama2 7b et 13b ont donné des réponses dans un bon format, avec une seule étiquette, Falcon ne s'est pas comporté comme nous le souhaitions. Pour pallier ce problème, nous considérons la première émotion mentionnée dans la sortie. Il est également important de noter que nous n'avons pas voulu utiliser ChatGPT d'OpenAI car nous n'avons pas un contrôle clair sur la version du modèle, la taille et l'approche utilisée derrière l'API, mais aussi parce que nous voulions utiliser des systèmes open source pour pouvoir diffuser nos modèles à la communauté.

La fenêtre contextuelle constitue une autre limitation. En ERC, la taille du contexte est essentielle, mais avec les LLMs, l'ajout d'exemples dans la requête pour effectuer un apprentissage frugal prendrait beaucoup de place dans le contexte global, la requête faisant partie du contexte. Ceci explique notre décision de ne considérer que l'apprentissage en un seul essai pour les LLMs, même s'il conviendrait de considérer également l'apprentissage frugal sur cette tâche spécifique.

Taille et efficacité du modèle. Notre modèle est efficace. Il donne des résultats état-de-l'art en macroF1 et de bons résultats en microF1, alors qu'il s'entraîne relativement vite et ne nécessite pas beaucoup d'époques pour converger. Nous pensons que cette efficacité, ainsi que la mémoire limitée nécessaire à l'apprentissage, est due à la rétro-propagation en deux étapes et au fait que nous utilisons des représentations intégrées à l'énoncé avec des Transformers au niveau de la phrase. Ainsi, notre modèle peut traiter efficacement de longs contextes conversationnels avec un coût limité en mémoire. En outre, la table 4 montre la différence entre les modèles que nous avons utilisés, en termes de taille, de paramètres et de nombre de couches. Notre modèle est relativement petit si l'on considère les

	Transformers		LLMs			Notre approche
Modèle	MiniLM	MPNet	Llama2-7b	Llama2-13b	Falcon-7b	SentEmoContext
Tokens	1bn+	1bn+	2T	2T	1.5T	4M
Taille	80 MB	420 MB	13 GB	25 GB	15 GB	604,8 MB
Paramètres	22M	110M	7B	13B	7B	157M

TABLE 4 – Aperçu de la taille des modèles. Les modèles LLaMA sont deux versions de LLaMA 2. MiniLM et MPNet sont les mêmes que ceux présentés en table 3.

avancées récentes et les travaux connexes en ERC, mais aussi par rapport aux LLMs.

Représentation relative des étiquettes. Notre approche apprend deux fois à partir des données, d'abord en utilisant un cadre supervisé, puis en tenant compte des distances relatives entre les représentations, en mettant à jour par la *triplet loss*. Cela permet d'utiliser notre modèle pour différents jeux de données de conversations avec différentes étiquettes. La seule exigence pour étendre la portée de ce modèle serait de considérer une autre stratégie d'échantillonnage des triplets en ignorant les étiquettes, telle que la stratégie *batch-hard* (Do *et al.*, 2019).

7 Conclusion

Dans cet article, nous présentons notre modèle SentEmoContext issu d'une approche combinant la représentation au niveau de l'énoncé, le *metric learning* et les réseaux siamois à l'aide de la *triplet loss*. Ce modèle représente efficacement le contexte conversationnel, atteint un score état-de-l'art en macroF1 de 57.71%, et un microF1 satisfaisant de 57.75% en ERC sur DailyDialog. Nous proposons également d'utiliser le coefficient de corrélation de Matthew afin de mieux évaluer cette tâche.

Avec SentEmoContext , nous utilisons l'apprentissage contrastif avec un échantillonage pour limiter le déséquilibre des classes. Nous utilisons Sentence BERT pour minimiser la mémoire nécessaire tout en représentant l'ensemble du contexte conversationnel. Cela conduit à un apprentissage plus robuste et plus efficace qui ne nécessite pas beaucoup d'époques pour obtenir des résultats satisfaisants. Nous montrons également que les LLMs open source de taille modeste sont en retard en ERC, car cette tâche nécessite d'incorporer beaucoup de contexte dans la requête et n'est pas spécifiquement pertinente pour les modèles génératifs.

Dans nos travaux futurs, nous envisageons d'appliquer cette approche à des données conversationnelles proposant des étiquettes légèrement différentes, car notre modèle apprend les émotions de manière relative. Nous prévoyons donc de l'adapter à un cadre davantage lié au méta-apprentissage.

Remerciements

Les expériences présentées dans cet article ont été réalisées sur le banc d'essai Grid'5000, soutenu par un groupement d'intérêt scientifique hébergé par Inria et comprenant le CNRS, RENATER et plusieurs universités ainsi que d'autres organisations (voir https://www.grid5000.fr).

Références

ANTONIOU A., EDWARDS H. & STORKEY A. (2019). How to train your maml. Seventh International Conference on Learning Representations, ICLR 2019.

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015*, **1409**.

BALDI P., BRUNAK S., CHAUVIN Y., ANDERSEN C. & NIELSEN H. (2000). Assessing the accuracy of prediction algorithms for classification : An overview. *Bioinformatics (Oxford, England)*, **16**, 412–24.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

BUSSO C., BULUT M., LEE C.-C., KAZEMZADEH A., MOWER E., KIM S., CHANG J. N., LEE S. & NARAYANAN S. S. (2008). Iemocap : interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, **42**(4), 335–359.

CHECHIK G., SHARMA V., SHALIT U. & BENGIO S. (2010). Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.*, **11**, 1109–1135.

CRAMÉR H. (1946). *Mathematical Methods of Statistics (PMS-9), Volume 9*. Princeton : Princeton University Press. DOI : doi :10.1515/9781400883868.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (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)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DO T.-T., TRAN T., REID I., KUMAR V., HOANG T. & CARNEIRO G. (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*, p. 10404–10413.

DUMOULIN V., HOULSBY N., EVCI U., ZHAI X., GOROSHIN R., GELLY S. & LAROCHELLE H. (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)*.

EKMAN P., SORENSON E. R. & FRIESEN W. V. (1969). Pan-cultural elements in facial displays of emotion. DOI : https ://doi.org/10.1126/science.164.3875.86.

FINN C., ABBEEL P. & LEVINE S. (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, p. 1126–1135 : JMLR.org.

GHOSAL D., MAJUMDER N., GELBUKH A., MIHALCEA R. & PORIA S. (2020). COSMIC : COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 2470–2481, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.224.

GHOSAL D., MAJUMDER N., MIHALCEA R. & PORIA S. (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*, p. 1435–1449, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.124.

GHOSAL D., MAJUMDER N., PORIA S., CHHAYA N. & GELBUKH A. (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)*, p. 154–164, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1015.

GUIBON G., LABEAU M., FLAMEIN H., LEFEUVRE L. & CLAVEL C. (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.

GUIBON G., LABEAU M., LEFEUVRE L. & CLAVEL C. (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. DOI : 10.1609/aaai.v37i6.25940.

HAYALE W., NEGI P. S. & MAHOOR M. H. (2023). Deep siamese neural networks for facial expression recognition in the wild. *IEEE Transactions on Affective Computing*, **14**(2), 1148–1158. DOI : 10.1109/TAFFC.2021.3077248.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780. DOI : 10.1162/neco.1997.9.8.1735.

HOSPEDALES T., ANTONIOU A., MICAELLI P. & STORKEY A. (2022). Meta-learning in neural networks : A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**(9), 5149–5169.

HU G., LIN T.-E., ZHAO Y., LU G., WU Y. & LI Y. (2022). UniMSE : Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 7837–7851, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.534.

ISHIWATARI T., YASUDA Y., MIYAZAKI T. & GOTO J. (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)*, p. 7360–7370, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.597.

JIN M., ZHENG Y., LI Y.-F., GONG C., ZHOU C. & PAN S. (2021). Multi-scale contrastive siamese networks for self-supervised graph representation learning. In Z.-H. ZHOU, Éd., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, p. 1477–1483 : International Joint Conferences on Artificial Intelligence Organization. Main Track, DOI : 10.24963/ijcai.2021/204.

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

KHOSLA P., TETERWAK P., WANG C., SARNA A., TIAN Y., ISOLA P., MASCHINOT A., LIU C. & KRISHNAN D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, **33**, 18661–18673.

KOCH G., ZEMEL R. & SALAKHUTDINOV R. (2015). Siamese neural networks for one-shot image recognition.

LEE B. & CHOI Y. S. (2021). Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 443–455, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.36.

LI J., LIN Z., FU P. & WANG W. (2021). Past, present, and future : Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association*

for Computational Linguistics : EMNLP 2021, p. 1204–1214, Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-emnlp.104.

LI Y., SU H., SHEN X., LI W., CAO Z. & NIU S. (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)*, p. 986–995, Taipei, Taiwan : Asian Federation of Natural Language Processing.

LI Z., TANG F., ZHAO M. & ZHU Y. (2022). EmoCaps : Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics : ACL 2022*, p. 1610–1618, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.126.

LIANG C., XU J., LIN Y., YANG C. & WANG Y. (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)*, p. 148–157, Online only : Association for Computational Linguistics.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach.

LORRAINE J., VICOL P. & DUVENAUD D. (2020). Optimizing millions of hyperparameters by implicit differentiation. In S. CHIAPPA & R. CALANDRA, Éds., *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 de *Proceedings of Machine Learning Research*, p. 1540–1552 : PMLR.

MAJUMDER N., PORIA S., HAZARIKA D., MIHALCEA R., GELBUKH A. & CAMBRIA E. (2019). Dialoguernn : An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**(01), 6818–6825. DOI : 10.1609/aaai.v33i01.33016818.

MATTHEWS B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et biophysica acta*, **405 2**, 442–51.

MISHRA N., ROHANINEJAD M., CHEN X. & ABBEEL P. (2017). A simple neural attentive meta-learner. In *International Conference on Learning Representations*.

ONG D., SU J., CHEN B., LUU A. T., NARENDRANATH A., LI Y., SUN S., LIN Y. & WANG H. (2022). Is discourse role important for emotion recognition in conversation ? *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**(10), 11121–11129. DOI : 10.1609/aaai.v36i10.21361.

PEARSON K. (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.

PENEDO G., MALARTIC Q., HESSLOW D., COJOCARU R., CAPPELLI A., ALOBEIDLI H., PANNIER B., ALMAZROUEI E. & LAUNAY J. (2023). The refinedweb dataset for falcon llm : Outperforming curated corpora with web data, and web data only.

PEREIRA P., MONIZ H., DIAS I. & CARVALHO J. P. (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*, p. 228–236, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.wassa-1.21.

PLUTCHIK R. (2001). The Nature of Emotions. *American Scientist*, **89**(4), 344. DOI : 10.1511/2001.4.344.

PORIA S., CAMBRIA E., HAZARIKA D., MAJUMDER N., ZADEH A. & MORENCY L.-P. (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), p. 873–883, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1081.

PORIA S., HAZARIKA D., MAJUMDER N., NAIK G., CAMBRIA E. & MIHALCEA R. (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*, p. 527–536, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1050.

RAVI S. & LAROCHELLE H. (2016). Optimization as a model for few-shot learning. In *International Conference on Learning Representations*.

REN F. & XUE S. (2020). Intention detection based on siamese neural network with triplet loss. *IEEE Access*, **8**, 82242–82254. DOI : 10.1109/ACCESS.2020.2991484.

RUMELHART D. E., HINTON G. E. & WILLIAMS R. J. (1985). *Learning internal representations by error propagation*. Rapport interne, California Univ San Diego La Jolla Inst for Cognitive Science.

SCHROFF F., KALENICHENKO D. & PHILBIN J. (2015). Facenet : A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 815–823. DOI : 10.1109/CVPR.2015.7298682.

SCHULTZ M. & JOACHIMS T. (2003). Learning a distance metric from relative comparisons. In S. THRUN, L. SAUL & B. SCHÖLKOPF, Éds., *Advances in Neural Information Processing Systems*, volume 16 : MIT Press.

SHEN W., WU S., YANG Y. & QUAN X. (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)*, p. 1551–1560, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.123.

SNELL J., SWERSKY K. & ZEMEL R. (2017). Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 4080–4090, Red Hook, NY, USA : Curran Associates Inc.

SONG K., TAN X., QIN T., LU J. & LIU T.-Y. (2020). Mpnet : Masked and permuted pre-training for language understanding. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 16857–16867 : Curran Associates, Inc.

SONG X., HUANG L., XUE H. & HU S. (2022). Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 5197–5206, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.347.

SRIVASTAVA A. *et al.* (2022). Beyond the imitation game : Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv :2206.04615*.

SUNG F., YANG Y., ZHANG L., XIANG T., TORR P. H. & HOSPEDALES T. M. (2018). Learning to compare : Relation network for few-shot learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 1199–1208. DOI : 10.1109/CVPR.2018.00131.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S., BIKEL D., BLECHER L., FERRER C. C., CHEN M., CUCURULL G., ESIOBU D., FERNANDES J., FU J., FU W., FULLER B., GAO C., GOSWAMI V., GOYAL N., HARTSHORN A., HOSSEINI S., HOU R., INAN H., KARDAS M., KERKEZ V., KHABSA M., KLOUMANN I., KORENEV A., KOURA P. S., LACHAUX M.-A., LAVRIL T., LEE J., LISKOVICH D., LU Y., MAO Y., MARTINET X., MIHAYLOV T., MISHRA P., MOLYBOG I., NIE

Y., POULTON A., REIZENSTEIN J., RUNGTA R., SALADI K., SCHELTEN A., SILVA R., SMITH E. M., SUBRAMANIAN R., TAN X. E., TANG B., TAYLOR R., WILLIAMS A., KUAN J. X., XU P., YAN Z., ZAROV I., ZHANG Y., FAN A., KAMBADUR M., NARANG S., RODRIGUEZ A., STOJNIC R., EDUNOV S. & SCIALOM T. (2023). Llama 2 : Open foundation and fine-tuned chat models.

TU G., WEN J., LIU C., JIANG D. & CAMBRIA E. (2022). Context- and sentiment-aware networks for emotion recognition in conversation. *IEEE Transactions on Artificial Intelligence*, **3**(5), 699–708. DOI : 10.1109/TAI.2022.3149234.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need.

VINYALS O., BLUNDELL C., LILLICRAP T., KAVUKCUOGLU K. & WIERSTRA D. (2016). Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, p. 3637–3645, Red Hook, NY, USA : Curran Associates Inc.

WANG W., WEI F., DONG L., BAO H., YANG N. & ZHOU M. (2020). Minilm : Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 5776–5788 : Curran Associates, Inc.

ZHANG D., CHEN F. & CHEN X. (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)*, p. 7395–7408, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.408.

ZHAO W., ZHAO Y. & LU X. (2022). Cauain : Causal aware interaction network for emotion recognition in conversations. In L. D. RAEDT, Éd., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, p. 4524–4530 : International Joint Conferences on Artificial Intelligence Organization. Main Track, DOI : 10.24963/ijcai.2022/628.

ZHONG P., WANG D. & MIAO C. (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)*, p. 165–176, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1016.

ZHU L., PERGOLA G., GUI L., ZHOU D. & HE Y. (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)*, p. 1571–1582, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.125.