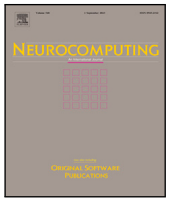
Neurocomputing 577 (2024) 127370

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Self-supervised utterance order prediction for emotion recognition in conversations

Dazhi Jiang [a,b], Hao Liu [b], Geng Tu [b], Runguo Wei [b], Erik Cambria [c,*]

[a] Guangdong Provincial Key Laboratory of Information Security Technology, Shantou, China
[b] Department of Computer Science, Shantou University, Shantou, China
[c] School of Computer Science and Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Communicated by S. Poria

Keywords:
ERC
Self-supervised learning
Utterance order prediction

ABSTRACT

As the order of the utterances in a conversation changes, the meaning of the utterance also changes, and sometimes, this will cause different semantics or emotions. However, the existing representation learning models do not pay close attention to capturing the internal semantic differences of utterance caused by the change of utterance order. Based on this, we build a self-supervised utterance order prediction approach to learn the logical order of utterance, which helps understand the deep semantic relationship between adjacent utterances. Specially, the utterance binary composed of two adjacent utterances, which are ordered or disordered, is fed to the self-supervised model so that the self-supervised model can obtain firm representation learning ability for the semantic differences of the adjacent sentences. The self-supervised method is applied to the downstream conversation emotion recognition task to test the value of the approach. The features extracted from the self-supervised model are fused with the multimodal features to obtain a richer utterance representation. After that, emotion recognition models are applied to two different datasets. The experiment results show that our proposed approach outperforms the current state of the art on ERC benchmark datasets.

1. Introduction

Affective computing has witnessed a sort of renaissance due to the recent developments in artificial intelligence [1]. Emotion Recognition in Conversations (ERC), in particular, is an affective computing task that is becoming increasing popular. It aims at recognizing the emotion of each utterance spoken in conversations, and the research can be used in various related applications, such as building effective dialogue systems [2,3], aiding social viewpoint mining [4,5], and building intelligent medical systems [6,7]. Current research on ERC focuses on the emotional information of the speaker in the current emotional state by analyzing contextual information and establishing different contexts for different speakers or using multimodal data to support this task.

Despite recent progress, two major issues still remain unaddressed: (1) How to ensure emotional consistency. (2) The creation of contextual information. The current research works are roughly divided into two categories: the first is to obtain the context representation of utterances based on a temporal neural network, and the second is to obtain long-distance information based on the graph networks. DialogueCRN [8] extracts and integrates emotion cues by building context reasoning networks.

ICON [9] and DialogueRNN [10] obtain utterance dependencies in different contexts by modeling different speakers with recurrent neural networks. To overcome the weakness of such networks in dealing with long information, they use the attention mechanism. On the other hand, DialogueGCN [11] obtains context information by constructing a directed graph, where the nodes denote utterance and the edges denote the relationship between utterances pair containing two types of relationships. RGAT [12] employs the relation-aware graph attention networks with embedding utterance relational position to take sequential information into account.

All these approaches, however, ignore an important fact: when the utterance order changes, the utterance meaning also changes, which may have different utterance emotions. Fig. 1 shows one such example. The change in utterance order leads to a change in the inner meaning of the utterance, resulting in a difference in speakers' emotions. In this paper, we investigate the impact of utterance orders on the ERC task and present a self-supervised model to predicting the utterance order. The model can capture the potential semantic information for the adjacent utterances by predicting whether the adjacent utterance is ordered.

* Corresponding author.
E-mail addresses: dzjiang@stu.edu.cn (D. Jiang), 20hliu2@stu.edu.cn (H. Liu), 19gtu@stu.edu.cn (G. Tu), 20rgwei@stu.edu.cn (R. Wei), cambria@ntu.edu.sg (E. Cambria).

https://doi.org/10.1016/j.neucom.2024.127370
Received 27 August 2022; Received in revised form 16 December 2023; Accepted 3 February 2024
Available online 6 February 2024

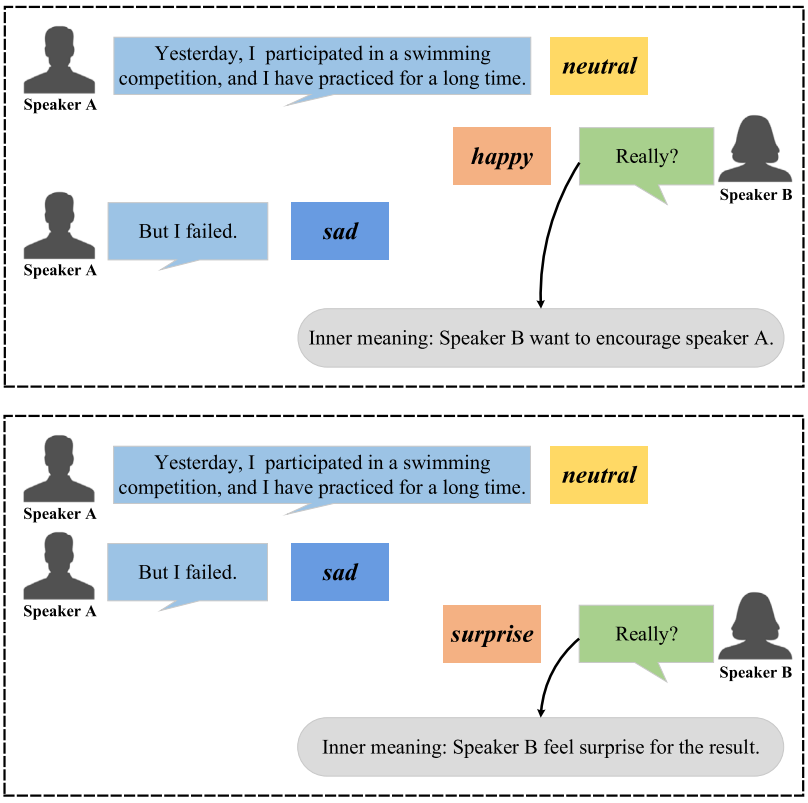


Fig. 1. The utterance order of speaker A and speaker B change. This change leads to dramatic differences in meaning of utterance spoken by speaker B which result in different emotion of this utterance.

When there is an emotion-shift in adjacent utterances, the potential semantic information of the utterances has a more significant impact on emotions. Two adjacent utterances are sampled to form an utterance pair used in the pre-trained model. Specifically, we first process word embedding of utterance pairs using the BERT-large model, after which the utterance pair features are fed into the transformer for further feature extraction. Then, the model predicts utterance pairs' order. After that, the model is applied to the downstream emotion recognition models. In this paper, we use DialogueRNN [10], BiERU [13], HiTrans [14], DialogueCRN [8], DAG [15]. The intermediate representations of the pre-trained model are obtained to enrich each model's local features of utterance. In the experimental stage, extensive experiments are conducted on emotion datasets IEMOCAP [16], and MELD [17].

To validate the proposed approach, the self-supervised model is applied to multiple emotion recognition models to improve their ability to recognize emotion, and the results are highly competitive compared to the latest results. The main contributions of this paper are as follows:

(1) We propose a self-supervised utterance order prediction method for conversation. It is the first time that a self-supervised approach has been used for ERC.
(2) The model can enhance the semantic understanding of utterances in the conversation context by predicting whether the adjacent utterances are orderly. Thus, the features extracted by the self-supervised model are more effective on the downstream ERC task.
(3) The experiments are conducted on several emotion recognition models on IEMOCAP and MELD datasets. The results, which demonstrate the effectiveness of our proposed method on IEMOCAP, outperform recent state-of-the-art methods.

The remainder of the paper is organized as follows: Section 2 lists related works; Section 3 describes the proposed approach; Section 4 presents experiments and discusses results; finally, Section 5 offers concluding remarks.

2. Related work

Affective analysis can be used in dialogue systems [18], public opinion analysis [19,20], ElectroEncephaloGraphy [21,22], emotion cause entailment [23,24], finance [25], and more. Among them, ERC has been receiving much attention in the field of sentiment analysis about its potential applications in conversational systems [26,27]. At the same time, more and more datasets are proposed for conversation emotion recognition research, such as IEMOCAP [16], MELD [17], DailyDialog [28], EmoryNLP [29]. These benchmark conversational emotion datasets are mainly used for multimodal tasks. And none of them can be applied to emotion reasoning because of the lack of the necessary annotations for the detailed information required for reasoning tasks. None of them contain aspect-level and topic-level emotional annotations.

In recent years, the mainstream approaches for conversational emotion recognition are based on temporal networks and graphical convolutional network models. Due to the context propagation problem of conversations in recurrent neural networks, Deepanway Ghosal proposed DialogueGCN [30] by adding a graph structure to convolutional neural networks. However, the graph neural network represented by DialogueGCN collects information within a specific window, which ignores sequence information in a large span. In contrast, DAG [15] solves this problem by treating the utterances in a conversation as a directed acyclic graph. I-GCN [31] utilizes incremental graph structure to imitate the dynamical conversation. SKIER [32] construct a neurosymbolic framework with fusing symbolic dependency knowledge, concept-level commonsense, and sentiment knowledge.

In addition to the traditional construction of neural networks, a commonsense knowledge graph can also mine emotional commonsense in a conversation [33,34]. Nie [35] embeds commonsense and potential topic information into long dialogue emotion detection. Using contextual information, the influence of the speaker's state can be considered.

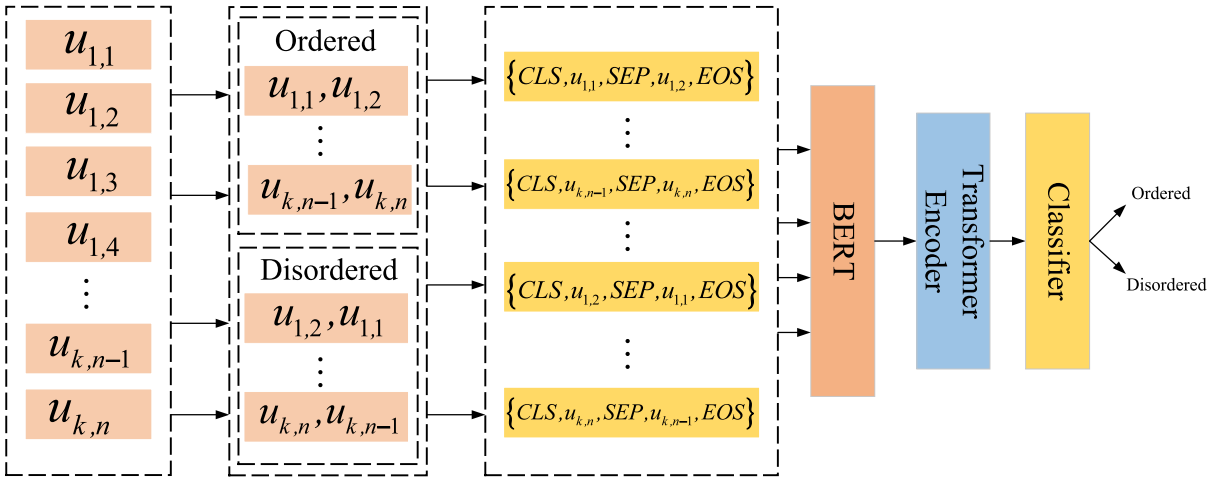


Fig. 2. The overall architecture of the self-supervised model where $u_{i,j}$ and $u_{i,j+1}$ two adjacent utterances from the same conversation i, where $i \in [1,k]$, $j \in [1,n]$. This model predicts whether two adjacent utterances are orderly or unorderly.

CMN [36], and ICON [9] model different speakers separately through GRU and obtain the interaction information between different speakers. Besides, multi-task [37–39] methods, and fuzzy networks [40] are also very effective for some multimodal emotion recognition studies. Self-supervised learning can obtain information on unlabeled data through pretext tasks. Self-supervised learning is now heavily explored in many fields of research using deep learning networks. In computer vision, pervasive images and visual features are learned from unlabeled data through various self-supervised tasks. For example, Gidaris et al. [41] propose to predict the rotated picture's rotation angle; to accomplish this, the model should learn some intrinsic knowledge about the picture. Image coloring of old photographs [42] also allowed the network to learn information about the picture, such as relevant color knowledge. Ishan et al. [43] can learn a visual feature by determining whether a video sequence is correct temporal order. To obtain the representation of temporal commonsense, Rowan et al. [44] construct a model to learn multimodal knowledge by watching a video with speech.

Self-supervised learning is also primarily used for emotion classification. Yu et al. [45] improved the emotion classification result by generating unimodal data labels from multimodal data through a self-supervised learning strategy. As for natural language processing, predicting central words or adjacent sequences can be used as a self-supervised task for learning word embeddings; predicting the masking words given a sequence of words can be used as an auxiliary task for language modeling. Devlin [46] proposed the masked language and next sentence prediction task to learn sentence embeddings. In building dialogue systems, Wu [47] proposed the task of utterance order consistency detection, which is acted as a supervised signal for dialogue generation. In BART [48], several self-supervised tasks such as sentence permutation and document rotation. In contrast, we consider whether the two consecutive utterances of the dialogue are ordered and propose a self-supervised task of utterance order prediction.

3. Method

In this section, we give a detailed description of the self-supervised utterance order prediction approach. This paper aims to predict the order of adjacent utterances so that we can pay more attention to modeling inter-utterance consistency without considering the topic and speaker information. Meanwhile, we have implemented several experiments for the downstream ERC tasks.

3.1. Self-supervise utterance order predict model

For our inter-sentence order prediction, which are two examples of positive and negative instance: $(u_{i,j}, u_{i,j+1})$ is the positive instance, and $(u_{i,j+1}, u_{i,j})$ is the negative instance where $u_{i,j}$ and $u_{i,j+1}$ are two adjacent utterances from the same conversation i, and $i \in [1,k]$, $j \in [1,n]$. The pre-trained model can further improve the performance of downstream sentiment analysis tasks.

The ERC task aims to recognize the hidden emotion of every utterance in a conversation, which requires utilizing contextual information and the speaker's information to enhance the model's performance in the conversation. Therefore, the pre-trained model obtained by the utterance order prediction approach can utilize the utterance context information to gain more information at the utterance level. When the model simultaneously utilizes the feature extracted from the pre-trained model and f_{text}, the model can focus on more details in the sentence.

Firstly, the pretext is defined as follows: given a set with N utterance pairs $\{(u_{1,1}, u_{1,2}), \ldots, (u_{k,n-1}, u_{k,n}), (u_{1,2}, u_{1,1}) \ , \cdots, (u_{k,n-1}, u_{k,n})\}$, the goal is to predict each utterance pair as positive or negative. The architecture of our pre-trained model is revealed in Fig. 2. The model is composed of three section: an encoder layer, a feature extract layer, and a prediction layer. The encoder layer transforms the utterances into a representation, while the feature extract layer further combines the utterance representation information and finally predicts the order of utterances through the predicted layer. The specific process is exhibited in Fig. 2.

The bert-large model is used as the encoder layer. Specifically, insert the tokens CLS, SEP, EOS into the utterance pair $(u_{i,j}, u_{i,j+1})$ to transform the utterance sequence into $\{CLS, u_{i,j}, SEP, u_{i,j+1}, EOS\}$, then, this sequence is then fed into the bert-large model and the output vector is the representation of the utterance pair and used in the next layer of the network, where the dimensionality of this vector is 1024. In the feature extract layer, we obtain the semantic information of utterance with a transformer. Transformer captures prior knowledge about utterance word order through discourse modeling, which allows the pre-trained model to learn more fine-grained utterance-level semantic information.

The representation of utterance obtained from the encoder layer is the input for the transformer. In the encoder layer of the transformer, we also add the positional encoding embedding to model the relative position of the utterance in the conversations. We use eight transformers encoder layers. The number of attention heads is 2, the dimension of the hidden layer is 2048, and the dropout rate is 0.2. We only use the encoder of the transformer because, in the emotion analysis task, the transformer is often employed for encoding utterances.

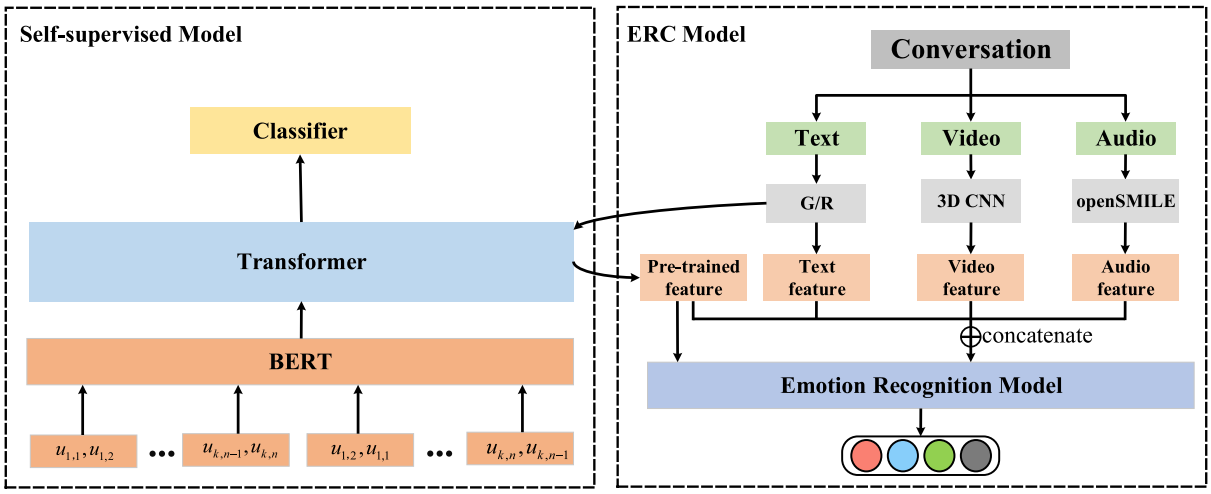


Fig. 3. ERC model architecture with the self-supervised task, where the utterance's emotion is the final output. The pre-trained model is used after completing the proposed self-supervised training. The representation of the utterance obtained from the last layer of the transformer fused with the multimodal features and fed into the downstream emotion recognition model. G/R denotes text features extracted using GloVe and RoBERTa pre-trained models.

The final representation of the utterance pair is obtained after the feature extract layer. And this representation is fed into the prediction layer with the MLP layer to produce the probability of whether the utterance pair is ordered or not:

$$p^{r_{u_{i,j}}} = softmax\left(linear\left(r_{u_{i,j}}\right)\right) \tag{1}$$

where the $r_{u_{i,j}}$ denotes the representation of $(u_{i,j}, u_{i,j+1})$ extracted by the Transformer. The $p^{r_{u_{i,j}}}$ is the probability of classification, $p^{r_{u_{i,j}}} \in [0, 1]$.

In this section, the self-supervised model is trained on the IEMOCAP and MELD datasets for emotion recognition.

3.2. ERC framework

In this section, we will specify how to apply our pre-trained model to downstream emotion recognition tasks. To demonstrate the effectiveness of the self-supervised utterance order prediction approach, we combine it with multiple emotion recognition models rather than with a particular model. Fig. 3 shows the combination of our proposed approach with the downstream emotion recognition task.

Firstly, we extract the three modal feature. **Textual Features Extraction:** GloVe [49], and RoBERTa [50] are used to extract text features. For RoBERTa, the output of the last layer is treated as the text feature. **Visual Features Extraction:** To get visual features,we use 3D-CNN to model spatiotemporal features of each utterance video. 3D-CNN helps to understand emotional concepts such as smiling or frowning, often spread across multiple video frames with no predefined spatial location. The model contains three blocks, each with two convolutional layers followed by max-pooling. We capture the visual features for each utterance with dimension. The is used as the representation of video features. **Audio Features Extraction:** We first formatted the audio of each voice video as a 16-bit PCM WAV file and used the open-source software openSMILE. Specifically, we use the IS13 ComParE1extractor which provides 6373 features for each utterance. Then the features are normalized, followed by L2-based feature selection, regarded as a presentation with dimension.

For multimodal data, the representation extracted from the pre-trained module is fused with the original multimodal representation to obtain the new feature representation for the downstream emotion recognition model.

$$u = concatanete\left(f_t, f_{text}, f_{video}, f_{audio}\right) \tag{2}$$

$$f_{emo} = ERCmodel\,(u) \tag{3}$$

where the f_t is the textual feature extracted by the pre-trained model. The $ERCmodel$ denotes the downstream ERC model. The representation of utterance u is the input of the ERC model with removing the classification layer. The representation f_{emo} is obtained through the ERC model that can be used for classification. To further improve the classification result of the final representation of the utterance, the features f_{emo} and f_t extracted by the pre-trained model are further fused as the final representation of the utterance:

$$u_o = r_1 * f_t + r_2 * f_{emo} \tag{4}$$

where r_1 and r_2 are the trainable parameters. u_o is the final utterance representation. Finally, the u_o is fed into the classification layer for emotion category prediction, and the probabilities of all candidate emotion labels are calculated:

$$\hat{y} = sofrmax\left(Wu_o + b\right) \tag{5}$$

where the W and b are trainable parameters. The model is trained by optimally minimizing the cross entropy loss of the ERC, for the single session, our objective function is:

$$loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right) \tag{6}$$

where i denotes the ith utterance, N is the number of utterances in this conversation, and y_i and $\hat{y}_i$ denote the expected emotion label and the predicted probability distribution for the emotion class of the ith utterance.

4. Experiment

4.1. Datasets

Our experiments perform emotion recognition experiments on the following two datasets.

IEMOCAP [16]: The IEMOCAP dataset consists of text, audio, and video. It contains two types of scenes, improvised and scripted scenes, and is divided into five sessions. All the corpus in the dataset has six emotions: anger, happiness, sadness, neutral, excitement, and frustration.

MELD [17]: MELD is a multimodal dataset. A total of 7 emotions and 3 sentiments are included. The labels we use in this paper are emotion labels. The emotion labels are anger, disgust, neutral, happiness, sadness, surprise, and fear.

The distribution of the number of training sets and test sets for two of these datasets is shown in Table 1.

Table 1
The dataset distribution.

Datasets	Conversations			Utterances		
	Train	Val	Test	Train	Val	Test
IEMOCAP	120	12	31	5810		1623
MELD	1039	114	280	9989	1109	5610

Datasets	Classes	Evaluation
IEMOCAP	6	Accuracy and Weighted F1
MELD	7	Accuracy and Weighted F1

4.2. Baselines and evaluations

- DialogueRNN [10]: It dynamically analyzes the state of each speaker during a conversation and models the context of each person's discourse for emotion analysis.
- BiERU [13]: A framework for conversational emotion analysis called bidirectional sentiment recursive unit is proposed.
- HiTrans [14]: A transformer-based network for modeling the context and speaker sensitivity. BERT is utilized as a low-level transformer to generate local corpus representations.
- DialogueCRN [8]: It is designed to iteratively execute multiple rounds of reasoning modules to extract affective cues. And the conversational context is fully understood in terms of human emotion.
- DAG-ERC [15]: A directed acyclic neural network is designed for the downstream emotion recognition task in conversation.

For the IEMOCAP and MELD datasets, like most current research works, we leverage average accuracy (ACC) and weighted F1 (F1) to evaluate our experimental results.

4.3. Result analysis

Table 2 unveils the experimental results that each model uses the multimodal feature where the textual feature is extracted by GloVe or RoBERTa respectively. Among them, the self represents the self-supervised utterance order prediction approach.

IEMOCAP: In Table 2, when using GloVe extracts the textual feature, the best results come from the self+BiERU model with ACC of 65.62% and F1 score of 64.56%. Meanwhile, we also use RoBERTa to re-extract the text features. The experimental results show that using the features extracted by RoBERTa brings about a considerable improvement. The results in Table 2 indicate that the best result comes from self+DAG-ERC with ACC of 69.01% and F1 score of 69.01%, which are state-of-the-art performance. The average conversation length is 50 on the IEMOCAP dataset. When the conversation is longer, more information can be obtained by predicting the utterance order task.

MELD: As shown in Table 2, the best result comes from the self + DialogueCRN model with ACC of 66.82% and F1 of 65.72%, while utilizing RoBERTa extracts the textual features. Compared to the DialogueCRN model, it improved by 1.25% on F1. Many conversations on the MELD dataset have more than 5 participants, making the speaker's information more variable. Due to this reason, approaches to self-supervised utterance prediction perform less well on the MELD dataset than on the IEMOCAP dataset.

Additionally, When we used RoBERTA instead of the BERT module in the pre-training stage as the feature extraction module, the results are shown in Table 3, which is similar to the results in Table 3.

4.4. Ablation study

To better understand the role of utterance order, we conduct several ablation studies on IEMOCAP and MELD datasets when the textual feature is extracted by RoBERTa. When the Transformer inserted into the model is not loaded with pre-trained weights, there is no information related to utterance order in the utterance representations extracted by the Transformer. Take BiERU as an example, the results are shown in Table 4. When combining Transformer, which does not perform pre-training tasks, with BiERU, the results are not significantly different from those of the original BiERU on the IEMOCAP and MELD datasets. This indicates that the transformer encoder, without being pre-trained, does not affect the downstream sentiment analysis model. Furthermore, when combining the pre-trained Transformer with BiERU, ACC and F1 improved by 0.86% and 1.55% on the IEMOCAP dataset, respectively. On the MELD dataset, ACC and F1 improved by 0.46% and 0.59%. These results illustrate the effectiveness of self-supervised utterance order prediction. These results suggest that the pre-trained model contains prior knowledge regarding utterance meanings influenced by shifts in discourse order, which empowers the emotion recognition model to pay more attention to the potential semantic information of the sentence when capturing context. In the presence of an emotion shift in neighboring utterances, it is crucial to focus on the subtle changes in utterance semantics.

4.5. Significance test

The significance test results of each model and each dataset are displayed in Table 4 (t-test). The results demonstrate a significant difference between our method and the comparison models (see Table 5).

4.6. Case study

A conversation sampled from the IEMOCAP dataset is shown in Fig. 4. Fig. 4 consists of two parts: The left part contains two utterances that are ordered and disordered. The right part is a hot map of the utterance two attention score. The aim is to predict the emotion label of the utterance 2 spoken by speaker B. Due to the lack of the ability to understand the change of intrinsic meaning caused by different utterance orders, some comparative models can easily to mistakenly identify the emotional label as frustrated or neutral. Fig. 4 shows that when the model does not use the features extracted by the self-supervise module, the model focuses on the words "problem" and "going" for utterance 2, which makes the model insufficient to capture the detailed information of the sentence.

When the order of utterances 1–2 changes, the intrinsic meaning will change. Fig. 4 shows that the model has higher attention scores for the words "problem", "satellites", and "down" in utterance 2. The ERC models can get more Fine-grained information-related emotion when the multimodal features are fused with the feature extracted from our proposed self-supervised model. Speaker A's emotion is identified as the angry label.

4.7. Error analysis

By ranking the weighted F1 score of each emotional category on MELD, and observing the proportion of the emotional sample with the lowest F1 in the dataset. We found that the F1 of fear is 0.2521, accounting for 0.9% of the samples, which shows that our approach has a low recognition ability for small sample emotions. In addition, misclassifications often occur on similar emotions, such as happy and excited or frustrated and angry. For example, on the IEMOCAP dataset, excited utterances are incorrectly predicted as Happy in 55.08% of the samples with incorrect predictions, while Happy utterances are incorrectly predicted as excited in 60% of the samples that were incorrectly predicted.

Table 2
Comparative performance of models with the textual feature extracted by GloVe(G) and RoBERTa(R). The ACC, F1 is average accuracy and weighted F1, respectively.

Model	$IEMOCAP_G$		$IEMOCAP_R$		$MELD_G$		$MELD_R$	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
DialogueRNN	63.40	62.75	64.20	64.21	59.66	57.35	65.90	63.90
self+DialogueRNN	**64.33**	**63.98**	**65.37**	**65.38**	**61.07**	**58.51**	**66.25**	**64.51**
BiERU	65.25	64.20	64.70	64.46	58.85	55.21	64.37	63.31
self+BiERU	**65.62**	**64.56**	**65.68**	**65.74**	**59.73**	**56.14**	**66.36**	**64.64**
HiTrans	-	-	64.03	64.07	-	-	62.26	61.89
self+Hitrans	-	-	**64.96**	**64.79**	-	-	**64.75**	**63.89**
DialogueCRN	63.83	63.74	66.17	66.21	59.92	57.76	65.44	64.47
self+DialogueCRN	**64.02**	**64.05**	**67.90**	**68.09**	**61.23**	**57.90**	**66.82**	**65.72**
DAG-ERC	-	-	68.00	67.83	-	-	63.87	63.48
self+DAG-ERC	-	-	**69.01**	**69.01**	-	-	**63.98**	**63.57**

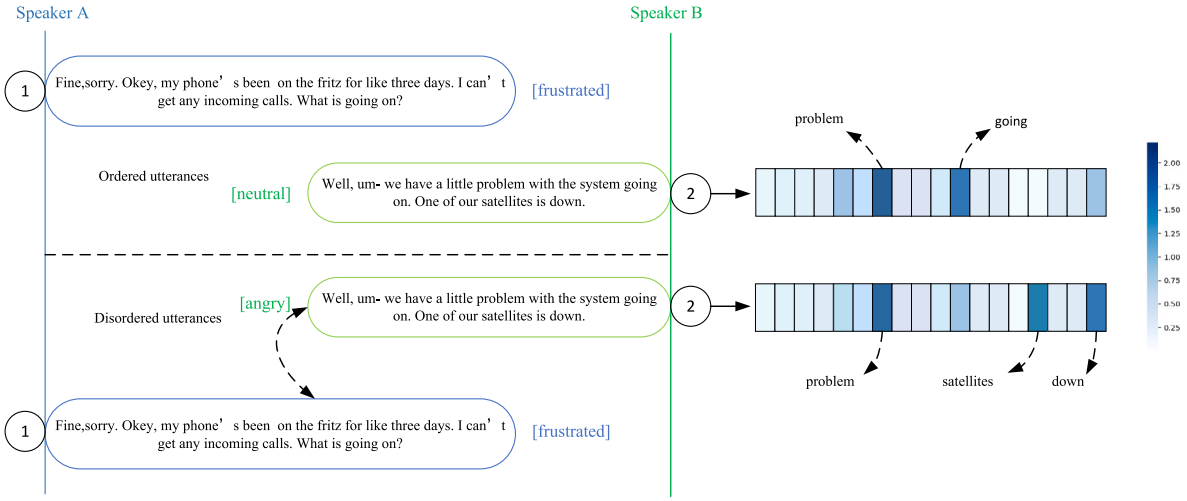


Fig. 4. A conversation passage from IEMOCAP dataset for case study.

Table 3
The self-supervised task and ERC task adopts the RoBERTa as backbone.

Model	$IEMOCAP_R$		$MELD_R$	
	ACC	F1	ACC	F1
self+DialogueRNN	65.13	65.42	66.78	64.73
self+BiERU	65.13	64.95	64.82	65.10
self+Hitrans	65.25	65.56	64.21	64.26
self+DialogueCRN	68.21	68.18	66.97	65.70
self+DAG-ERC	68.70	68.76	63.75	63.47

Table 4
Experimental results of ablation studies on IEMOCAP and MELD datasets.

Model	IEMOCAP		MELD	
	ACC	F1	ACC	F1
BiERU	64.70	64.46	64.37	63.31
Self+BiERU W/O Pre-train	64.82	64.19	65.90	64.05
Self+BiERU W/ Pre-train	65.68	65.74	66.36	64.64

Table 5
The significance test result (p-Value).

	DialogueRNN	BiERU	HiTrans
IEMOCAP	4.10e−04	3.67e−06	2.63e−04
MELD	7.29e−04	2.49e−05	8.10e−04

	DialogueCRN	DAG-ERC
IEMOCAP	1.19e−04	4.97e−03
MELD	2.36e−05	8.94e−03

Similarly, the neutral utterances are predicted as frustrated in 50% of the incorrectly predicted samples, and frustrated utterances are predicted as neutral in 42.61% of the incorrectly predicted samples. And there is a similar situation in the MELD dataset. This may be the utterances with similar emotion labels having similar semantic information.

5. Conclusion

In this paper, an approach that uses a self-supervised pretext task for conversational emotion recognition is proposed. We present the self-supervised task for predicting utterance order and its implementation in detail. In dialogue emotion recognition, the change of utterance order will lead to dramatic differences in meaning, which lead to the change of utterance of emotion. For example, in Fig. 1, we notice that the change in utterance order is vital for recognizing emotions. So, the pretext task of utterance order prediction is utilized to obtain utterance consistency. The experiments are performed on two available ERC datasets and the results indicate the proposed approach improves recognition performance compared with those baselines. With comprehensive evaluation and ablation analysis, we have confirmed that fusing the information obtained from our module with the utterance is conducive to enriching the representation of emotional features. Besides, this paper serves as the foundation for our future research, which will build upon the characteristics of conversation. Specifically, we plan to conduct additional pre-training work using emotional word masks and speaker sentence judgments, and combine them together.

Declaration of informed consent

Informed consent was obtained from all individual participants included in the study.

CRediT authorship contribution statement

Dazhi Jiang: Writing – original draft. **Hao Liu:** Writing – review & editing. **Geng Tu:** Data curation. **Runguo Wei:** Formal analysis. **Erik Cambria:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

The authors would like to respect and thank all reviewers for their constructive and helpful review. This research is funded by the National Natural Science Foundation of China (62372283, 62206163), Science and Technology Major Project of Guangdong Province, China (STKJ2021005, STKJ202209002, STKJ2023076), Natural Science Foundation of Guangdong Province, China (2019A1515010943).

References

[1] E. Cambria, R. Mao, M. Chen, Z. Wang, S.-B. Ho, Seven pillars for the future of artificial intelligence, IEEE Intell. Syst. 38 (6) (2023) 62–69.
[2] S. Sabour, C. Zheng, M. Huang, Cem: Commonsense-aware empathetic response generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, (no. 10) 2022, pp. 11229–11237.
[3] K. Zhang, Y. Li, J. Wang, E. Cambria, X. Li, Real-time video emotion recognition based on reinforcement learning and domain knowledge, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2021) 1034–1047.
[4] A. Chatterjee, K.N. Narahari, M. Joshi, P. Agrawal, Semeval-2019 task 3: Emocontext contextual emotion detection in text, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 39–48.
[5] A. Valdivia, V. Luzón, E. Cambria, F. Herrera, Consensus vote models for detecting and filtering neutrality in sentiment analysis, Inf. Fusion 44 (2018) 126–135.
[6] F.A. Pujol, H. Mora, A. Martínez, Emotion recognition to improve e-healthcare systems in smart cities, in: The International Research & Innovation Forum, Springer, 2019, pp. 245–254.
[7] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly available pretrained language models for mental healthcare, in: LREC, 2022, pp. 7184–7190, arXiv preprint arXiv:2110.15621.
[8] D. Hu, L. Wei, X. Huai, Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations, 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), 2021, pp. 7042–7052.
[9] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, Icon: Interactive conversational memory network for multimodal emotion detection, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2594–2604.
[10] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, (no. 01) 2019, pp. 6818–6825.
[11] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, 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), 2017, pp. 873–883.
[12] T. Ishiwatari, Y. Yasuda, T. Miyazaki, J. Goto, 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, 2020, pp. 7360–7370.
[13] W. Li, W. Shao, S. Ji, E. Cambria, BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis, Neurocomputing 467 (2022) 73–82.
[14] J. Li, D. Ji, F. Li, M. Zhang, Y. Liu, Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4190–4200.
[15] W. Shen, S. Wu, Y. Yang, X. Quan, 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), 2021, pp. 1551–1560.
[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335–359.
[17] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: A multimodal multi-party dataset for emotion recognition in conversations, in: ACL, 2019, pp. 527–536.
[18] E. Cambria, N. Howard, J. Hsu, A. Hussain, Sentic blending: Scalable multimodal fusion for continuous interpretation of semantics and sentics, in: IEEE SSCI, Singapore, 2013, pp. 108–117.
[19] E. Cambria, A. Hussain, Sentic album: Content-, concept-, and context-based online personal photo management system, Cogn. Comput. 4 (4) (2012) 477–496.
[20] F. Xing, F. Pallucchini, E. Cambria, Cognitive-inspired domain adaptation of sentiment lexicons, Inf. Process. Manage. 56 (3) (2019) 554–564.
[21] D. Huang, S. Zhou, D. Jiang, Generator-based domain adaptation method with knowledge free for cross-subject EEG emotion recognition, Cogn. Comput. (2022) 1–12.
[22] S. Zhou, D. Huang, C. Liu, D. Jiang, Objectivity meets subjectivity: A subjective and objective feature fused neural network for emotion recognition, Appl. Soft Comput. 122 (2022) 108889.
[23] D. Jiang, H. Liu, G. Tu, R. Wei, Window transformer for dialogue document: a joint framework for causal emotion entailment, Int. J. Mach. Learn. Cybern. (2023) 1–11.
[24] H.T. Nguyen, P.H. Duong, E. Cambria, Learning short-text semantic similarity with word embeddings and external knowledge sources, Knowl.-Based Syst. 182 (104842) (2019).
[25] F. Xing, E. Cambria, R. Welsch, Intelligent asset allocation via market sentiment views, IEEE Comput. Intell. Mag. 13 (4) (2018) 25–34.
[26] S. Poria, N. Majumder, R. Mihalcea, E. Hovy, Emotion recognition in conversation: Research challenges, datasets, and recent advances, IEEE Access 7 (2019) 100943–100953.
[27] Y. Ma, K.L. Nguyen, F.Z. Xing, E. Cambria, A survey on empathetic dialogue systems, Inf. Fusion 64 (2020) 50–70.
[28] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, S. Niu, DailyDialog: A manually labelled multi-turn dialogue dataset, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2017, pp. 986–995.
[29] S.M. Zahiri, J.D. Choi, Emotion detection on TV show transcripts with sequence-based convolutional neural networks, in: AAAI Workshops, 2018.
[30] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, DialogueGCN: A graph convolutional neural network for emotion recognition in conversation, in: EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2020.
[31] W. Nie, R. Chang, M. Ren, Y. Su, A. Liu, I-GCN: Incremental graph convolution network for conversation emotion detection, IEEE Trans. Multimed. 24 (2022) 4471–4481, http://dx.doi.org/10.1109/TMM.2021.3118881.
[32] W. Li, L. Zhu, R. Mao, E. Cambria, SKIER: A symbolic knowledge integrated model for conversational emotion recognition, in: AAAI, 2023, pp. 13121–13129.
[33] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, S. Poria, COSMIC: COmmonSense knowledge for emotion identification in conversations, in: Findings of the Association for Computational Linguistics, EMNLP 2020, 2020, pp. 2470–2481.
[34] G. Tu, J. Wen, C. Liu, D. Jiang, E. Cambria, Context- and sentiment-aware networks for emotion recognition in conversation, IEEE Trans. Artif. Intell. 3 (5) (2022) 699–708.
[35] W. Nie, Y. Bao, Y. Zhao, A. Liu, Long dialogue emotion detection based on commonsense knowledge graph guidance, IEEE Trans. Multimed. (2023) 1–15, http://dx.doi.org/10.1109/TMM.2023.3267295.
[36] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: NAACL, 2018, pp. 2122–2132.
[37] G. Tu, J. Wen, H. Liu, S. Chen, L. Zheng, D. Jiang, Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models, Knowl.-Based Syst. 235 (2022) 107598.
[38] G. Xiao, G. Tu, L. Zheng, T. Zhou, X. Li, S.H. Ahmed, D. Jiang, Multimodality sentiment analysis in social internet of things based on hierarchical attentions and CSAT-TCN with MBM network, IEEE Internet Things J. 8 (16) (2020) 12748–12757.
[39] D. Jiang, R. Wei, H. Liu, J. Wen, G. Tu, L. Zheng, E. Cambria, A multitask learning framework for multimodal sentiment analysis, in: 2021 International Conference on Data Mining Workshops, ICDMW, 2021, pp. 151–157, http://dx.doi.org/10.1109/ICDMW53433.2021.00025.
[40] D. Jiang, H. Liu, R. Wei, G. Tu, CSAT-FTCN: A fuzzy-oriented model with contextual self-attention network for multimodal emotion recognition, Cogn. Comput. (2023) 1–10.

[41] S. Gidaris, P. Singh, N. Komodakis, Unsupervised representation learning by predicting image rotations, in: International Conference on Learning Representations, 2018.
[42] R. Zhang, P. Isola, A.A. Efros, Colorful image colorization, in: European Conference on Computer Vision, Springer, 2016, pp. 649–666.
[43] I. Misra, C.L. Zitnick, M. Hebert, Shuffle and learn: Unsupervised learning using temporal order verification, in: European Conference on Computer Vision, Springer, 2016, pp. 527–544.
[44] R. Zellers, X. Lu, J. Hessel, Y. Yu, J.S. Park, J. Cao, A. Farhadi, Y. Choi, Merlot: Multimodal neural script knowledge models, Adv. Neural Inf. Process. Syst. 34 (2021) 23634–23651.
[45] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, (no. 12) 2021, pp. 10790–10797.
[46] J.D.M.-W.C. Kenton, L.K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[47] J. Wu, X. Wang, W.Y. Wang, Self-supervised dialogue learning, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3857–3867.
[48] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880.
[49] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.
[50] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019.

Dazhi Jiang received his BA in Computer Science from the China University of Geoscience (Wuhan) in 2004. He obtained his PhD from the State Key Laboratory of Software Engineering, Wuhan University, China in 2009. Since then, he has been with the Department of Computer Science, Shantou University, China where he was a Professor. His research interests include affective computing, deep learning, data mining and applications of artificial intelligence.

Hao Liu is a graduate student of the Department of Computer Science at Shantou University. His current research interests include affective computing and deep learning.

Geng Tu is a graduate student of the Department of Computer Science at Shantou University. His current research interests include affective computing and deep learning.

Runguo Wei is currently pursuing the master's degree with the Department of Computer Science at Shantou University, China. His current research focuses on affective computing and machine learning, etc.

Erik Cambria is the Founder of SenticNet, a Singapore-based company offering B2B sentiment analysis services, and an Associate Professor at NTU, where he also holds the appointment of Provost Chair in Computer Science and Engineering. His research focuses on the ensemble application of symbolic and subsymbolic AI to natural language processing tasks such as sentiment analysis, dialogue systems, and financial forecasting. Erik is recipient of many awards, e.g., the 2019 IEEE Outstanding Early Career Award, he was listed among the 2018 AI's 10 to Watch, and was featured in Forbes as one of the 5 People Building Our AI Future.