

Title	文脈、感情の動態、および話者パーソナリティのモデリングを取り入れた感情認
Author(s)	XUE, JIEYING
Citation	
Issue Date	2025-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/20078
Rights	
Description	Supervisor: NGUYEN, Minh Le, 先端科学技術研究科, 博士

Doctoral Dissertation

Emotion Detection with Context, Emotional Dynamics, and Speaker Personality Modeling

XUE Jieying

Supervisor : NGUYEN Le Minh

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Information Science

September, 2025

Abstract

Our research encompasses primarily two interrelated areas: Emotion Recognition in Conversations (ERC) and multilingual multi-label emotion detection. The former aims to identify the emotional state of each utterance in a dialogue, while the latter addresses the detection of multiple emotions across languages within a given sample.

Within the ERC domain, we explore several critical yet underexplored dimensions: In emotion context modeling, traditional sequential models often capture only local emotional dependencies, overlooking long-range emotional influences that may arise between distant parts of a conversation. We argue that emotional states can be affected and transmitted by speakers and utterances throughout the conversation, regardless of their positional distance. In response to this limitation, we propose the ***Long-range dependencY emotionS Model (LYSM)***, which employs self-attention mechanisms to capture emotional dynamics throughout entire conversations, allowing the system to integrate emotional dependencies from both nearby and distant utterances. Experimental results on four benchmark datasets confirm its strong generalizability and effectiveness.

In utterance representation, unlike isolated sentence-level emotion tasks, the core challenge in dialogue emotion recognition lies in effectively representing a target utterance within its conversational context. Although pre-trained language models (PLMs) such as RoBERTa offer strong capabilities for context modeling, existing PLM-based approaches often fail to fully exploit their potential for fine-grained contextual encoding. To overcome this limitations, we introduce ***Accumulating Word Representations in Multi-level Context Integration for ERC Task (AccWR)***, which aggregates multi-level contextual word information before inputting it into the PLM. This enriched contextual word aggregation enhances both semantic understanding and the model’s focus on the target utterance. Experimental results show that AccWR consistently outperforms baseline models on four benchmark datasets and demonstrates strong potential for broader applications such as

response generation and semantic parsing.

For speaker modeling, ERC typically relies on spoken dialogues transcribed by automatic speech recognition systems. Individual traits such as linguistic style and personality significantly influence emotional expression, yet prior work often depends on implicitly learned speaker features, limiting interpretability and cross-domain generalization. To address this, we propose *BiosERC: Integrating **B**iography-Based **S**peaker Representations with Large Language Models for **E**motion **R**ecognition in **C**onversations*, which employs LLMs with prompt-based techniques to extract explicit speaker profiles as external knowledge. These biography-based representations enhance the emotional understanding of each speaker, leading to more accurate and nuanced recognition, especially in complex or multi-party conversations. BiosERC achieves competitive or state-of-the-art (SOTA) performance on three benchmark datasets.

In multi-level context modeling, the spontaneous nature of conversations makes it difficult to capture transient and dynamically evolving emotional states using only contextual discourse and static speaker profiles. To address this, we propose *TraceERC: **T**racking **R**elational **A**wareness of Contextual, **C**haracter, and **E**motional States in **E**motion **R**ecognition in **C**onversations*, which leverages LLMs to jointly encode dialogue context, speaker personality, and dynamic emotional cues. These enriched representations enable emotion predictions that are sensitive to both conversational flow and individual speaker characteristics, enhancing emotional understanding and adaptability. As one of the first LLM-based ERC models to incorporate contextual learning and in-context fine-tuning, TraceERC achieves SOTA results on MELD and strong performance across benchmarks.

In the multilingual domain, our team participated in SemEval-2025 Task 11, tackling both multi-label classification (Track A) and emotion intensity detection (Track B). We developed a generation-based framework leveraging multilingual PLMs and LLMs to support both high- and low-resource languages.

Keywords: Emotion Recognition in Conversations, Multilingual Multi-label Emotion Detection, Large Language Models, Speaker Personality Modeling, Emotion Context Modeling

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my academic advisor, Professor Nguyen Le Minh of the Japan Advanced Institute of Science and Technology (JAIST), for his continuous guidance and support throughout my research journey. His insightful mentorship helped me define a meaningful research direction and taught me how to navigate and overcome academic challenges with clarity and confidence.

I am also sincerely grateful to the members of my thesis committee: Professor Kiyooki Shirai, Professor Shogo Okada, Associate Professor Naoya Inoue, Professor Shinobu Hasegawa, and Professor Ken Satoh of the National Institute of Informatics, for their invaluable comments and suggestions. Their critical feedback not only highlighted important areas for improvement but also deepened my understanding of the research topic.

I would like to thank JAIST for its generous financial support, which enabled me to present my work at international conferences. These opportunities significantly broadened my academic perspective and strengthened my skills in scholarly communication and international collaboration. My sincere appreciation also goes to all the staff members at JAIST for their unwavering kindness and support. Their thoughtful assistance created a welcoming and comfortable environment that allowed me to focus fully on my research and personal growth. I am especially thankful to all members of the Nguyen Laboratory for their friendship, encouragement, and collaboration. My time at JAIST, both in academic endeavors and daily life has been among the most rewarding and memorable periods of my academic career.

Finally, I would like to express my deepest gratitude to my family. Their unconditional love, patience, and unwavering belief in me have been the foundation of my journey. Without their support, none of this would have been possible.

Contents

Abstract	i
Acknowledgments	iii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Literature Review	4
1.3 Research Objectives and Contributions	5
1.4 Dissertation Outline	9
2 Emotional Dependency Modeling	12
2.1 Motivation and Objectives	12
2.2 Related Works	13
2.3 Methodology	14
2.3.1 Task definition.	15
2.3.2 COSMIC Framework	15
2.3.3 LYSM Architecture	17
2.3.4 Conditional Random Field (CRF)	18
2.4 Experiment and Result Analysis	19
2.4.1 Experimental Setup	19
2.4.2 Main Results	20
2.4.3 Ablation Study	21
2.4.4 Improvement Example.	22

2.4.5	Emotional Dependency.	25
2.5	Chapter Summary	25
3	Utterance Modeling	27
3.1	Motivation and Objectives	27
3.2	Related Works	28
3.3	Methodology	29
3.3.1	Baseline	30
3.3.2	Our Proposed Model (AccWR).	32
3.4	Experiment and Result Analysis	33
3.4.1	Experimental Setup.	33
3.4.2	Main Results	34
3.4.3	Ablation Study	35
3.4.4	Convergence Time	36
3.5	Chapter Summary	37
4	Speaker Modeling	38
4.1	Background and Motivation	38
4.2	Related Works	40
4.3	Methodology	41
4.3.1	Intra-inter ERC (baseline)	41
4.3.2	BiosERC: Integrating Biography Speakers Supported by LLMs for ERC Tasks	43
4.4	Experiment	47
4.5	Result Analysis	48
4.5.1	Main Results	48
4.5.2	Ablation Study	49
4.5.3	Conversation Length	51
4.5.4	Case Study	51
4.6	Chapter Summary	52

5	TraceERC: Tracking Relational Awareness of Contextual, Character, and Emotional States in Emotion Recognition in Conversations	55
5.1	Background and Motivation	55
5.2	Related Works	58
5.3	Methodology	59
5.3.1	Problem Definition	60
5.3.2	Baseline - LLM-based ERC	60
5.3.3	TraceERC	61
5.3.4	Speaker Personality Extraction.	63
5.3.5	Learning Scenarios	65
5.4	Experiment	70
5.5	Result Analysis	71
5.5.1	Comparison with Previous Works	71
5.5.2	Ablation Study	73
5.5.3	Analysis of Emotional Context Quality and Its Impact on Performance	79
5.5.4	Error Analysis	80
5.5.5	Case Study	83
5.6	Chapter Summary	86
6	Cross-Lingual Multi-Label Emotion Detection Using Generative Models	87
6.1	Motivation and Objectives	87
6.2	Related Works	88
6.3	Methodology	90
6.3.1	System Overview	90
6.3.2	Methods	90
6.4	Experiment	92
6.5	Result Analysis	95
6.5.1	Track A: Multi-label Emotion Detection.	95

6.5.2	Track B: Emotion Intensity.	95
6.5.3	Result Analyses	96
6.6	Chapter Summary	98
7	Conclusion and Future Work	100
7.1	Conclusion and Future Work	100
7.1.1	Conclusion	100
7.2	Future Work	102
	Publications and Awards	104

List of Figures

1.1	Overview of the Thesis Framework	11
2.1	Example illustrating long-range emotional dependencies in a conversation from the IEMOCAP dataset.	13
2.2	The architecture of the proposed <i>Long-range dependencY emotionS Model</i> (<i>LYSM</i>), built on top of the COSMIC.	17
2.3	Performance comparison between LYSM and COSMIC on the IEMOCAP, stratified by the number of utterances per conversation (#conversation). .	23
2.4	Comparison of weighted average F1 scores between COSMIC and LYSM on the IEMOCAP dev set.	24
2.5	Heatmap visualization of dependencies between pairs of emotional states in a conversation.	25
3.1	An example of the ERC task and a comparison of our approach with pre- vious BERT-based methods.	27
3.2	Ablation study on the IEMOCAP development set evaluating different word vector aggregation methods. The red bar indicates the median per- formance for each configuration.	36
3.3	Loss and performance comparison between our proposed method and the ([CLS]) baseline on the IEMOCAP development set. The shaded area around the line graph represents the variability across ten different runs.	36
4.1	Overview of the Proposed BiosERC framework	39
4.2	Overview of our BiosERC model architecture.	45

4.3	Performance comparison between our BERT-based BiosERC model and the baseline on the MELD development set, illustrating variability across 10 random runs.	49
4.4	Performance comparison respect to length of conversation (number of utterance) on the MELD development set (variability across 10 random runs).	52
5.1	Overview of our TraceERC framework architecture.	57
5.2	Idea of our TraceERC framework compared to the baseline model.	62
5.3	Architecture of our TraceERC framework.	63
5.4	Length distribution comparison of different speaker description variants on the MELD dataset.	75
5.5	Performance comparison of our TraceERC (excluding emotional context information) with the baseline method across four types of speaker descriptions on the MELD development set.	75
5.6	Dependence of performance between the emotional context model and our TRACEERC model. The blue dashed line shows the equivalent alignment between the x - and y -axes of the graph.	79
5.7	Impact of emotional context quality on model performance in the local context setting	81
5.8	Comparison of confusion matrices between our TRACEERC (right) and the baseline system (left) on the IEMOCAP dataset.	82
5.9	Comparison of confusion matrices between our TRACEERC (right) and the baseline system (left) on the MELD dataset.	82
5.10	Comparison of confusion matrices between our TRACEERC (right) and the baseline system (left) after mapping to sentiment labels on the IEMOCAP dataset.	83

5.11	Comparison of confusion matrices between our TRACEERC (right) and the baseline system (left) after mapping to sentiment labels on the MELD dataset.	83
6.1	Distribution of improved samples between <i>base</i> and <i>pairwise</i> strategies, categorized by the number of emotions (track A).	97
6.2	Distribution of improved samples between <i>base</i> and <i>pairwise</i> strategies, categorized by the number of emotions (track B).	97
6.3	Overall performance of the pairwise strategy across all emotional labels and languages (Track B).	98

List of Tables

1.1	Statistical information on all ERC datasets. The character $\#$ denotes the size of the set.	4
2.1	Performance comparison between methods.	21
2.2	Ablation study on the IEMOCAP dataset.	22
2.3	Improvement example collected in IEMOCAP dataset. The green and red labels indicate the correct and incorrect model predictions, respectively. .	23
3.1	Performance comparison between our proposed method and existing SOTA approaches.	35
4.1	Prompting template used to extract speaker characteristics from a conversation using an LLM.	44
4.2	Prompting nput template using speaker description and content of conversation for fine-tuning LLMs.	47
4.3	Performance comparison between our proposed method and prior works on the test sets. The column $\#T.Params.$ denotes the number of trainable parameters. Symbols \ddagger and \dagger indicate statistically significant improvements over the baseline according to a t-test, at levels $p < 0.01$ and $p < 0.05$, respectively.	48
4.4	Performance comparison among variants of BiosERC on the MELD development set.	50

4.5	Case study of improvement examples from the MELD dataset. Red and green labels indicate incorrect and correct model predictions, respectively.	53
5.1	Prompting input template and example for the baseline model.	61
5.2	TraceERC prompting input template and example using speaker description and emotional context information.	66
5.3	Performance comparison between our proposed method and previous works on the test sets. Column $\#T.Params.$ refers to the number of trainable parameters. The marked star (*) indicates the preprints (unofficial works). The abbreviation “ <i>w/o EC</i> ” stands for the setting without using emotional context information. The notation \ddagger indicates the significant difference (t-test) with the baseline in levels $p < 0.05$.	72
5.4	Performance comparison between our proposed method and previous works on the test sets in the <i>in-context learning</i> paradigm. The marked star (*) indicates the preprints (unofficial works).	73
5.5	Prompting template for generating different types of speaker descriptions and their corresponding output examples.	74
5.6	Performance comparison between our proposed method and the baseline model on the development set under the <i>in-context learning</i> and <i>instruction fine-tuning</i> paradigms. The notation \ddagger denotes the significant difference (t-test) from the baseline in levels $p < 0.05$.	77
5.7	Performance comparison among variants of BiosERC on the MELD development set.	80
5.8	Case study of improvement examples collected from two datasets. The red and green labels refer to the incorrect and correct prediction by the models, respectively.	84
5.9	Case study of improvement examples collected from two datasets. Red and green labels indicate incorrect and correct model predictions, respectively.	85

5.10	Case study of improvement examples collected from two datasets. Red and green labels refer to the incorrect and correct model predictions, respectively.	85
6.1	Instruction prompting using the <i>base</i> template (track A).	92
6.2	Examples of output format for text generation.	92
6.3	Instruction prompting templates for the <i>base</i> (top) and <i>pairwise</i> (bottom) strategies in Track A.	92
6.4	Instruction prompting template for the <i>base</i> (top) and <i>pairwise</i> (bottom) strategies in Track B.	93
6.5	Results of Sub-task A. For a fair comparison, the <i>average</i> column is computed based on all languages except for <i>orm</i> , <i>ron</i> , <i>ptbr</i> , and <i>ptmz</i> , as these languages are missing in some settings. The red-highlighted values indicates the best-performing configurations used for the final test submission. The asterisk (*) marks results obtained during the post-evaluation phase.	94
6.6	Results of Sub-task B. The meanings of the denotations and colors are the same as in Table 6.5.	94
7.1	Performance Comparison of Proposed and Existing Methods. The columns EmoDept, SpkrMd, LocalCtx, and GlobalCtx refer to the Emotional dependencies, Speaker, Local context, and Global context modeling.	102

Chapter 1

Introduction

1.1 Background and Motivation

As a pivotal area within natural language processing (NLP), emotion recognition tasks have received extensive attention over the past decades, driven by the continuous advancement of artificial intelligence [13]. Among its various subfields, Emotion Recognition in Conversations (ERC) has emerged as a particularly important area due to its broad range of applications, including mental health care [57], intelligent systems [53], and social media analysis [35, 80, 45], among others. In parallel, the accelerating pace of global digitalization has led to increasing reliance on social media platforms for communication across diverse linguistic and cultural communities. Against this backdrop, multilingual and multi-label emotion detection has become an increasingly critical research focus [51]. Therefore, our research on text-based emotion recognition primarily focuses on two key subfields: ERC and multilingual multi-label emotion detection.

In terms of ERC, the task diverges significantly from traditional sentiment analysis [52], as conversational data typically consists of spoken dialogue transcriptions characterized by fragmented utterances and non-standard grammatical structures. Consequently, ERC necessitates a more nuanced understanding of dialogue-specific challenges, including conversational context [19, 54, 15, 43, 37], speaker-specific information [90, 31, 33], and the intricate emotional dependencies across utterances [18, 30]. These complexities often

result in ambiguous or even misleading emotional interpretations when utterances are analyzed in isolation. It is well established that sentences removed from their dialogue context are prone to divergent emotional interpretations, particularly in the case of short or reactive utterances, which are common in conversational exchanges.

For the purpose of modeling the conversational context, early studies [15] employed sequential architectures such as LSTM or GRU models to process utterances or entire conversations sequentially in order. To better reflect the intrinsic dependencies among utterances, the graph-based approaches [30, 63] represent conversations as nodes and edges of the graph to effectively model the underlying structure in the conversation. Transformer-based techniques [48, 33] have made great contributions by directly adopting or modifying original transformer blocks to establish long-range emotional correlations in conversational scenarios. However, relying solely on contextual information is insufficient for accurate emotion recognition, as speaker-specific characteristics also play a vital role. For instance, individuals with an optimistic personality tend to exhibit positive emotions throughout a conversation, while those with a more pessimistic disposition may convey negative sentiments even through semantically neutral utterances. Notably, speakers who frequently employ sarcasm often express emotions that contradict the literal meaning of their words, making their emotional intent difficult to infer without understanding their underlying personality traits.

Although speaker modeling has received increasing attention in recent ERC research [90, 31, 33], progress has been hindered by the lack of datasets annotated with explicit personality attributes. Consequently, most approaches rely on implicit modeling of inter-speaker and intra-speaker relationships, which leaves the true value and utility of speaker-specific information underexplored. Furthermore, the absence of explicit speaker features reduces the interpretability and generalizability of such models across other dialogue-related tasks. In particular, speaker personality—an inherently stable and influential factor—remains an underutilized resource in current ERC systems.

Moreover, emotional dependencies between utterances have been significantly underexamined. Emotional states can propagate through interactions between speakers, regard-

less of the physical distance between utterances in the dialogue. The overall emotional tone of a conversation provides critical cues for accurately identifying individual emotions.

Finally, the spontaneous and dynamic nature of real-world conversations introduces additional challenges. Unpredictable events—such as a speaker undergoing a sudden personal incident or participating in a dialogue while carrying unrelated emotional burdens—are common and can substantially influence the emotional dynamics of a conversation. Therefore, modeling dynamic emotional states in conjunction with speaker personality is essential for achieving robust and context-aware emotion recognition.

Therefore, the guiding philosophy behind our research, we believe that emotion recognition models should not merely learn to label emotions from data, but should instead be designed to mimic the way humans reason about emotions in conversation. When people interpret the emotional content of a dialogue, they naturally consider a variety of interrelated factors: the conversational context and background, the personality and communication style of the speaker, and the emotional trajectory that has unfolded over the course of the interaction. Our core objective is to incorporate these human-like reasoning processes into the model design. By enabling models to attend to the same types of information that humans rely on—such as speaker traits, contextual cues, and emotional transitions—we aim to overcome the limitations of conventional ERC systems. This perspective motivates our multi-context framework, which seeks not only to improve accuracy, but also to bring emotion recognition closer to human-level interpretability and cognitive alignment.

Dataset. In all sections related to ERC, we primarily focus on evaluating the performance of our proposed architecture using four benchmark datasets (as shown in Table 1.1).

- **IEMOCAP** [5]: the dataset of six different emotion categories collected from conversations of ten different speakers, each conversation contains utterances of two persons.
- **DailyDialog** [37]: the largest multi-utterance dialogue dataset collected in daily life

conversations, including seven different emotion categories. Following the previous work experimental setup, we ignore the label *neutral* when computing the evaluation score because this label is highly imbalanced in 83% of utterances across the whole dataset.

- **MELD** [55] and **EmoryNLP**: the datasets of seven different types of emotions scraped from TV shows.

The IEMOCAP is the dataset that contains long dialogues with an average of around 50 utterances per conversation, while DailyDialog is the dataset that contains many topics in conversation.

Table 1.1: Statistical information on all ERC datasets. The character # denotes the size of the set.

Dataset	# dialogues			# utterances		
	train	dev	test	train	dev	test
IEMOCAP	108	12	31	5,163	647	1,623
DailyDialog	11,118	1,000	1,000	87,823	7,912	7,836
MELD	1,039	114	280	9,989	1,109	2,610
EmoryNLP	659	89	79	7,551	954	984

1.2 Literature Review

Existing approaches for ERC approaches can be broadly categorized into four primary paradigms: recurrent neural network (RNN)-based, graph neural network (GNN)-based, transformer-based, and pre-trained language model (PLM)-based models [19, 15, 43, 37]. Early studies on modeling conversational context [15] primarily employed sequential architectures such as LSTM or GRU models to process utterances or entire conversations sequentially in a sequential manner. To better capture the intrinsic dependencies among utterances, the GNN-based approaches [30, 63] model conversations as graphs, where utterances serve as nodes and their relationships as edges, effectively representing the underlying conversational structure. Transformer-based techniques [48, 33] have advanced

the field by directly adopting or modifying transformer architectures to capture long-range emotional dependencies within conversations. Leveraging the strong representation capabilities of PLMs, previous works [31] utilize PLMs as utterance encoders, fine-tuning them to improve contextual representation and enhance ERC performance. Additionally, knowledge-enhanced methods [15] integrate external commonsense and domain knowledge into hierarchical transformer models to better capture complex conversational interactions.

1.3 Research Objectives and Contributions

To address the aforementioned challenges, our research is guided by the following objectives:

Emotional Dependencies Modeling: Regardless of the distance between utterances or the overall length of the dialogue, emotional states can be transmitted and influenced through both speakers and utterances. To effectively capture the subtle and complex emotional dependencies present in conversations, particularly those spanning long-range utterances. We propose a novel architecture, the *Long-range dependency emotionS Model (LYSM)*, which is designed to enhance ERC performance by learning the interactions between emotional states across dialogue contexts. This enables the generation of richer and more contextually grounded emotion representations, thereby improving the model’s ability to recognize emotional correlations between utterances.

We also conducted experiments to evaluate the performance of our proposed model compared with previous methods on four popular benchmark datasets: IEMOCAP, DailyDialog, EmoryNLP, and MELD. Experimental results showed that our proposed model works effectively and achieves competitive results with the current SOTA results, and outperform the baseline models on all experimental datasets.

Utterance Modeling: Since the meaning of an utterance in dialogue often depends on its surrounding context rather than its isolated semantics, a central challenge in ERC lies in effectively modeling the target utterance within its conversational context. To

address this, we proposed ***Accumulating Word Representations in Multi-level Context Integration for ERC Task (AccWR)***, a dual-focus mechanism that incorporates both the contextual information from surrounding utterances and emphasizes the accumulation of word-level representations of the focal utterance. Instead of relying on special markers such as [CLS], we extract the hidden states of all constituent words in the focal utterance and aggregate them using a Multi-Layer Perceptron (MLP) to construct a more expressive sentence representation. This approach enables a more comprehensive utilization of the representational capacity of PLMs, while facilitating selective attention to salient emotional cues embedded in the dialogue context.

Experimental results on four widely-used benchmark datasets demonstrate that our method not only outperforms previous approaches, but also achieves performance comparable to recent SOTA models. Moreover, the proposed mechanism shows strong potential for broader application in context-sensitive sentence modeling tasks, such as dialogue response generation, multi-turn question answering, and conversational semantic parsing.

Speaker Modeling: To capture speaker-specific traits, we proposed BiosERC, a novel model that accurately extracts each speaker’s personality characteristics in a dialogue, including lexical preferences, personality traits, and emotional fluctuation patterns. This approach addresses the absence of explicit speaker information in existing ERC datasets and enables a deeper understanding of individual personas and emotional transitions, thereby enhancing the modeling of emotional dynamics. Our method utilizes prompting techniques with LLMs to extract biographical information about speakers, showing strong adaptability across various conversational scenarios.

Importantly, the personality extraction mechanism we introduce is explicit and verifiable, offering a new perspective for ERC. It is interpretable, scalable, and applicable to a wide range of dialogue analysis tasks. We conducted extensive experiments on three benchmark datasets: IEMOCAP, MELD, and EmoryNLP. The results consistently demonstrate that our method achieves SOTA performance, validating the effectiveness of the proposed model.

Multi-level Context Modeling (Semantic, Emotional, and Speaker Contexts):

In the preceding chapters, we conducted an in-depth exploration of emotional dependency modeling, sentence modeling, and speaker modeling. However, given the inherently spontaneous and unpredictable nature of conversations, which may occur at any time or place, speakers may experience sudden and unanticipated emotional shifts. These shifts can be triggered by personal memories or unexpected events, resulting in significant alterations in emotional trajectories. Such multi-source uncertainty greatly intensifies the challenges associated with ERC. Nevertheless, most existing ERC research focuses primarily on static conversational information and tends to neglect the dynamic emotional changes that occur during interactions. Specifically, it fails to capture the evolving trends and causal factors of emotional states at each time step. Furthermore, due to the unstructured nature of dialogue and the absence of speech and visual modalities, text-only ERC approaches struggle to capture fine-grained emotional details, making it particularly difficult to promptly detect emotional shifts and disambiguate between similar emotion labels. Therefore, it is essential to model the temporality of emotions throughout the dialogue by incorporating contextual information available at each conversational point.

To address these limitations, we formalize three critical aspects that influence human emotional expression in conversations: utterance content, static speaker-specific traits, and dynamic emotional cues arising during interactions. Based on these dimensions, we proposed TraceERC: Tracking Relations Aware of Context, Character, and Emotional State for ERC. This framework integrates both dynamic emotional states and static personality traits, enabling the joint modeling of dependencies between these factors. It bridges a major gap in existing research that neglected the crucial relationship between speaker personality and emotional relativity.

More specifically, TraceERC leverages the powerful contextual modeling capabilities of LLMs to extract speaker personality traits from the dialogue, encode emotional atmosphere and trends, and generate supplementary knowledge enriched with emotional and contextual information. These features are used to construct context-aware, emotion-labeled exemplars tailored to the specific dialogue scenario and speaker characteristics, thereby improving the model’s ability to interpret emotional dynamics accurately.

Moreover, as one of the pioneering studies exploring in-context learning and instruction tuning for LLM-based ERC, we conduct comprehensive evaluations to assess the contribution of each component. Experimental results across three benchmark datasets demonstrate that TraceERC achieves competitive performance and establishes a SOTA on the MELD dataset. These results validate the effectiveness of our approach. In addition to its performance, the simplicity and modularity of TraceERC make it highly adaptable to other dialogue-related NLP tasks, highlighting its broad applicability and practical potential.

Cross-Lingual Sentiment Analysis: With the rapid expansion of social media and the increasing prevalence of digital communication worldwide, online platforms have become the primary channels for information exchange and opinion sharing [51]. This shift has led to a growing demand for cross-lingual sentiment detection systems capable of analyzing multilingual user-generated content [51, 74]. Consequently, multilingual sentence-level sentiment analysis has emerged as a critical task for tracking public sentiment across languages and regions. A key challenge in text-based sentiment recognition is that a single sentence often conveys multiple emotions with varying intensities [12]. This complexity is compounded by the limited availability of annotated training data—particularly for low-resource languages—and the suboptimal performance of existing PLMs for these languages.

To address these challenges, this study explores the integration of PLMs such as XML-RoBERTa for multi-label sentiment analysis across both high-resource languages (e.g., English, Chinese) and low-resource languages (e.g., several African languages). We formulate multi-label sentiment recognition as a text generation task and leverage the multilingual capabilities of PLMs to improve semantic understanding and sentiment tone identification, especially in underrepresented languages. Furthermore, we introduce two complementary strategies—the *base* and *pairwise* approaches—to tackle the inherent complexity of multi-label classification. Our experimental results demonstrate that models trained on combined multilingual datasets outperform those trained on individual monolingual datasets, underscoring the benefits of cross-lingual knowledge transfer.

We contribute to two tracks of SemEval-2025 Task 11: Bridging the Gap in Text-Based Sentiment Detection. Extensive evaluations, including ablation studies, confirm the effectiveness of our proposed methods in addressing the two subtasks: multi-label sentiment classification (Track A) and sentiment intensity detection (Track B). In Track A, our system ranks within the top four for ten languages and achieves first place in Hindi. In Track B, we secure top-five rankings in seven languages, further validating the robustness and generalizability of our approach.

1.4 Dissertation Outline

The remainder of this thesis is organized as follows:

- **Chapter 2** introduces the emotional dependency modeling framework, with a particular focus on capturing emotional dependencies between both short- and long-range utterances. The chapter provides a detailed description of the model architecture and presents experimental results that validate the effectiveness of each component.
- **Chapter 3** presents an utterance utterance modeling architecture. It introduces a mechanism that jointly leverages contextual utterances and word-level representations of the focal sentence. This design enhances the model’s ability to capture emotional cues while fully utilizing the contextual modeling capacity of PLMs. The chapter also reports extensive experimental results across four benchmark datasets, demonstrating the effectiveness and generalizability of the proposed method in comparison with recent SOTA approaches.
- **Chapter 4** proposes a speaker modeling framework that leverages LLMs to extract explicit speaker-related information—such as personality traits—to enrich ERC with speaker-specific external knowledge. This chapter provides a comprehensive evaluation by comparing our approach with SOTA methods and validates the effectiveness of each model component through detailed ablation studies.

- **Chapter 5** builds upon the content of the preceding chapters by further introducing dynamic emotional context and proposing TraceERC: Tracking Relational Awareness of Contextual, Character, and Emotional States in Emotion Recognition in Conversations. This model constitutes a comprehensive framework that integrates multi-level contextual information, encompassing the utterance level, static speaker level, and dynamic emotional level. Extensive experimental evaluations are conducted to demonstrate the effectiveness and superiority of each component within the proposed architecture.
- **Chapter 6** presents our achievements in multilingual, sentence-level, multi-label sentiment recognition, with a particular focus on both high-resource and low-resource languages. By participating in SemEval-2025 Task 11, we demonstrated the effectiveness of our proposed methods.
- **Chapter 7** concludes the thesis by summarizing the key findings, contributions, limitations, and potential directions for future research.

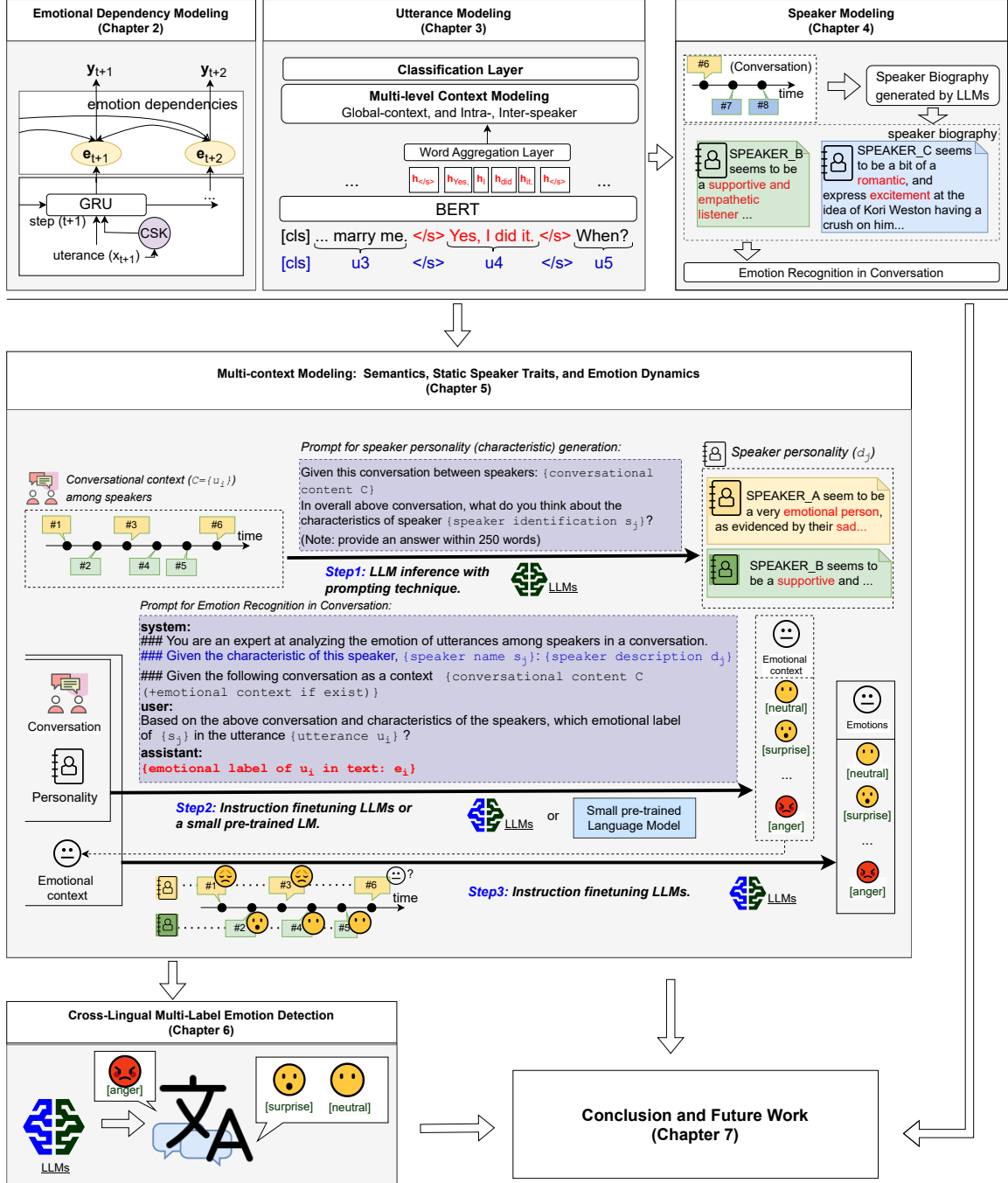


Figure 1.1: Overview of the Thesis Framework

Chapter 2

Emotional Dependency Modeling

2.1 Motivation and Objectives

The relationship between emotional states has been a crucial yet significantly overlooked aspect in ERC [65, 18]. SOTA models for ERC primarily focus on integrating prior knowledge through external commonsense extractors [34, 15, 92, 3] or leveraging PLMs [60, 18] to construct utterance representations that are fused with the surrounding conversational context. However, these architectures implicitly treat the emotional states of utterances as a temporally ordered sequence and model the dependencies between them in a forward or bidirectional manner. This sequential assumption, which focuses on emotion transitions between adjacent utterances, fails to capture potentially strong long-range dependencies between emotional states that may occur across distant parts of the dialogue. Our intuition is that emotional states can be influenced and transmitted across speakers and utterances throughout a conversation, regardless of their distance or the overall length of the conversation. For instance, an expression of anger by a speaker early in a conversation may reappear or be referenced in a seemingly unrelated utterance much later, interspersed with many turns from other speakers. Traditional sequential modeling tends to overlook such long-range or discontinuous emotional dependencies.

As shown in Figure 2.1, the emotional states in a conversation can propagate not only through adjacent utterances (e.g., u3, u4, u5) but also across long-range distance

utterances from the same or different speakers (e.g., u3, u8, u33). Although the two utterances from the same speaker (e.g., u5, u33) may appear emotionally neutral when interpreted in isolation at the literal level, they in fact convey positive emotions such as excitement and happiness within the broader conversational context. This observation highlights that the emotional meaning of an utterance can be shaped by the emotional tone of preceding or even distant utterances—regardless of their distance or position in the dialogue.

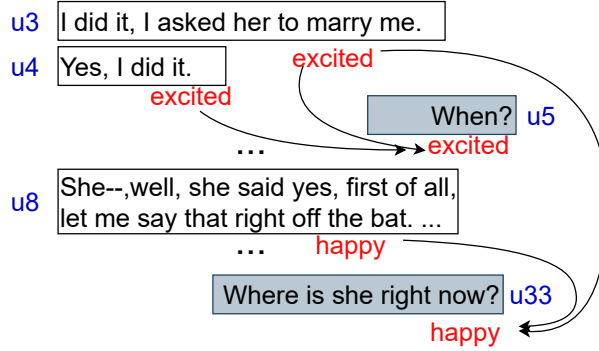


Figure 2.1: Example illustrating long-range emotional dependencies in a conversation from the IEMOCAP dataset.

To effectively capture these subtle and complex emotional dependencies, our model enhances the interactions between emotional states across the conversation, enabling richer and more context-aware emotional representations. To this end, we propose a novel architecture, the *Long-range dependency emotionS Model (LYSM)*, which leverages a self-attention mechanism to explicitly model the dependencies among emotional states. By learning attention-based interactions between emotional state pairs across the dialogue, LYSM facilitates nuanced emotional reasoning and enhances the robust baseline model based on the COSMIC [15].

2.2 Related Works

Recent SOTA models in ERC have primarily focus on integrating prior commonsense knowledge [34, 15, 92, 3] or leveraging PLMs [60, 18] to construct utterance representations. Among them, our base model, COSMIC [15], stands out as one of the top-

performing frameworks, employing a commonsense knowledge extractor to capture additional features such as speaker intentions and reactions. However, emotional dependencies, which are a crucial aspect of ERC, have received limited attention in prior research. Only a few studies have explicitly addressed this issue, though recent models that incorporate emotional dependency structures have demonstrated substantial improvements in ERC performance [18, 30, 65]. These approaches typically apply a CRF layer, which is widely used in sequence labeling tasks in NLP. Among existing studies, EmotionFlow [65] is most closely related to our work. However, its modeling is restricted to emotional relations between adjacent utterances, thereby limiting its ability to model long-range affective dependencies. In contrast, our proposed model, LYSM, leverages a self-attention mechanism to capture dependencies among emotional states across the entire conversation, enabling the modeling of richer and more comprehensive emotional interactions. To demonstrate the effectiveness of our approach, we conduct comparative experiments against CRF-based models.

2.3 Methodology

In this section, we present the architecture of our proposed model, LYSM, which is built upon the COSMIC framework. To effectively capture the strong relationships among emotions in a conversation, we incorporate a self-attention mechanism [71] and compare its performance with that of a CRF, inspired by prior work [65] that adapted CRF on top of COSMIC.

The overall system comprises two main components: (1) a conversational encoding module, which transforms utterances in a conversation into hidden vector representations, and (2) an emotional dependency encoding module, which learns the influence of emotional relationships across the conversation.

2.3.1 Task definition.

Given a conversation consisting of a sequence of utterances $[(u_t, p_t)]_{t=1}^N$, where u_t denotes the t -th utterance and p_t its corresponding speaker, the objective is to identify the emotion label (y_t) for each utterance from a predefined set of emotion categories (e.g., *happy*, *sad*, etc.).

2.3.2 COSMIC Framework

COSMIC [15] aims to incorporate prior commonsense knowledge into the emotion recognition process. The knowledge features are extracted by COMET [3], an external tool built upon a commonsense knowledge graph. The main contribution part of COSMIC is its identification of several commonsense dimensions that are particularly relevant to conversational emotion recognition, including *the speaker’s intent* $\mathcal{IS}_{cs}(u_t)$, *the effect of speaker* $\mathcal{ES}_{cs}(u_t)$, *the reaction of speaker* $\mathcal{RS}_{cs}(u_t)$, *the effect of listeners* $\mathcal{EL}_{cs}(u_t)$, and *the reaction of listeners* $\mathcal{RL}_{cs}(u_t)$.

To obtain utterance-level sentence representations, COSMIC first fine-tunes a PLMs (e.g. RoBERTa) on an emotion classification task without considering the context. This fine-tuned model is then used to generate continuous vector representations for utterances. Following the conventional BERT-based architecture, a special classification token [CLS] is prepended to each input sentence to capture its overall meaning. The final representation for each utterance (\mathbf{x}_t) is derived by averaging the last four hidden layers corresponding to the [CLS] token.

To model the sequential features in the conversation, this framework employs GRU cells [8] to represent hidden states that influence the emotion of sentences. Five distinct types of states are constructed sequentially alongside the utterances in the conversation: *context state*, *internal state*, *external state*, *intent state*, and *emotion state*. These states are encoded by five separated GRU cells, namely GRU_C , GRU_Q , GRU_R , GRU_I , and GRU_E , respectively. For the mathematical formulation, the context vector (c_t) is first computed based on the current sentence vector (x_t), the previous internal state ($q_{s(u_t),t-1}$), and the

previous external state ($r_{s(u_t),t-1}$):

$$c_t = GRU_C(c_{t-1}, (x_t \oplus q_{s(u_t),t-1} \oplus r_{s(u_t),t-1})) \quad (2.1)$$

where \oplus denotes the concatenation operator. Subsequently, a soft attention vector (a_t) is introduced to update the internal and external hidden vectors:

$$u_i = \tanh(W_s c_i + b_s), \quad i \in [1, t-1] \quad (2.2)$$

$$\alpha_i = \sum_{j=1}^{t-1} \left(\frac{\exp(u_i^\top x_j)}{\sum_j \exp(u_j^\top x_j)} \right) c_j \quad (2.3)$$

where W_s, b_s are learnable parameters. The internal state $q_{s(u_t),t}$, external state $r_{s(u_t),t}$, intent state $i_{s(u_t),t}$, and emotion state e_t are then updated based on their previous values, incorporating commonsense knowledge and soft attention vectors:

$$q_{s(u_t),t} = GRU_Q(q_{s(u_t),t-1}, (a_t \oplus \mathcal{ES}_{cs}(u_t))) \quad (2.4)$$

$$r_{s(u_t),t} = GRU_R(r_{s(u_t),t-1}, (a_t \oplus \mathcal{RS}_{cs}(u_t))) \quad (2.5)$$

$$i_{s(u_t),t} = GRU_I(i_{s(u_t),t-1}, (\mathcal{IS}_{cs}(u_t) \oplus q_{s(u_t),t})) \quad (2.6)$$

$$e_t = GRU_E(e_{t-1}, x_t \oplus q_{s(u_t),t} \oplus r_{s(u_t),t} \oplus i_{s(u_t),t}) \quad (2.7)$$

In addition, the states of listeners are updated for each utterance (u_t) by replacing $\mathcal{ES}_{cs}(u_t)$ and $\mathcal{RS}_{cs}(u_t)$ in Equations 2.4 and 2.5, respectively, with $\mathcal{EL}_{cs}(u_t)$ and $\mathcal{RL}_{cs}(u_t)$. Then, the emotion label probabilities for the current utterance are calculated using a *softmax* layer applied to the emotion vector e_t :

$$p_t = \text{softmax}(W_e e_t + b_e) \quad (2.8)$$

where W_e, b_e are learnable parameters. Finally, the probabilities (p_t) for all sentences in the conversation are forwarded to compute the negative log-likelihood loss, and the model is trained using the backpropagation algorithm.

2.3.3 LYSM Architecture

To jointly encode the semantic representations of utterances along with their associated speaker’s emotional states to enhance the overall performance of ERC, we propose LYSM, a novel architecture designed to model emotional dependencies across utterances in a conversation.

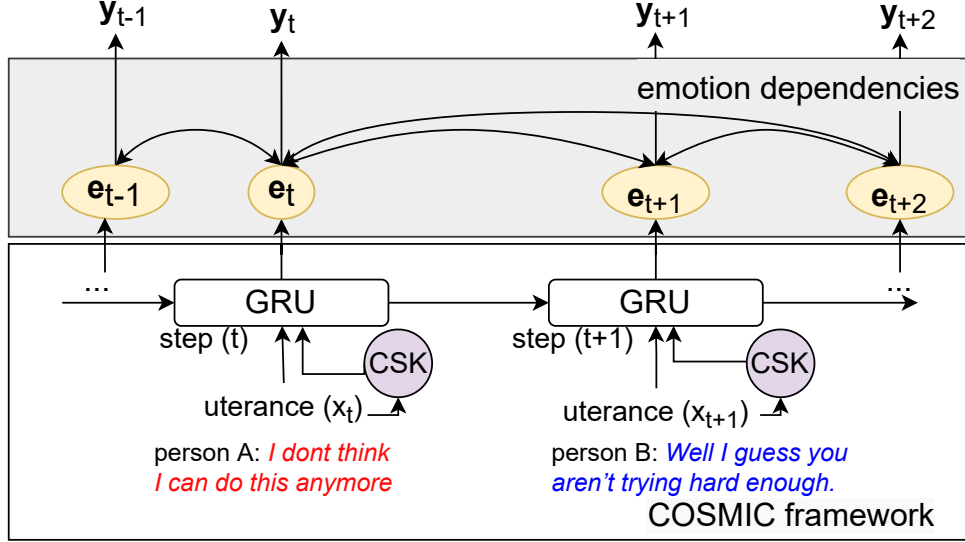


Figure 2.2: The architecture of the proposed *Long-range dependency emotionS Model (LYSM)*, built on top of the COSMIC.

As illustrated in Figure 2.2, LYSM extends the SOTA COSMIC framework [15], which serves as a strong baseline. COSMIC captures commonsense-informed features such as speaker intent and reaction through a dedicated commonsense knowledge module (denoted as *CSK* in the figure). While COSMIC models emotion dynamics through a recurrent structure (e.g., GRU), it lacks mechanisms to explicitly capture long-range emotional dependencies across distant utterances. We hypothesize that emotional states are strongly interdependent throughout the dialogue, the emotional tone set early in a conversation can significantly influence the emotions expressed in later utterances, even if they are temporally or speaker-distant. To address this limitation, LYSM retains COSMIC’s capability to generate commonsense-enhanced utterance representations while introducing a Transformer encoder layer [71] to model emotional dependencies (e_t) across the entire conversation.

Formally, the sequence of emotional states ($e = [e_t]_{t=1}^N$) obtained via Equation 2.7, is fed into the Transformer encoder.

$$\begin{aligned}
g_j^q, g_j^k, g_j^v &= eW_j^q, eW_j^k, eW_j^v, \quad j \in [1, \#heads] \\
head_j &= \text{softmax}\left(\frac{g_j^q \cdot (g_j^k)^\top}{\sqrt{d_h}}\right) g_j^v \\
g_{mul} &= (head_1 \oplus head_2 \oplus \dots \oplus head_{\#heads})W^o \\
g_{norm} &= \text{LayerNorm}(g_{mul} + e) \\
e' &= \text{LayerNorm}(\text{FFW}(g_{norm}) + g_{norm})
\end{aligned} \tag{2.9}$$

where $\#heads$ denotes the number of attention heads in the Multi-head layer, d_h is the dimensionality of each head, and LayerNorm and FeedForward (FFW) refer to the same operations as described in Transformer[71]. Finally, the updated emotional state vector $e' = [e'_t]_{t=1}^N$ compute the emotion label probabilities by a *softmax* layer, similar to Equation 2.8:

$$p_t = \text{softmax}(W_e e'_t + b_e) \tag{2.10}$$

This allows the model to generate contextualized emotion representations that incorporate long-range dependencies, enabling richer and more globally consistent emotional reasoning.

2.3.4 Conditional Random Field (CRF)

This architecture is commonly applied to sequence labeling tasks such as part-of-speech (POS) tagging and named entity recognition [44]. To model the dependencies between emotional states in conversations, previous works [65, 18] have incorporated a CRF layer

as the final component of the neural network.

$$\begin{aligned}\text{score}(e, y) &= \sum_{t=1}^N (W_{em}e_t + b_{em})[y_t] + \sum_{t=0}^N (W_{tr}[y_t, y_{t+1}]) \\ p(y|e) &= \frac{\exp(\text{score}(e, y))}{\sum_{y'} \exp(\text{score}(e, y'))}\end{aligned}\tag{2.11}$$

where y_0 and y_{N+1} represent special start and end emotional labels; $[\cdot]$ denotes the matrix selection operator based on row and column indexes; W_{em}, b_{em} and W_{tr} are the learnable weights for the emission and transition scores; and the variable y' is a candidate emotional label sequence drawn from the set of possible emotion flows. By applying the CRF layer, the model is trained to maximize the log-likelihood of gold emotion sequence.

Given its effectiveness in capturing short-range dependencies, we adapt this architecture to the COSMIC framework as a strong baseline for comparison with our proposed model, LYSM. Specifically, we treat the emotion vector representation (e_t) as the emission score for each utterance, while the transition scores, which model the dependencies between adjacent emotional states, are randomly initialized and subsequently learned in an end-to-end manner. However, unlike CRF-based approaches, which are limited to modeling local transitions between consecutive emotional states, LYSM leverages a self-attention mechanism to capture long-range emotional dependencies across the entire conversation. This enables a more global and context-aware understanding of emotional dynamics, which we demonstrate to be more effective through comparative experiments.

2.4 Experiment and Result Analysis

2.4.1 Experimental Setup

Since our proposed model is constructed built upon the COSMIC framework, we first conducted experiments using the outputs of COSMIC following these steps: fine-tuning the PLMs to obtain utterance representations and extract commonsense knowledge fea-

tures. The resulting continuous feature vectors, which are equivalent to the input features of the COSMIC framework, are then used as inputs to our LYSM architecture. These experiments aim to evaluate the effectiveness of the LYSM architecture in comparison with the original COSMIC framework. In addition, to assess our model’s performance in capturing emotional transference, we also conducted experiments incorporating a CRF layer into our architecture, allowing for a direct comparison with the EmotionFlow model [65].

For each dataset mentioned above, we run the experiments ten times using different random seeds and evaluate the performance of our proposed model using the Weighted Average F1 score [15]. The best-performing model on the development set is selected for evaluation on the test set. We then report the mean performance across runs and compare it with the results reported in previous work on the same datasets.

2.4.2 Main Results

We conducted experiments on four aforementioned datasets, and the results are presented in Table 7.1. Our proposed LYSM model consistently improved the performance of the COSMIC framework across all datasets. Specifically, on the IEMOCAP dataset, LYSM improved a 0.19 increase in F1 score over the original COSMIC model. On the DailyDialog dataset, our model improved Macro F1 score by 0.27 and Micro F1 score by 0.21, respectively. For the EmoryNLP and MELD datasets, we conducted experiments under the standard seven-class emotion recognition setting, where LYSM yielded F1 improvements of 0.23 and 0.19, respectively. These results demonstrate that our LYSM architecture generalizes well across datasets and the emotion dependency modeling component is effective when incorporated into the COSMIC framework.

Compared to EmotionFlow, which models emotion transitions using a CRF layer, our approach leverages a self-attention mechanism to capture long-range emotional dependencies across the entire conversation. This enables a more comprehensive and context-aware understanding of emotional dynamics. For a fair comparison between the CRF

Table 2.1: Performance comparison between methods.

Methods	IEMOCAP	DailyDialog		MELD	EmoryNLP
	W-Avg F1	Macro F1	Micro F1	W-Avg F1	W-Avg F1
DialogueRNN re-product [16]	62.57	41.80	55.95	57.03	31.70
EmoBerta [28]	67.42	-	-	65.61	-
EmotionFlow [65]	65.05	-	-	-	-
COSMIC [15]	65.28	51.05	58.48	65.21	38.11
LYSM (ours)	65.47	51.32	58.69	65.40	38.34

layer and our proposed self-attention mechanism, we also conducted ablation studies, as detailed in Section 2.4.3. Additionally, while our LYSM model achieved competitive performance with EmoBERTa on the MELD dataset, it showed lower performance on IEMOCAP. We attribute this gap primarily to model scale: EmoBERTa fine-tunes RoBERTa-large [42] with 355 million parameters and a training time of 9 minutes per epoch, whereas our LYSM model employs a fixed, fine-tuned RoBERTa-base encoder with only 17 million trainable parameters, requiring just 30 seconds per epoch. This highlights LYSM’s efficiency–performance trade-off and its potential for lightweight, real-time applications.

2.4.3 Ablation Study

CRF Layer In this experiment, we evaluate the effectiveness of the emotional dependencies modeling component. While our LYSM architecture employs a self-attention mechanism, previous works [65, 18] have proposed using a CRF layer to model emotion transfer between sequences of utterances in a conversation. To enable a fair comparison we conducted additional experiments by incorporating a CRF layer (+CRF) into both the COSMIC and our proposed model, as shown in Table 2.2.

Although we followed the same experimental setup as COSMIC[15], minor discrepancies in the reproduced its results can be attributed to differences in experimental environments, such as library versions or computing infrastructure. Consistent with prior findings, our results confirmed that integrating a CRF layer into COSMIC yields an improvement of 0.32 in F1 score. However, our LYSM, which is based on a self-attention mechanism,

Table 2.2: Ablation study on the IEMOCAP dataset.

Methods	IEMOCAP
EmotionFlow [65]	65.05
COSMIC	64.50
COSMIC +CRF	64.82
LYSM	65.47
LYSM +CRF	65.43

achieves a significantly larger performance gain, with an improvement of 0.97 in F1 score. We argue that this advantage arises from the self-attention mechanism’s ability to model emotional dependencies between both adjacent and non-adjacent utterances, whereas the CRF layer is inherently limited to modeling local transitions. In addition, we applied a CRF layer on top of the LYSM architecture. However, this did not lead to additional performance improvements, likely because the emotional dependency information had already been effectively captured by the self-attention mechanism. These findings demonstrate the effectiveness of the proposed LYSM architecture and highlight the critical role of long-range emotional dependencies in conversational emotion recognition.

Conversation Length. We conducted an analytical experiment to examine the impact of conversation length on the performance of the ERC system (Figure 2.3). The results indicate that the performance of both the COSMIC framework and our proposed LYSM model tends to decline as the conversation length increases. Nevertheless, LYSM consistently outperforms the baseline model across all conversation length groups, demonstrating the generalizability of our approach. Notably, the performance gap becomes more prominent in longer conversations, providing further evidence of LYSM’s effectiveness in capturing long-range emotional dependencies.

2.4.4 Improvement Example.

Based on our analysis of prediction outputs on the IEMOCAP dataset, we observed that LYSM architecture generally performs more effectively in conversations involving strong emotional expressions. To better illustrate the improvements brought by the LYSM

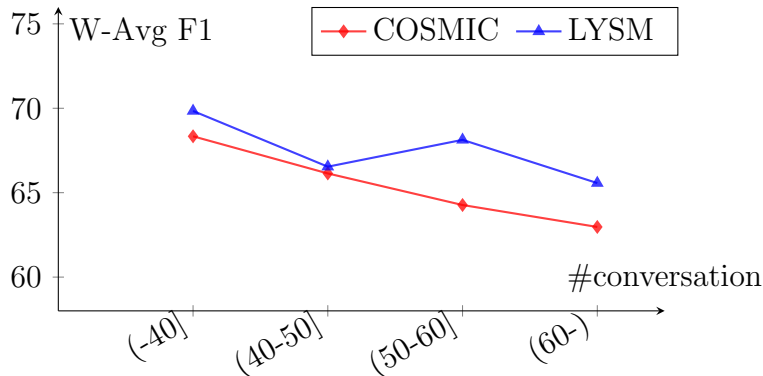


Figure 2.3: Performance comparison between LYSM and COSMIC on the IEMOCAP, stratified by the number of utterances per conversation (#conversation).

Table 2.3: Improvement example collected in IEMOCAP dataset. The green and red labels indicate the correct and incorrect model predictions, respectively.

Id	Utterance	Label	LYSM	COSMIC
u23	S1: You infuriate me sometimes. Do you know that? God.	angry	angry	frustrated
u24	S1: Isn't it your business, too, if dad – if I tell dad and he throws a fit about it? I mean, you have such a talent for ignoring things.	angry	angry	frustrated
u25	S2: I ignore what I got to ignore. I mean, the girl is Larry's girl.	angry	frustrated	frustrated
u26	S1: She is not Larry's girl!	angry	angry	frustrated
u27	S2: From your father's point of view he's not dead and she's still his girl. Now, you can go on from there if you know where to go, Chris, but I don't know. So what can I do for you?	angry	frustrated	frustrated
u28	S1: I don't know why it is but everytime I reach out for something I- that I want, I have to pull back because I might hurt somebody else. My whole bloody life; time after time after time.	frustrated	frustrated	angry
u29	S2: Well, you're a considerate fella, there's nothing wrong in that	neutral	neutral	neutral
u30	S1: To hell with that!	angry	angry	angry

architecture, we present representative examples in Table 5.8, highlighting differences in predictions between LYSM and the COSMIC model. In the selected conversation, two speakers are engaged in a negatively charged exchange, with many utterances labeled as **angry**. We found that for sentences containing strongly emotional words (e.g., *hell* in utterance u30), both LYSM and COSMIC produced correct predictions. However, for utterances with more neutral or ambiguous wording (e.g., u23, u24, u26), the COSMIC framework tended to predict milder negative emotions such as **frustrated**. In contrast, LYSM, benefiting from its stronger modeling of contextual emotional dependencies, was able to assign more accurate sentiment labels in line with the overall emotional trajectory of the conversation.

Learning Process. We reproduced the COSMIC framework and conducted experiments to compare its performance with our LYSM architecture on the IEMOCAP dataset, focusing on weighted average F1 scores (Figure 2.4). The results show that the weighted

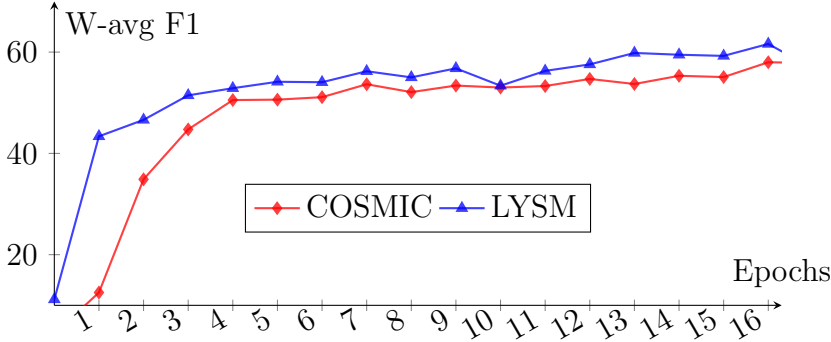


Figure 2.4: Comparison of weighted average F1 scores between COSMIC and LYSM on the IEMOCAP dev set.

average F1 scores of our proposed architecture on the development set consistently exceed those of the COSMIC framework across most training epochs. These findings demonstrate the robustness and effectiveness of the LYSM architecture, highlighting its superior generalization capability over the baseline model.

2.4.5 Emotional Dependency.

In our LYSM architecture, we aim to model the dependencies between pairs of emotional state across utterances to improve the performance of sentiment analysis system. To investigate how these dependencies manifest in practice, we visualize the emotional state dependencies constructed between utterances within a conversation (Figure 2.5). This figure shows the scaled self-attention in the LYSM architecture, computed in Equation 2.9. The title of each row or column in the heatmap represents an utterance (u_i) in the conversation. Darker colors indicate higher attention scores. The analysis reveals that emotional dependencies influence not only adjacent utterances but also distant ones. For example, the emotion *happy* in utterance 34 (column u34) is affected by contextual cues spanning the entire dialogue, including those from the conversation’s beginning.

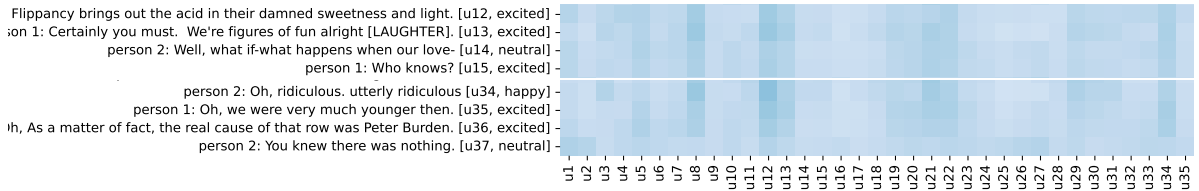


Figure 2.5: Heatmap visualization of dependencies between pairs of emotional states in a conversation.

Moreover, emotionally distinct utterances that differ from the surrounding emotional context tend to receive more attention from the model. Notable examples include utterances u12, u13, and u34. These observations provide strong evidence that capturing emotional dependencies, particularly long-range ones, is critical for building effective emotion recognition systems.

2.5 Chapter Summary

In this work, we investigated the role of emotional dependency features in conversation-level emotion recognition tasks. To this end, we proposed LYSM, an effective model that integrates a self-attention mechanism into the COSMIC framework, enhancing performance and achieving competitive results across four benchmark datasets: IEMOCAP,

DailyDialog, EmoryNLP, and MELD. The proposed model is simple yet powerful, and its modular design allows for easy integration into other architectures within the sentiment recognition domain. For future work, we plan to extend the self-attention mechanism to explicitly model emotional dependencies for individual speakers in multi-party conversations. Additionally, we aim to explore the influence of speaker personality traits on sentiment dynamics, with the goal of further improving the contextual understanding of emotions in dialogue systems.

Chapter 3

Utterance Modeling

3.1 Motivation and Objectives

Understanding the meaning of a sentence in a conversation mostly relies not on its isolated semantics, but on the surrounding conversational context. Therefore, one of the fundamental challenges in ERC lies in representing the target utterance given its conversational context.

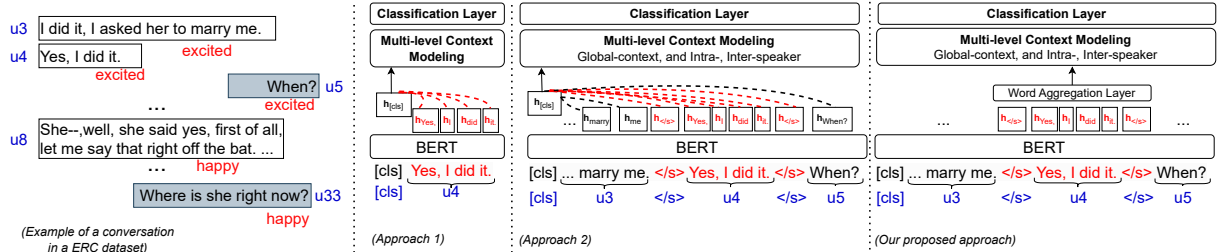


Figure 3.1: An example of the ERC task and a comparison of our approach with previous BERT-based methods.

Owing to the powerful context modeling capabilities of PLMs such as RoBERTa, prior research has commonly adopted two major strategies to incorporate contextual information into utterance representations, as illustrated in Figure 3.1: (1) Encoding each utterance individually using a PLM, followed by the use of additional components to inject contextual information into the representation of the focal utterance. This approach enhances the semantic clarity and completeness of the utterance vector by integrating

surrounding information in post-processing stages [34, 63, 43, 15]. (2) Concatenating the target utterance with its local context into a single input sequence, which is then fed into a PLM to obtain a more contextually enriched representation [28, 33].

While the first approach partially benefits to some extent from the context modeling ability of PLMs, it fails to fully exploit their potential, as each utterance is encoded in isolation prior to the contextual integration. In contrast, the second approach leverages the full contextualization capability of PLMs more effectively. However, existing methods under this paradigm typically append special tokens (e.g., [CLS]) or self-attention mechanisms to mark the location of focal sentence [28, 33]. Such design requires the model to implicitly learn to focus on the focal sentence rather than on the auxiliary tokens, making it difficult to explicitly verify whether the attention mechanism truly captures the intended semantic focus.

To address these limitations, we propose a novel sentence modeling approach named AccWR, which directly integrates word-level information from the target utterance while incorporating surrounding utterances as contextual input as shown in Figure 3.1. Specifically, we concatenate the local context with the focal utterance and feed this sequence into a PLM (e.g., RoBERTa). Instead of relying on special tokens to extract the focal utterance representation, we explicitly aggregate the hidden states of its constituent words using a MLP to form the final sentence vector. This design offers two key advantages: (1) it retains the full contextual modeling capacity of PLMs by presenting them with coherent local conversational input, and (2) it enables precise focus on the word-level semantics of the target utterance, thereby avoiding the ambiguity and potential masking effects introduced by special token-based methods. Consequently, our model facilitates more effective and interpretable representation learning for the focal utterance in context.

3.2 Related Works

For modeling the conversational context, RNNs [20] have been widely adopted to capture the sequential structure and interactions among utterances in a dialogue [19, 54, 47]. In

parallel, graph convolution network [11] have been utilized to uncover the latent structural relationships between utterances [30]. With the advent of the self-attention mechanism [41], a variety of attention-based methods have been proposed to infer the utterance-level emotions by modeling implicit dependencies across dialogue turns, such as DialogueEIN [43], CoMPM [31], MuCDN [90], and DAG [63].

Furthermore, the remarkable ability of PLMs, such as RoBERTa, to capture nuanced contextual semantics has led to their widespread adoption in recent ERC studies [14]. By leveraging large-scale pre-training on diverse textual corpora, PLMs offer powerful representations that effectively encode both local utterance semantics and broader conversational context. Recent works [14] have built upon PLMs to incorporate not only surrounding utterances but also external commonsense knowledge into the utterance-level representations [63, 43, 15, 36].

In contrast to prior approaches, our method integrates local contextual information by concatenating surrounding utterances [28, 33], while simultaneously leveraging global context and speaker-related dependencies following the architecture of DialogueEIN [43]. We propose a novel representation strategy that accumulates word-level embeddings across the entire utterance, instead of relying solely on the CLS token. Building upon recent SOTA methods, our framework utilizes a PLM and introduces an effectiveness method to enhance utterance representations through fine-grained word-level accumulation within the broader conversational context.

3.3 Methodology

To enhance sentence modeling in the ERC task, we propose a sentence modeling mechanism grounded in on word-level representations. Unlike previous work that primarily rely on sentence-level tokens (e.g., [CLS]) or global sequence embeddings, our method explicitly emphasizes the contribution of individual word vectors within the target sentence. Specifically, we concatenate surrounding utterances as local context and input the entire sequence into a PLM model such as RoBERTa. The hidden states corresponding to each

word in the focal utterance are then extracted and aggregated via a MLP to obtain the final utterance representation (Figure 3.1). This word-level strategy allows the PLM to model nuanced contextual interactions while maintaining focus on the focal utterance, thereby yielding more precise and semantically rich representations. It also avoids the limitations of approaches that rely on special tokens or implicit attention mechanisms to identify the target utterance within a concatenated sequence. Furthermore, inspired by recent SOTA methods DialogueEIN [43], our framework integrates multi-level contextual information, including both local and global dialogue context, as well as speaker-related dependencies, further enhancing the model’s capacity for accurate emotion recognition.

3.3.1 Baseline

We implemented our proposed method by extending the recent SOTA DialogueEIN framework [43]. The baseline DialogueEIN consists of three principal components: utterance representation, multi-level context modeling (including *semantic* and *emotional interaction networks* [43]), and a classification layer:

- *Utterance Representation.* This component transforms a raw conversation (C), containing of several utterances, from textual form into hidden vector representations. Following the BERT architecture [14], each utterance (s_i) is concatenated with a [CLS] token and passed through a PLM. Especially, the input can be expanded to include local context by concatenating a fixed window of surrounding utterances [28, 31, 36]. In this setting, s_i is replaced by a concatenated sequence of nearby utterances, formatted as: [CLS], s_{i-w} , s_{i-w+1} ..., $</s>$, s_i , $</s>$, ... s_{i+w} , where w denotes the local contextual window size (e.g., $w = 2$). The hidden representation of the [CLS] token is then used as the final utterance representation u_i , where i is the index of the utterance in the conversation, $0 \leq i < |C|$.

$$h_{[CLS]}, h_{words} = \text{RoBERTa}(s_i) \quad (3.1)$$

$$u_i = h_{[CLS]} \quad (3.2)$$

Here, h_{words} denotes the hidden states of all words¹ in the text input. The resulting utterance representations are then passed to the subsequent component for contextual modeling.

- *Multi-level Context Modeling.* This module captures various dependencies among utterances in a conversation. Given the sentence-level representations (u_i), multiple types of relational dependencies are modeled using the attention mechanism [71]. These include *global-context*, *inter-speaker*, and *intra-speaker* relations [43].

$$q_h, k_h, v_h = u_i W_h^q, u_i W_h^k, u_i W_h^v \quad (3.3)$$

$$head_h = \text{Attention}(q_h, k_h, v_h, mask) \quad (3.4)$$

$$h_{\text{Multi-head}} = \text{concat}([head_h]_{0 < h \leq H}) W^o \quad (3.5)$$

where W^* are learnable parameters, H is the number of heads in the multi-head attention layer, each head contains a query, key, and value vectors with dimension d_h , a relational-specific *mask* controls which utterance pairs are considered for attention computation. Attention scores are derived using the scaled dot-product mechanism:

$$\text{Attention}(q, k, v, mask) = \text{softmax}\left(\frac{q \cdot k^T}{\sqrt{d_h}} + mask\right) \cdot v$$

Following DialogueEIN [43], we define distinct *mask* for each relational type, such as global context mask (connecting all utterances), an intra-speaker mask (connecting utterances from the same speaker) and an inter-speaker mask (connecting utterances from different speakers). This process yields three context-aware vectors (from Equation 4.6) h_{global} , h_{intra} , h_{inter} to model the different conversational dependencies, which are subsequently forwarded to the classification layer.

- *Classification.* This component aggregates the original utterance representation

¹Subword distinctions are omitted for simplicity, without loss of generality.

with the context-aware vectors to predict the emotion label:

$$\begin{aligned} h_{out} &= u_i W^u + h_{global} W^g + h_{intra} W^i + h_{inter} W^t \\ e_{out} &= \text{softmax}(h_{out}) \end{aligned} \quad (3.6)$$

where W^* are learnable weight matrices. The output vector (e_{out}) represents the predicted emotion distribution, which is used to compute the training loss via the *Cross-Entropy* function.

3.3.2 Our Proposed Model (AccWR).

In this work, we propose a novel approach that emphasizes the accumulation of word-level vector representations, a perspective that has not been thoroughly explored in prior research. Traditional ERC models, such as DialogueEIN [43] (Approach 1 in Figure 3.1), each utterance is encoded independently, thus failing to fully exploit the contextual modeling capabilities of PLMs. This results in limited interaction between the focal utterance and its conversational context during the representation learning phase. To address this, Some earlier works [28, 33, 36] local context integration by concatenating surrounding utterances (within a fixed-size window) with the target utterance into a single input sequence, placing a special [CLS] token, and subsequently relying on self-attention to differentiate between the target utterance and its context (Approach 2 in Figure 3.1). These models then rely on the PLM’s self-attention mechanism to learn context-aware representations, using the hidden state of the [CLS] token as the final utterance vector. However, these methods absorbed all contextual information into a single global token, neglecting the contribution of individual words.

To overcome this limitation, our proposed method focuses on the explicit aggregation of word-level hidden representations extracted by PLMs, termed *AccWR*. Our mechanism fully leverages the representational capacity of PLMs by integrating contextual information at the word level. Specifically, we investigate the following strategies for aggregating word-level representations into an utterance-level vector:

- *Average pooling.* This is a simple yet effective strategy where we compute the average of the hidden states corresponding to all words in the considering sentence. The utterance vector in Equation 6.4 is replaced by:

$$u_i = \text{average}(h_{words}^{s_i}) \quad (3.7)$$

where $h_{words}^{s_i}$ denotes the set of hidden vectors corresponding to all word in the currently considered sentence (s_i).

- *Bidirectional LSTM.* To better capture the sequential structure, we apply a Bi-LSTM network [20] over the word embeddings. The output hidden states are then averaged to form the utterance vector:

$$u_i = \text{average}(\text{BiLSTM}(h_{words}^{s_i})) \quad (3.8)$$

- *Multilayer Perceptron.* To introduce non-linearity and allow the model to learn task-specific transformations of the averaged word vectors, We employ an MLP with the *tanh* activation function.

$$u_i = \tanh(\text{average}(h_{words}^{s_i}) \cdot W^s) \quad (3.9)$$

where W^s is a learnable parameter matrix. This enables a more flexible mapping from word-level semantics to the utterance-level emotion representation space.

3.4 Experiment and Result Analysis

3.4.1 Experimental Setup.

For each of the aforementioned datasets, we conducted ten independent runs using distinct random seeds to ensure the robustness of our performance evaluation. Following prior works [15, 43], we report the Weighted Average F1 score [15] for IEMOCAP,

EmoryNLP, MELD datasets, and the Micro F1 score for the DailyDialog dataset. Subsequently, we present the mean performance across the ten runs and compare it against existing SOTA methods.

In the training process, we used the PLM `roberta-large`² for all experiments. To reduce computational cost and prevent overfitting, we froze the BERT embedding layer and first ten Transformer layers of RoBERTa, updating only the final two layers during training. The *Cross-Entropy* loss function was employed with class weights to deal with the imbalance in emotional label classes.

Regarding hyperparameter settings, we set the learning rate to $1e^{-5}$ for the IEMOCAP dataset and $5e^{-6}$ for the remaining datasets. The dropout rate was fixed at 0.2, and the number of training epochs was set to 30. The local context window size (w) was selected in $\{2, 4\}$. The model achieving the best performance on the development set was chosen for final evaluation on the test set.

3.4.2 Main Results

Our proposed method demonstrates competitive performance compared to recent SOTA approaches (Table 7.1). Among the methods that do not incorporate external knowledge bases or additional auxiliary data, our model achieves the highest performance on the EmoryNLP dataset and performs comparably to the best-performing models on MELD. While our method adopts the multi-level context modeling framework introduced in DialogueEIN [43], it differs significantly in how it models local context, particularly through the accumulation of word-level representations. We attribute the performance gap between our model and DialogueEIN on the IEMOCAP dataset primarily to differences in implementation details and library environments, rather than fundamental architectural disparities. Notably, results marked with an asterisk (*) indicate the use of external knowledge bases, which offer additional contextual information that is not utilized in our approach. Despite the absence of such external knowledge, our method still yields

²<https://huggingface.co/roberta-large>

promising results, performing competitively with the SOTA method, such as SKIER [36].

Moreover, we observe that the average pooling strategy for word-level representations achieves performance comparable to that of the MLP-based approach across most datasets. This indicates that the primary performance gains are attributed to our proposed word accumulation mechanism, rather than to the complexity of the aggregation function itself.

Table 3.1: Performance comparison between our proposed method and existing SOTA approaches.

Methods	IEMOCAP W-F1	EmoryNLP W-F1	MELD W-F1	DailyDialog MicroF1
HiTrans [33]	64.50	36.75	61.94	-
SKAIG [34]	66.96	38.88	65.18	59.75
DAG [63]	68.03	39.02	63.65	59.33
DialogXL [62]	65.94	34.73	62.14	54.93
DialogueEIN [43]	68.93	38.92	65.37	62.58
COSMIC [15]*	65.28	38.11	65.21	58.48
AccWR (ours)				
- using MLP	67.65	39.33	64.58	59.22
- using AVERAGE	67.78	38.21	64.99	59.40
- using ALL WORDS (MLP)	66.18	39.06	65.18	59.75

3.4.3 Ablation Study

We presented the performance distribution over ten independent runs in the box plot (Figure 4.1) to compare our proposed word-level aggregation methods with the baseline model that employs the [CLS] token for utterance representation. All aggregation-based methods consistently outperform the [CLS]-based baseline, demonstrating the effectiveness of explicitly accumulating word-level representations for contextualized sentence modeling. Among them, the MLP-based methods achieves the highest performance, yielding improvements of 0.45 F1 over the average pooling method, 0.75 F1 over the BiLSTM method, and a substantial gain of 2.53 F1 over the baseline. Our approach leverages the contextual capacity of pre-trained language models while preserving a clear focus on the target utterance, avoiding the potential noise introduced by concatenated contextual inputs.

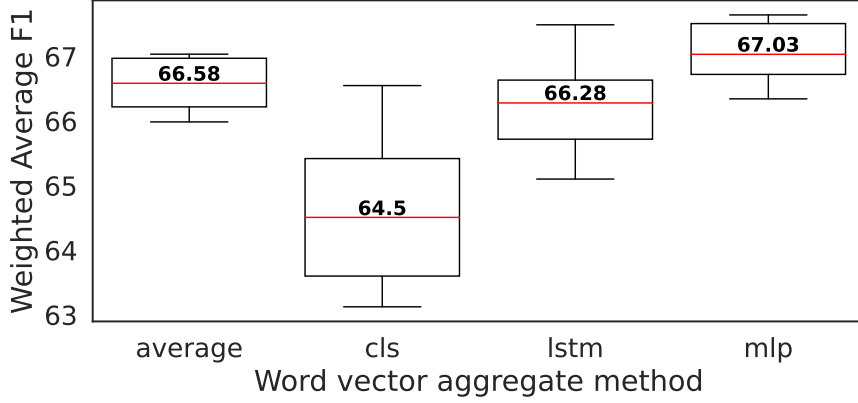


Figure 3.2: Ablation study on the IEMOCAP development set evaluating different word vector aggregation methods. The red bar indicates the median performance for each configuration.

3.4.4 Convergence Time

Figure 3.3 illustrates the development performance on the IEMOCAP dataset, comparing different word-level aggregation strategies. Our proposed mechanism demonstrates

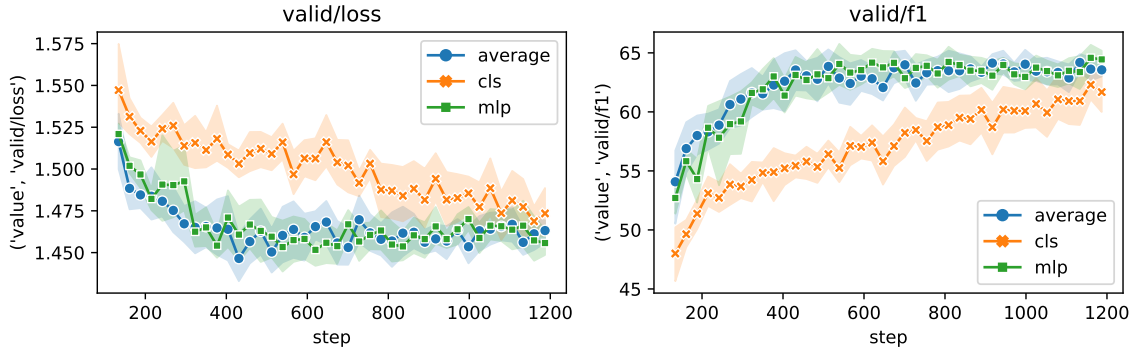


Figure 3.3: Loss and performance comparison between our proposed method and the ([CLS]) baseline on the IEMOCAP development set. The shaded area around the line graph represents the variability across ten different runs.

consistently superior performance over the baseline across both the *Average* and *MLP* settings. In addition, it facilitates faster convergence, with notable performance stabilization occurring between steps 400 and 600 (Figure 3.3). We attribute this efficiency to the fact that, unlike the setting of baseline approach, which relies on the [CLS] token and thus requires additional learning to align the attention mechanism with the focal sentence. Our proposed mechanism, directly aggregates word-level representations, enabling the model to utilize its learning capacity more effectively from the outset.

3.5 Chapter Summary

In this work, we proposed an efficient mechanism for accumulating word representation of an utterance in emotion recognition in conversation. Experimental results on four widely used benchmark datasets demonstrate that our method not only outperforms previous approaches but also achieves performance competitive with recent state-of-the-art models. Besides, the proposed mechanism shows strong potential for broader applicability in tasks that require context-aware sentence representation, such as conversational response generation, multi-turn question answering, and dialogue-based semantic parsing.

Chapter 4

Speaker Modeling

4.1 Background and Motivation

In ERC, the input text typically consists of spoken dialogues transcribed by automatic speech recognition systems. These transcripts often contain colloquial expressions, incomplete sentences, and non-standard grammar, all of which introduce additional complexity to emotion identification within dialogue. To address these challenges, recent approaches have sought to enrich utterance representations by incorporating acoustic features [23, 64] or visual knowledge [1] to enrich utterance vector representation. However, in purely text-based ERC tasks where such multimodal cues are unavailable, accurately identifying speakers’ emotions becomes significantly more challenging.

Prior research has demonstrated that personality traits strongly influence an individual’s emotional expression and habitual word choices during conversation [47, 31, 2, 56, 83, 17, 63, 26]. Consequently, even identical utterances may convey distinct emotional meanings depending on the speaker, the distinctiveness of a speaker’s personality, habitual speech patterns, vocabulary choices, and modes of expression can lead to distinct emotional nuances and implied meanings [57]. For instance, a sarcastic individual might deliver praise with an ironic tone, an emotionally expressive person may exhibit heightened intensity in responses, while a composed speaker tends to provide more neutral feedback. Therefore, speaker modeling has emerged as a key direction in recent ERC research. Ear-

lier studies have employed GRUs [47, 31, 2], GNNs [56, 83], and self-attention mechanisms [17, 63, 26] to encode rich speaker-specific information, including intra- (within a speaker) and inter-speaker (between speakers) features.

Unfortunately, in previous research, speaker-specific information has been learned implicitly during the training process, posing challenges in validating its correctness and adapting models to diverse domains. Explicit information, such as speaker characteristics, has not received sufficient attention due to the scarcity of annotated personality data and the high costs associated with manual labeling, despite the critical role of personality-related knowledge in understanding conversational emotions. This limitation hinders the integration of personality-related knowledge into ERC models.

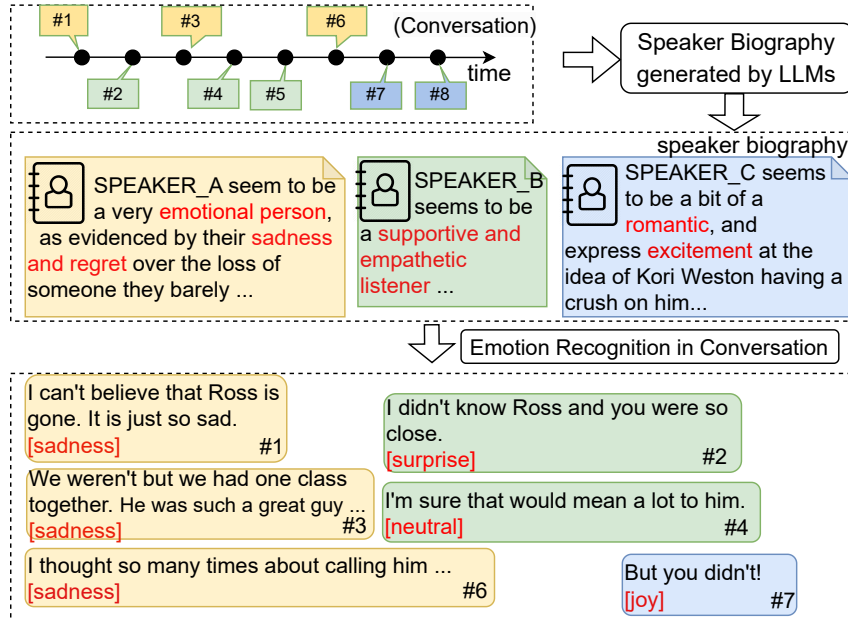


Figure 4.1: Overview of the Proposed BiosERC framework

To tackle the aforementioned limitations, we propose BiosERC, a novel framework designed to explicitly extract and leverage speaker personality traits to enhance ERC systems. In contrast to previous methodologies [31, 2, 17, 63, 26], BiosERC is capable of accurately identifying individual speaker personalities within dialogue (Figure 4.1). Specifically, BiosERC utilizes LLMs with a prompting technique [75, 66] to extract descriptions of interlocutor features as supplementary knowledge, which are then injected into the emotion recognition process within conversations. As illustrated in Figure 4.1,

the conversation features three distinct speakers, each contributing differing viewpoints and displaying clearly differentiated emotional expressions. The incorporation of speaker descriptions enables the model to more effectively grasp each individual’s communicative role within the dialogue. For instance, SPEAKER A is portrayed as experiencing a mix of *sadness and regret*, which is reflected in consistently sorrowful utterances. In contrast, adopts a more passive and *empathetic* stance, offering emotional support primarily through reactions to SPEAKER A, rather than initiating new content. SPEAKER C responds with notable *excitement* upon overhearing the interaction, introducing a sharp emotional contrast.

Integrating biographical information substantially enriches the emotional context associated with each speaker, offering valuable prior knowledge that enhances the model’s ability to perform nuanced and context-aware emotion recognition—particularly in emotionally complex dialogues. This uniqueness empowers the system to better capture personal characteristics and habitual expression patterns, enabling a deeper understanding of emotional transitions grounded in speaker-specific traits. Furthermore, our approach to capturing speaker characteristics is both interpretable and easily adjustable, making it well-suited for broader applications in conversational analysis beyond ERC.

4.2 Related Works

Given the substantial influence of speaker identity on ERC performance, researchers have placed emphasis on speaker modeling. Models such as DialogueRNN [47] and COSMIC [15] leverage RNN to capture speaker-specific semantic context. Some researches have conceptualized conversations as graphs, incorporating speaker information either through predefined relationships between utterances [56, 83] or by modeling speakers directly as nodes within the graph structure. HiTrans exploits an auxiliary task that classifies whether two utterances belong to the same speaker to make the model speaker-sensitive. S+PAGE [39] employs a two-stream conversation Transformer architecture to extract both intra and inter-speaker contextual features.

While most existing studies have focused on modeling individual utterances or interactions between speakers to particularly extract intra- and inter-speaker relationships to derive speaker-related information [43, 26]. Regrettably, the exploration of speaker-specific characteristics remains relatively underdeveloped. These characteristics, however, are essential components of conversational understanding, as they offer a foundational layer for interpreting speaker behaviors and emotional expressions. To bridge this gap, we propose a novel method named BiosERC, a novel approach that leverages external tools to explicitly extract speaker characteristics and integrate them into the ERC process, thereby enriching the model’s ability to interpret emotional nuances grounded in individual speaker profiles.

4.3 Methodology

This section presents the architecture of our baseline model for the ERC task, which incorporates both intra- and inter-speaker information in line with current state-of-the-art approaches [39, 43, 2]. We also introduce our proposed method, BiosERC, which enriches the model by incorporating speakers biography information to enhance emotional inference.

Formally, we define a conversation as: $\mathcal{C} = \{u_i\}_{0 \leq i < |\mathcal{C}|}$, where each individual utterance u_i is articulated by a speaker $p(u_i) \in \mathcal{S}$, with $\mathcal{S} = \{s_j\}_{0 \leq j < |\mathcal{S}|}$ denotes the set of all speakers involved in the conversation, and p is a mapping function that assigns each utterance to its corresponding speaker.

4.3.1 Intra-inter ERC (baseline)

Following prior SOTA frameworks in ERC [39, 2, 26], our baseline model consists of three core components: utterance vector representation, context modeling, and an emotion classification layer.

Utterance Vector Representation. To enhance the semantic richness of individual utterances, we adopt a context-aware representation approach by incorporating surround-

ing utterances within a fixed-size local window [28, 31, 36, 26]. Particularly, for encoding an utterance u_i , the input sequence is constructed as: “[CLS], $u_{i-w}, \dots, </s>, u_i, </s>, \dots, u_{i+w}$ ”, where w is the size of the local contextual window. The vector representation of u_i is then derived by aggregating the corresponding word-level embeddings, as described in [26]:

$$h^{cls}, h^{words} = \text{RoBERTa}([u_{i-w}, \dots, u_{i+w}]) \quad (4.1)$$

$$h^{utt} = [\tanh(\text{average}(h^{\text{words of } u_i}) \cdot W^u)]_{0 \leq i < |C|} \quad (4.2)$$

where $h^{\text{words of } u_i}$ represents the word vectors extracted from h^{words} at the positions corresponding to u_i ; h^{utt} refers to all utterance vectors in the conversation, and W^* is a learnable weights.

Context Modeling. To capture the contextual dependencies within a conversation, utterance vectors are integrated contextual information of whole conversation by attention mechanism:

$$\text{Attn}(q, k, v, M) = \text{softmax}\left(\frac{q \cdot k^\top}{\sqrt{d_t}} + M\right) \cdot v \quad (4.3)$$

$$q_t, k_t, v_t = h^{utt} W_t^q, h^{utt} W_t^k, h^{utt} W_t^v \quad (4.4)$$

$$head_t = \text{Attn}(q_t, k_t, v_t, M) \quad (4.5)$$

$$h_{\text{MultiHead}} = \text{concat}([head_t]_{0 \leq t \leq H}) W^o \quad (4.6)$$

where H is the number of attention heads; q_t, k_t, v_t represent the query, key, and value vectors corresponding to the utterances in different semantic subspaces, each with dimensionality d_t .

In detail, following the approach of DialogueEIN [43, 26], we construct relation matrices M to explicitly guide the attention mechanism in modeling interactions among utterances. Where $M_{ik} = 0$ if u_i and u_k should have interaction, $M_{ik} = -\infty$ if otherwise. For our baseline model, we implement three types of relation matrices to capture different contextual relationships:

- *global context*: all utterance pairs are connected, modeling full conversational flow.
- *intra-speaker context*: only utterances from the same speaker are connected, capturing individual speaker consistency.
- *inter-speaker context* only utterances between different speakers are connected, focusing on dialogue interaction and response dynamics.

Each attention variant yields a contextualized hidden representation: h^{contxt} , h^{intra} , and h^{inter} . These are subsequently passed into the classification component to infer the emotional states.

Classification. The classification component aims to integrate all the hidden features derived from the utterances and classify the corresponding emotion label. The speaker-specific hidden representation $h_i^{speaker}$ is computed as a weighted sum of the intra- and inter-speaker context vectors:

$$h_i^{speaker} = h_i^{intra}W^a + h_i^{inter}W^r \quad (4.7)$$

$$e_i^o = \text{softmax}(h_i^{utt}W^u + h_i^{contxt}W^g + h_i^{speaker}) \quad (4.8)$$

Subsequently, the predicted emotion vector e_i^o is evaluated using the *cross-entropy* loss against the ground-truth label of the i -th utterance, guiding the optimization of the model during training.

4.3.2 BiosERC: Integrating Biography Speakers Supported by LLMs for ERC Tasks

In this section, we outline the procedure for generating speaker biographies and introduce our BiosERC framework. The framework is built upon two widely adopted PLM models: a BERT-based encoder model (e.g., RoBERTa), and a decoder-only transformer-based LLM (e.g., LLaMA-2) [66]. Notably, we also introduce an effective strategy that incorporates speaker biographies by fine-tuning an LLM using a prompting technique. [58].

Biography of Speaker We introduce a prompting-based mechanism that enables LLMs to generate speaker-specific descriptions (d_j) for each speaker (u_j) in a given conversation \mathcal{C} . The output of this step is a set of speaker biographies for all participants in the conversation, denoted as $\mathcal{B} = \{d_j\}_{0 \leq j < |\mathcal{S}|}$.

$$d_j = \text{LLMs}(\text{prompting}(\mathcal{C}, s_j)) \quad (4.9)$$

Here, *LLMs* refers to large language models such as Llama2 [66], which can generalize a speaker’s biography based on their conversational behavior. The *prompting* function uses a template that includes two conversation instances (\mathcal{C}) and the target speaker identifier (s_j), thereby enabling the LLM to leverage its knowledge effectively (see Table 4.1). To prevent overly verbose descriptions, we constrain the output length by appending a “note” to the template. Consequently, we obtain concise yet informative persona of the speakers (\mathcal{B}) in each conversation, which are subsequently used to enhance speaker modeling in our ERC framework.

Table 4.1: Prompting template used to extract speaker characteristics from a conversation using an LLM.

Given this conversation between speakers: <code>{conversation content \mathcal{C}}</code> In overall above conversation, what do you think about the characteristics of speaker <code>{speaker identification s_j}</code> ? (Note: provide an answer within 250 words)

BERT-based BiosERC architecture In the proposed BiosERC model, we begin by encoding the speaker’s description using a RoBERTa to obtain a hidden vector representation, h_j^{desc} , for each speaker j :

$$h_j^{desc} = \text{RoBERTa}(d_j)[0] \quad (4.10)$$

where j represents the index of the speaker in the conversation, with $0 \leq j < |\mathcal{S}|$. Our proposed method, BiosERC, extends the baseline model by redefining the speaker’s hidden vector representation ($h_i^{speaker}$ in Equation 4.7), as shown in Figure 5.1. This architecture is designed to ispeaker personality information into the corresponding utter-

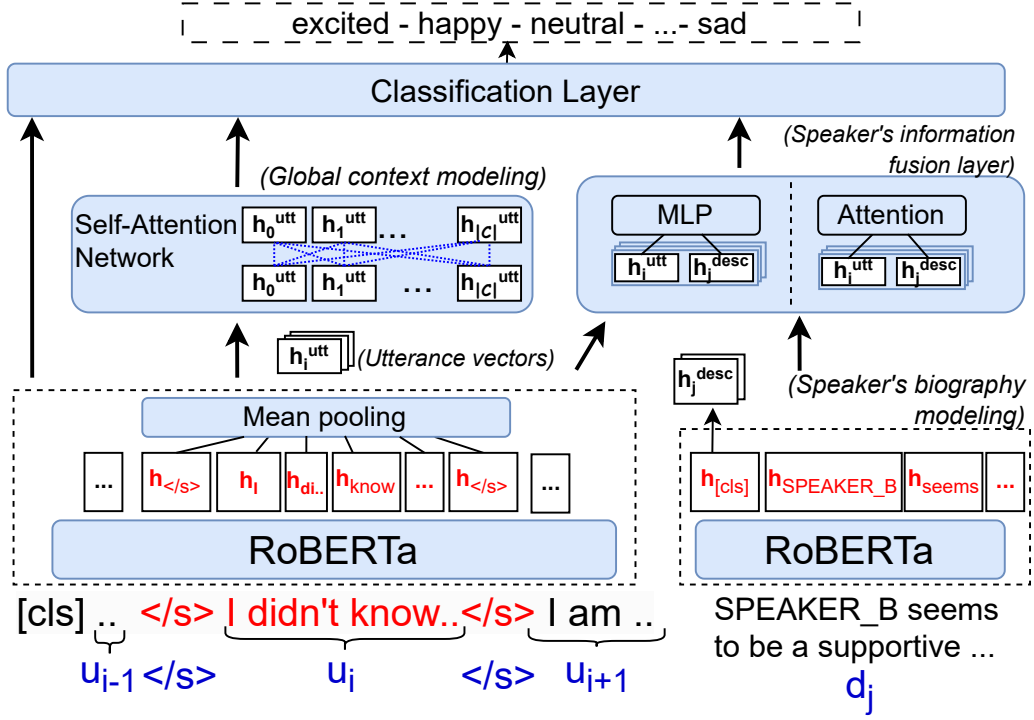


Figure 4.2: Overview of our BiosERC model architecture.

ances via a **multi-layer perceptron** network. This is achieved by replacing the original speaker-specific vector in Equation 4.7 with:

$$h_i^{speaker} = h_{p(u_i)}^{desc} W^{desc} + b^{desc} \quad (4.11)$$

where $p(u_i)$ denotes the speaker associated with the i -th utterance. Through this mechanism, all the utterances from the same speaker share a unified speaker vector representation, with the weights being updated during the training process. Finally, the utterance vector is fused with the speaker vector to supports emotional classification, ensuring that the speaker’s personality is integrated into the emotion recognition process.

BiosERC - Biography Injected by Attention Mechanism. We consider a variant of the BiosERC model designed to dynamically incorporate speaker-specific information into each utterance through an attention mechanism. By modeling the relationship between the current utterance and all individual speakers, the model enriches the utter-

ance representation with speaker-aware contextual information.

$$h_i^{fusion} = h_{p(u_i)}^{desc} W^p + h_i^{utt} \quad (4.12)$$

$$h^{desc} = \{h_j^{desc}\}_{0 \leq j < |S|} \quad (4.13)$$

$$h_i^{speaker} = \text{Attn}(h_i^{fusion}, h^{desc}, h^{desc}, \mathbf{0}) \quad (4.14)$$

The process begins by computing a fusion vector h_i^{fusion} , combining the speaker’s description and the corresponding utterance vector. Next, we aggregate all speaker description vectors (h^{desc}) and apply the attention mechanism to model the relationship between the current utterance and all speakers in a conversation. Finally, the speaker features are embedded into this vector, and the resulting speaker vector $h_i^{speaker}$ is used to replace the original each representation in Equation 4.7 of the baseline system. The attention mechanism ensures that the speaker’s influence is dynamically integrated into each utterance, improving the emotion recognition process.

LLM-based BiosERC + instruction fine-tuning (ft LLM) Building on the robust natural language understanding capabilities of LLMs [66], we incorporate speaker descriptions directly into the input prompts (as highlighted in blue in Table 6.1), rather than altering the model architecture. This approach follows the instruction fine-tuning paradigm [7], using a *causal language modeling* objective to train the LLM to generate emotional label text (as highlighted in red in Table 6.1):

$$x = \text{prompting}(u_i, s_j, d_j, \mathcal{C}, e_i) \quad (4.15)$$

$$\mathbb{P}(x) = \prod_{z=1}^{|x|} \mathbb{P}(x_z | x_0, x_1, \dots, x_{z-1}) \quad (4.16)$$

Here, x denotes the sequence of tokens in the prompting input, and z represents the corresponding token indices (as shown in Table 6.1). To reduce the number of trainable parameters, we adopt LoRA [22], a lightweight fine-tuning technique. The instruction-tuned LLM is trained to learn the distribution of emotional labels conditioned on the prompting input x . The fine-tuned model is trained to predict the emotional label (e_i) of

a given utterance, which is omitted from the input prompt.

Table 4.2: Prompting nput template using speaker description and content of conversation for fine-tuning LLMs.

<i>system</i>
You are an expert at analyzing the emotion of utterances among speakers in a conversation.
Given the characteristic of this speaker, {speaker name s_j }: {speaker description d_j }
Given the following conversation as a context {conversation \mathcal{C} }
<i>user</i>
Based on the above conversation and characteristics of the speakers, which emotional label of $\{s_j\}$ in the utterance {utterance u_i } ?
<i>assistant</i>
{emotional label of u_i in text: e_i }

4.4 Experiment

Implementation Details Leveraging the recent advancements in LLMs [66, 75], we utilize the Llama-2 model to procure personality description for each participant in the conversation. Specifically, we evaluate the effectiveness of speaker description information using two aforementioned PLMs: a BERT-based model with `roberta-large` and a transformer-based, decoder-only LLM model with Llama-2-13b. The best model is determined based on the development set of each dataset and is subsequently employed to evaluate the test set. For fine-tuning the BERT-based BiosERC (section 4.3.2), the hyper-parameters were selected as follows: the learning rate was chosen from $\{1e^{-5}; 5e^{-6}\}$; the dropout rate was set to 0.2, the number of epochs was 30; and the local context window size (w) was chosen from $\{2, 4\}$; The average scores across 10 independent runs are reported. For fine-tuning LLM-based BiosERC (section 4.3.2), the learning rate was selected from $\{2e^{-4}; 3e^{-4}\}$, and the number of epochs was set to 3. Due to computational costs, the average scores across 5 independent runs are reported.

Table 4.3: Performance comparison between our proposed method and prior works on the test sets. The column $\#T.Params.$ denotes the number of trainable parameters. Symbols \ddagger and \dagger indicate statistically significant improvements over the baseline according to a t-test, at levels $p < 0.01$ and $p < 0.05$, respectively.

Methods	$\#T.Params.$	IEMOCAP	EmoryNLP	MELD
HiTrans [33]		64.50	36.75	61.94
DAG [63]		68.03	39.02	63.65
DialogXL [62]		65.94	34.73	62.14
DialogueEIN [43]		68.93	38.92	65.37
SGED + DAG-ERC [2]		68.53	40.24	65.46
S+PAGE [39]		68.93	40.05	64.67
InstructERC [32] $+ (ft\ LLM)$		71.39	41.39	69.15
INTR/INTER ERC (baseline) [26]	189×10^6	67.65	39.33	64.58
BIOSERC _{BERT-based}	186×10^6	67.79	39.89 \dagger	65.51 \ddagger
BIOSERC $+ft\ LLM_{Llama-2-7b}$	80×10^6	69.02	41.44	68.72
BIOSERC $+ft\ LLM_{Llama-2-13b}$	125×10^6	71.19	41.68	69.83

4.5 Result Analysis

4.5.1 Main Results

Our approach demonstrated competitive performance compared to recent SOTA methods on three well-known benchmark datasets (Table 7.1) across both BERT-based and transformer-based decoder-only LLM models.

In comparison with the previous speaker-based methods (SGED + DAG-ERC [2], S+PAGE [39] and DialogueEIN [43]), the experimental results demonstrated the effectiveness of our proposed approach, and further affirm that speaker modeling using speaker descriptions outperforms the information derived by intra- and inter-speaker contexts. In addition, our BiosERC model achieved significant differences over the baseline system on both the EmoryNLP and MELD datasets, as shown clearly in Figure 4.3. Because the MELD and EmoryNLP are multiparty conversation datasets (the average number of interlocutors are 2.72 and 3.34, respectively), emotions in these datasets are influenced more by the distinct personalities of the speakers than in the IEMOCAP dataset.

Unlike previous method of fine-tuning an LLM, such as InstructERC [32], which considers speaker identifier as an auxiliary task and requires two-stage training. Our proposed

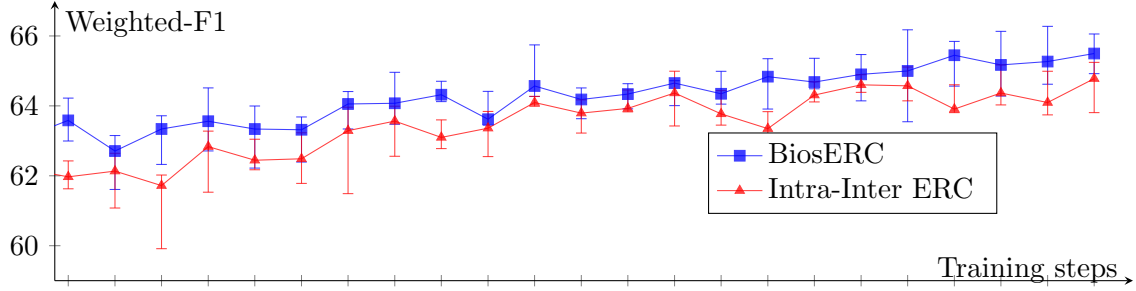


Figure 4.3: Performance comparison between our BERT-based BiosERC model and the baseline on the MELD development set, illustrating variability across 10 random runs.

simplifies the process and reduce time-consuming. Additionally, our method uses speaker descriptions generated by LLMs in natural language, facilitating easier customization for real-world applications (e.g., customer support staff can directly provide or modify characteristics generated by LLMs for their customers).

Similar to the BERT-based BiosERC, the LLM-based BiosERC demonstrates particular strength on multiparty datasets, such as EmoryNLP and MELD, where more than two speakers are involved in each conversation. By fine-tuning an LLM `Llama-2-13b`, the performance of our BiosERC increased by 1-4% in weighted F1 scores compared to the BERT-based models, achieved new SOTA performance on EmoryNLP and MELD datasets. Besides, by employing the lightweight training technique LoRA [22], the number of trainable parameters in LLM-based BiosERC was smaller than BERT-based BiosERC (which was fine-tuned only on last two layers), demonstrating the potential of LLM-based BiosERC for real applications.

4.5.2 Ablation Study

We conducted an ablation study to assess the effectiveness of incorporating speaker biographies into the overall system, which integrates multiple contextual components.

BiosERC Architecture. As shown in Table 4.4, it is apparent that our BERT-based BiosERC (row 3), which incorporates the speaker’s descriptions, demonstrates significant improvements in F1 score, outperforming the baseline system that relies solely on intra- and inter-speaker relationships. Additionally, by using the attention mecha-

nism to encode the speaker’s biography (row 2), BiosERC achieved high performance clearly outperformed the baseline model. Moreover, in the *BiosERC +fine-tuning LLM* setting (row 8), removing the speaker description (highlighted in blue in Table 6.1) from the input prompting (row 6), the performance significantly decreased by 1.05 F1 score. By fine-tuning the different LLM models, Llama-2-13b and Llama-2-7b, led to a slight performance drop of 0.52 F1 score (rows 7, 8). These results underscore the importance of incorporating speaker biography and validate the effectiveness of our proposed approach for speaker modeling.

Table 4.4: Performance comparison among variants of BiosERC on the MELD development set.

Methods	LLMs extracting bio.	Weighted-F1
1. INTRA/INTER ERC (baseline)	-	66.08 _(-1.19)
2. BiosERC <i>injecting bio. by attention</i>	Llama-2-chat-70b	66.71 [†] _(-0.56)
3. BiosERC	Llama-2-chat-70b	67.27[‡]
4. BiosERC	Llama-2-chat-7b	67.23 [‡] _(-0.04)
5. BiosERC	vicuna-33b-v1.3	66.96 [‡] _(-0.32)
6. BiosERC +ft LLM _{Llama-2-13b} <i>w/o speaker bio.</i>	-	69.17 _(-1.05)
7. BiosERC +ft LLM _{Llama-2-7b}	Llama-2-chat-70b	69.70 _(-0.52)
8. BiosERC +ft LLM _{Llama-2-13b}	Llama-2-chat-70b	70.22[†]

Speaker Biographies. We explored various currently popular LLMs for generating speaker biographies, including LLama-2-chat-70b, Llama-2-chat-7b [66], and vicuna-33b-v1.3 [91]. Among these, LLama-2-chat-70b produced the best outcomes. Upon further investigation, we observed that the Vicuna model struggled to provide speaker descriptions in certain challenging cases, such as when the conversation length was too short (e.g., fewer than three utterances) or when a speaker’s utterances were extremely brief (e.g., “Hmm”). These solid improvements, achieved through diverse biographies generated by various LLMs, highlight the versatility and effectiveness of extracting “speaker biographies”, demonstrating their potential in enhancing ERC tasks.

4.5.3 Conversation Length

To assess the impact of conversation length on model performance, we conducted an analysis on the MELD development set, as illustrated in Figure 4.4. Overall, our method consistently outperforms both intra- and inter-speaker baselines across conversations of varying lengths. Notably, the performance improvement for short dialogues (with a conversation length of fewer than 15 utterances) is significantly greater than for long dialogues. These results further emphasize the importance of “speaker characteristic” in short conversations, which tend to lack sufficient contextual information. In such cases, the speaker’s lexical choices become crucial, as they often carry explicit or implicit meaning. The LLM can extract the speaker’s characteristics by recognizing the explicit or implicit meaning conveyed in these statements. Additionally, since MELD is a multiparty dataset with many conversations involving more than three speakers, our observations highlight that “speaker characteristics” play a particularly significant role in multi-party short conversations where contextual information is limited.

4.5.4 Case Study

Our model improves emotion recognition accuracy, even in short conversations where contextual information is limited. As shown in Table 5.8, conversation 1041 is a short dialogue consisting of only five sentences. In this case, our model leverages two speaker descriptions to more accurately interpret SPEAKER_0’s discourse, correctly identifying the tone as positive rather than angry. And our architecture demonstrates enhanced capability in predicting emotions in shorter utterances, such as “*Yeah*” or “*Okay*”, where conventional models often struggle due to the limited semantic content.

Furthermore, our method effectively handles cases with high error rates at the beginning of conversations, as illustrated in Table 5.8 (u_0 , u_1 , u_2 in conversation 1061). In these early stages, contextual and speaker-specific information is minimal, leading the baseline model to consistently misclassify initial utterances. However, with the incorporation of our “speaker description”, the model is able to achieve more accurate predictions

from the very first utterances of the dialogue. These results demonstrate the robustness and adaptability of our model across a wide range of conversational settings, including those characterized by short lengths or complex contextual dynamics.

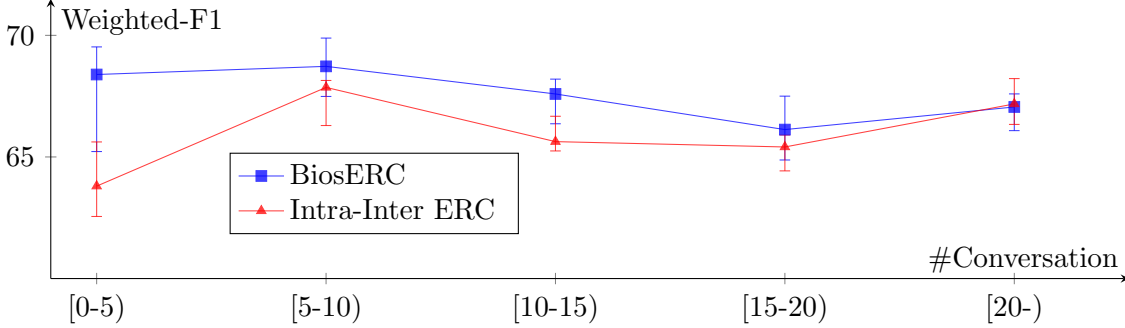


Figure 4.4: Performance comparison respect to length of conversation (number of utterance) on the MELD development set (variability across 10 random runs).

4.6 Chapter Summary

In conclusion, we proposed a novel mechanism for incorporating speaker characteristics into the ERC task—an area that has been underexplored in previous research. By investigating how interlocutor personalities influence emotional expressions, we leveraged this external knowledge as a unique and complementary feature to enhance model performance. Our experiments on three widely used benchmark datasets consistently yielded SOTA or competitive results, demonstrating the effectiveness of our approach. Moreover, our method remains simple yet highly adaptable, making it suitable for a broad range of conversation analysis tasks.

Limitation In this work, we introduced a new framework for speaker modeling that utilizes biographical information about interlocutors, automatically generated by LLMs. While effective, our BiosERC method incurs higher computational costs due to the additional inference step required by the LLM, in contrast to models such as *Intra-inter ERC*, which rely on latent speaker identity embeddings. Furthermore, this study does not address potential privacy concerns related to human data. In real-world applications, access to conversation histories must be explicitly granted and clearly communicated by the data

Table 4.5: Case study of improvement examples from the MELD dataset. Red and green labels indicate incorrect and correct model predictions, respectively.

Idx	Conversation 1041			Label	BiosERC	Baseline
	Speaker_0	Speaker_1				
d ₀	SPEAKER_0 in the conversation comes across as someone who is confident , friendly .. to create a relaxed atmosphere ...					
d ₁	SPEAKER_1 in the conversation comes across as a friendly .. have a strong sense of loyalty and trust in their relationships...					
u0	Hey Estelle,listen			neutral	neutral	neutral
u1		Well! Well! Well! Joey Tribbiani! So you		surprise	surprise	joy
u2	What are you talking about? I never left	came back huh?		surprise	surprise	anger
u3	you! You've always been my agent!			surprise	surprise	surprise
u4	Yeah!	Really?!		joy	joy	anger

Idx	Conversation 1061			Label	BiosERC	Baseline
	Speaker_0	Speaker_1	Speaker_2			
d ₀	SPEAKER_0 seems to be a very inquisitive and curious person SPEAKER_0 appears to be quite blunt and direct in his communication style, not mincing words or sugarcoating his thoughts.					
d ₁	SPEAKER_1 seems to be a humorous and light-hearted person SPEAKER_1 shares that they have only been with one person in their whole life, and this is met with surprise and disbelief by the other ...					
d ₂	SPEAKER_2 seems to be a humorous and light-hearted person ... SPEAKER_2 is someone who enjoys having fun and is not afraid to poke fun at themselves or others...					
u0	Well, what?			neutral	neutral	surprise
u1	What?			neutral	neutral	surprise
u2	What is it?			neutral	neutral	sadness
u3	That she left you?			surprise	surprise	sadness
u4	That she likes women?			neutral	sadness	sadness
u5	That she left you for another woman that likes women?			neutral	surprise	sadness
u6		Little louder, okay, I think there's a man on the twelfth floor in a coma that didn't quite hear you.		anger	neutral	anger
...						
u11		With Carol? Oh.		surprise	surprise	neutral
u12	So in your whole life, you've only been with one oh.			surprise	neutral	neutral
u13		Whoah, boy, hockey was a big mistake! There was a whole bunch of stuff we could've done tonight!		surprise	surprise	joy

owners. Nonetheless, we believe that with proper user agreements and privacy safeguards, obtaining such permission is feasible and ethically manageable.

Chapter 5

TraceERC: Tracking Relational Awareness of Contextual, Character, and Emotional States in Emotion Recognition in Conversations

5.1 Background and Motivation

Unlike non-conversational emotion recognition tasks, ERC presents unique challenges arising from the nature of dialogue data, which are typically transcriptions of spoken interactions. Such texts are often characterized by truncated statements and non-standardized grammar. Consequently, ERC tasks heavily rely on multiple facets of contextual information, including (i) the dialogue context, (ii) speaker-specific characteristics, and (iii) the dynamic flow of emotions throughout the interaction. While recent advancements have explored implicit speaker relationships, explicit speaker information (e.g., personality) remains largely underexplored. This omission hinders accurate modeling of speaker dynamics and poses challenges for model adaptability and generalization. In our previous

work, we have addressed this issue [77].

Another critical aspect is that due to the spontaneous and unpredictable nature of conversations. In particular, text-based emotion recognition, which limited by the absence of speech and visual modalities [72], may inadequate when relying solely on contextual utterances and static speaker information may be insufficient for capturing the transient fine-grained, and dynamically activated emotions at each timestep throughout the conversation [40]. Moreover, existing studies have overlooked the dependencies between a speaker’s personality and the relativity of emotional states. At the same time, the inherent spontaneity of dialogues and the limitations of text-based emotion recognition make it particularly challenging to detect emotional shifts promptly and to disambiguate similar emotion labels. Since dialogues can occur at any time and place, speakers may experience sudden and unforeseen emotional shifts triggered by personal memories or unexpected events, significantly altering the emotional trajectory (e.g., a sudden shift from sadness to angry, as illustrated in Figure 5.1). To address the challenge of modeling the temporality of emotions at each timestep throughout a conversation, it is essential to model dynamic emotional cues and the dependencies between emotional states, while also exploring the intimate connection between personality and expressed emotions.

As illustrated in Figure 5.1, we provide an example showing that emotions evolve dynamically throughout the course of a conversation. In analyzing nuanced emotional expressions such as in utterance #11, several critical cues should be considered: Initially, based on the speaker description, Speaker_0 is identified as an emotional person who initially feels “sadness” and “regret” due to the belief that Speaker_2 has passed away. Secondly, the emotional context indicates that the surrounding utterances (from #6 to #10) are “joyful”, suggesting that Speaker_2’s overall emotional atmosphere is “joyfully joking”. Thirdly, the emotional cue in Speaker_0’s previous utterance (#5) reveal that her current emotional state is “sadness”. Finally, reasoning from the conversational content allows for the inference that Speaker_0 realizes she has been deceived by Speaker_2. These combined factors lead to an emotional shift from “sadness” to “anger”. This example underscores the importance of incorporating both speaker personality traits and dynamic

emotional cues for accurately identifying emotions in complex dialogues.

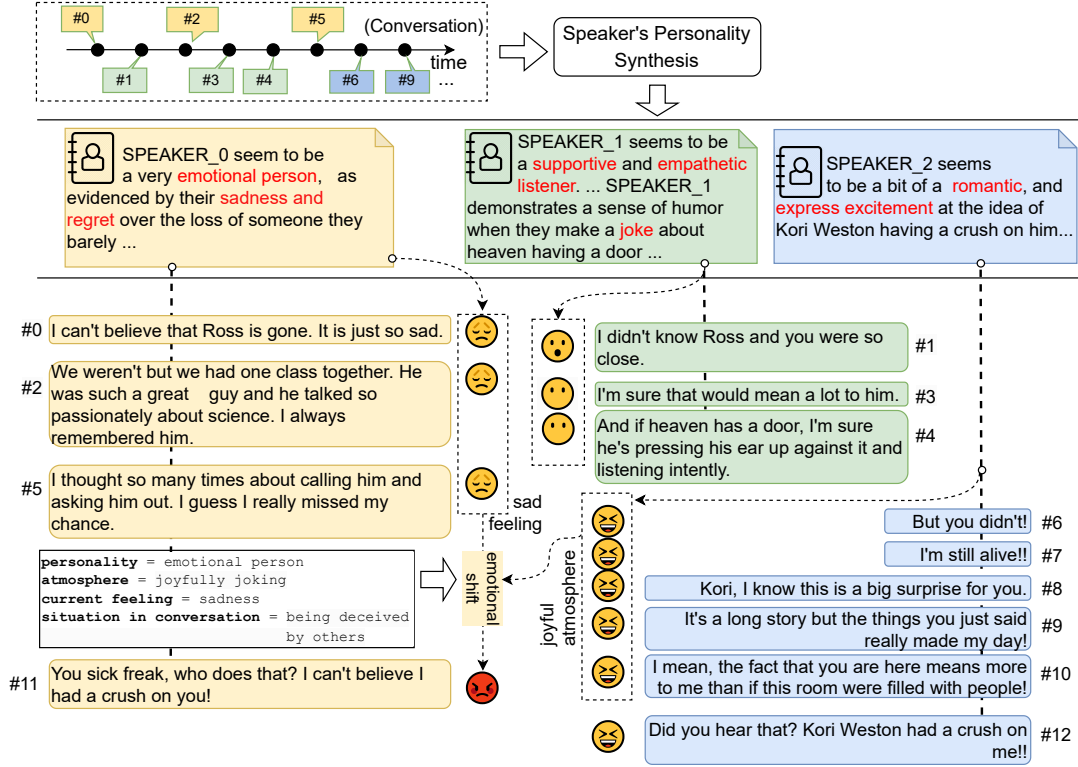


Figure 5.1: Overview of our TraceERC framework architecture.

To bridge the aforementioned gaps, we formalize the influence of three salient aspects on human emotions in conversations: conversational content, static speaker-specific characteristics, and dynamic emotional cues within interactions. Accordingly, we propose TraceERC: **T**racking **R**elational **A**wareness of Contextual, **C**haracter, and **E**motional States in **ERC** tasks. By harnessing the capabilities of LLMs, our method incorporates speaker-specific personality traits and encoded dynamic emotional atmosphere as supplementary knowledge. These features enable the generation of contextually enriched, emotion-tagged examples tailored to corresponding dialogue scenarios and speakers' unique profiles, enhancing the model's emotional understanding and adaptability. Furthermore, as one of the pioneering studies to explore both in-context learning and instruction fine-tuning within LLM-based ERC, we conduct rigorous evaluations to validate the contributions of each component.

5.2 Related Works

Emotional context The aforementioned methods primarily focus on modeling the semantic context of conversations, typically overlooking the emotional dependencies and distinguishing similar emotions. To address this gap, some studies [15, 73, 76, 25] have explored the emotional context by modeling the global evolution of emotion states using various methods. COSMIC [15] employs an additional GRU to monitor the progression of global emotional states over time. CAUXIT [27] as a semi-supervised framework, which leverages contextual conversations to enhance emotion prediction by recovering augmented emotional tags for masked utterances in unlabeled data. CESTa [73] framed the ERC task as a sequence tagging problem, leveraging conditional random field to model emotional consistency within conversations. LYSM [76] captures the dependencies between the emotional states by applying the self-attention mechanism, enabling effective modeling of long-range emotional transitions. DialogueEIN [43] employed emotion embeddings to represent candidate emotion categories and utilized a multi-head attention module to capture the emotional tendencies of each utterance. Furthermore, the depth of a model’s understanding of emotion labels plays a critical role in its performance on ERC tasks. IEIN injected explicit emotional labels into the ERC task and focused on learning iterative emotional interaction throughout the training process. To mitigate the challenges posed by emotion similarity, EACL [81] has focused on representing closely related emotion pairs, such as excited and happy or frustrated and sad. They proposed a novel emotion-anchored contrastive learning framework designed specifically for ERC, with the goal of generating more distinctive utterance representations, thereby effectively guiding the representation learning process. In contrast to previous approaches that implicitly model emotional dependencies or rely solely on static emotion embeddings, our dynamic iterative emotional interaction network emphasizes leveraging dynamically predicted emotional labels to explicitly model emotional dependencies. This approach enables tailored emotion understanding for different conversational contexts and speakers, serving as dynamic few-shot guidance to enhance emotional representations.

Large Language Model for ERC In recent years, LLMs and their variants have demonstrated remarkable efficacy in tasks such as in-context learning, few-shot prompting, and instruction following, spanning a wide array of downstream NLP applications. These capabilities offer a novel perspective to addressing the challenges posed by the lack of intricate annotations, such as speaker characteristics in ERC tasks. Several studies have also attempted to integrate LLM models with ERC tasks. For instance, InstructERC [32] reformulates the ERC task as a unified Seq2Seq paradigm by fine-tuning LLMs to transfer the inference ability of pre-train LLMs. MKFM [68] integrates multiple knowledge (e.g., co-reference, topics, emotional cause) generated by ChatGPT into a graph-based model. In contrast to these approaches, TraceERC is the novel framework to employ open-source LLMs for extracting crucial yet previously neglected speaker personality information. Additionally, it strengthens representation features through the integration of emotional states, pioneering the effective application of in-context learning in ERC tasks.

5.3 Methodology

In this section, we formalize the problem statement and elaborate on the motivation and core components of our proposed TraceERC, emphasizing its contributions and key distinctions from baseline systems. This work adopts an LLM-based approach in preference to prior methodologies relying on graph attention networks [62, 2] or customized architectures built upon small pre-trained models [43, 81]. Our motivation stems from the transformative potential of LLMs to serve as a foundational technology for general-purpose AI systems. We argue that their scalability, adaptability, and capacity for in-context learning make LLMs particularly well-suited for advancing ERC and extending its applicability to a broader range of downstream tasks.

5.3.1 Problem Definition

Formally, we define a conversation as $\mathcal{C} = \{u_i\}_{0 \leq i < |\mathcal{C}|}$, where each utterance u_i is articulated by a speaker $p(u_i) \in \mathcal{S}$, with $\mathcal{S} = \{s_j\}_{0 \leq j < |\mathcal{S}|}$ denoting the set of all speakers involved in the dialogue. Here, i and j denote the indices of an utterance and the interlocutor in a conversation, respectively, and p is a mapping function that associates each utterance with its corresponding speaker. Given a conversation \mathcal{C} as input, the target of an ERC system is to predict the emotional label (e_i) of each utterance, draw from a predefined label set (\mathcal{E}^{label}):

$$\hat{e}_i = \operatorname{argmax}_{e_i \in \mathcal{E}^{label}} \mathbb{P}(e_i \mid u_i, p(u_i), \mathcal{C}, \theta) \quad (5.1)$$

where \mathbb{P} represents the probability that the utterance u_i belongs to the emotional label e_i , conditioned on the conversation context and speaker information, and θ denotes the model hidden parameters.

5.3.2 Baseline - LLM-based ERC

In contrast to previous approaches based on pre-trained BERT-based architectures [62, 2, 43, 81], we are among the first to apply instruction fine-tuning [7] with LoRA [22] to train open-source LLMs for solving ERC tasks. Capitalizing on the robust natural language understanding abilities of LLMs [66], we adopt a prompt-based paradigm in which all necessary information is embedded directly into the input text (x), rather than designing the specific model architecture, a prompt template and example are shown in Table 5.1. By using an instruction fine-tuning technique, we train the LLM with a *causal language modeling* objective to produce emotional label text (highlighted in red in Table 5.1). The model is trained to learn a probability distribution over the textual output, conditioned on the given input prompt with gold emotion annotations. In the inference/testing phase, the emotional label prediction portion of the prompt (red text) is removed, and the model is tasked with generating the label based solely on the learned representation of

the preceding dialogue context.

$$x = \text{baseline-prompting}(u_i, s_j, \mathcal{C}, e_i) \quad (5.2)$$

$$\mathbb{P}(x) = \prod_{z=1}^{|x|} \mathbb{P}(x_z | x_0, x_1, \dots, x_{z-1}) \quad (5.3)$$

Table 5.1: Prompting input template and example for the baseline model.

<i>(baseline-prompting template)</i>
<i>system</i>
You are an expert at analyzing the emotion of utterances among speakers in a conversation.
Given the following conversation as a context { conversation \mathcal{C} }
<i>user</i>
Based on the above conversation and characteristics of the speakers, which emotional label of $\{s_j\}$ in the utterance { utterance u_i } ?
<i>assistant</i>
{emotional label of u_i in text: e_i }
<i>(example)</i>
<i>system</i>
You are an expert at analyzing the emotion of utterances among speakers in a conversation.
Given the following conversation as a context
SPEAKER_0: Thanks.
SPEAKER_1: Yep! Looks like it's gonna be a leeeean Christmas at the Dragon house this year.
SPEAKER_0: Enough!
SPEAKER_1: Lean-lean-lean!
<i>user</i>
Based on above conversation and characteristic of the speakers, which emotional label of SPEAKER_0 in the utterance "Enough!"?
<i>assistant</i>
Mad

5.3.3 TraceERC

In this section, we introduce our TraceERC framework, detailing the design and motivation underlying each of its core components. TraceERC is a multi-context-aware conversational emotion recognition model that integrates dialogue context, speaker personality traits, and dynamic emotional state cues. In response to the growing emphasis on explicit reasoning in machine learning systems [87], our approach leverages LLMs in a novel way to extract critical yet previously underutilized speaker-specific personality characteristics, bridging a key gap in ERC research. By integrating emotional states

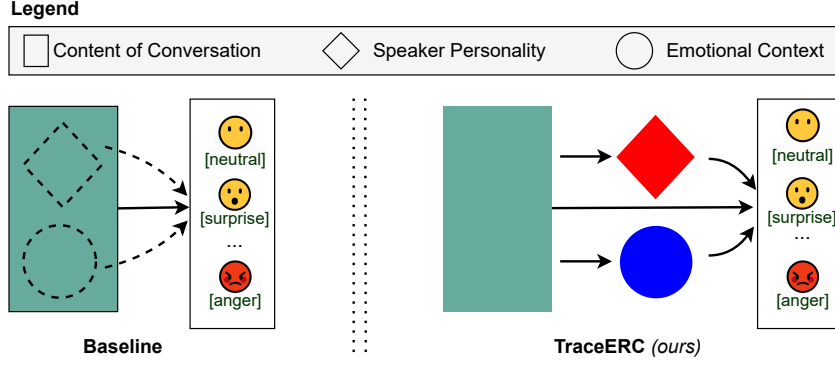


Figure 5.2: Idea of our TraceERC framework compared to the baseline model.

alongside personality traits, TraceERC refines representational features of emotion labels, culminating in a robust and comprehensive multi-context-aware ERC system. Compared to the previous works [43, 31, 62], our method departs from reliance on implicit feature learning and instead introduces a mechanism for explicitly instructing the model to utilize speaker and emotional context in performing ERC tasks. As illustrated in Figure 5.2, earlier approaches tend to encode speaker and emotional cues indirectly within the dialogue context, which constrains their contribution to final emotion predictions. TraceERC overcomes this limitation by explicitly incorporating these elements into the modeling process, ensuring their direct influence on prediction outcomes.

The proposed TraceERC model operates through three primary stages, as depicted in Figure 5.3.

- Stage 1 involves the use of prompt-based inference with LLMs to extract speaker-specific personality profiles from the dialogue. These profiles capture habitual lexical choices, personality traits, and emotional fluctuation tendencies unique to each speaker.
- Stage 2 applies instruction fine-tuning to LLMs or smaller PLMs to preliminary learn to generate the contextual emotional states. This stage incorporates two primary inputs: the dialogue context and the speaker description from stage 1. Through fine-tuning these inputs, the model generates preliminary emotion predictions enriched with speaker-aware insights.

- Stage 3 further fine-tunes the LLMs by integrating emotional context cues derived from previous stages. Specifically, the model ingests the dialogue content, synthesized speaker personality descriptions, and initial emotion likelihoods. These emotional contexts function as dynamic guidance examples within a few-shot learning mechanism, enhancing emotion representation and improving the model’s ability to recognize subtle emotional transitions.

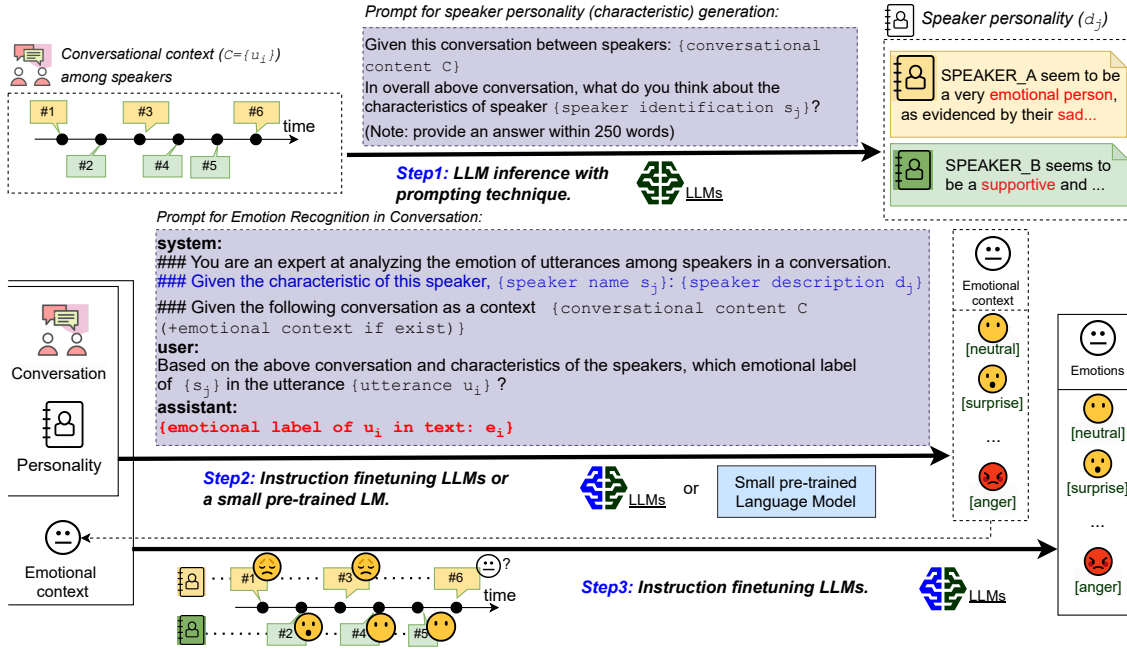


Figure 5.3: Architecture of our TraceERC framework.

5.3.4 Speaker Personality Extraction.

We begin by outlining the first stage of the TraceERC framework, which is dedicated to generating speaker-specific personality descriptions for each participant in a conversation. While previous studies [68, 32] have explored the extraction of various supplementary information from dialogues, our approach uniquely prioritizes speaker-related attributes. By systematically evaluating diverse prompting strategies, we aim to identify the most effective approach for extracting speaker-specific knowledge that are not explicitly present in annotated data. To achieve this, we propose a prompting-based mechanism that leverages

the capabilities of LLMs to generate speaker-specific descriptions(d_j) and interpretative rationales explaining emotional transitions. Given a conversation \mathcal{C} , this step produces a set of biographical profiles for all speakers (\mathcal{S}) involved in a conversation $\mathcal{B} = \{d_j\}_{0 \leq j < |\mathcal{S}|}$:

$$d_j = \text{LLMs}(\text{prompting}(\mathcal{C}, s_j)) \quad (5.4)$$

where *LLMs* refers to LLMs such as LLaMA2 [66], which are capable of summarizing the speaker’s personality traits and providing plausible explanations for their emotional dynamics based on dialogue content. The *prompting* function operates as a structured template that combines both conversation instance (\mathcal{C}) and speaker identification (s_j) information (an example of prompting input which is illustrated in Step 1 of Figure 5.3). To further enhance personality modeling, we introduce three types of speaker descriptions derived via prompting: (1) **DetailDesc** provides a detailed analysis of the speaker’s personality, offering a comprehensive perspective on the entire conversation. (2) **EmoTransDesc** captures the speaker’s personality while specifically focusing on analyzing emotional transitions. (3) **ConciseDesc** offers a concise and focused characterization of the speaker. The effectiveness and characteristics of these descriptions are systematically examined in the results and ablation studies presented in the results analysis section.

Emotional Context Extraction. In this work, emotional context is defined as the likelihood distribution of emotional labels associated with utterances surrounding a given target utterance (as shown in the template “*predicted emotion: e_i* ” in Table 6.1). To mitigate the impact of noise introduced by inaccurate early-stage predictions, these emotional states are extracted during training only after a few initial epochs, as determined by a designated hyperparameter, ensuring that the model has attained a more stable level of predictive performance. The integration of emotional context into our framework is motivated by the following key factors: (1) Enhancing emotional adaptability: enabling the model to anticipate and adjust to emotional transitions (e.g., from *sadness* to *happy*); (2) Providing a comprehensive emotional atmosphere: capturing the explicit emotional

trend of the entire conversation (e.g., whether the atmosphere is tense or pleasant) while analyzing dependencies between adjacent emotional states; (3) Contextualizing emotional labels with speaker-enriched emotional context: enriching each emotional label with concrete, contextually grounded examples from surrounding utterances, supplemented by speaker-specific personality traits. This few-shot-style guidance offers semantically rich interpretations of emotion categories, improving the model’s ability to guide the model’s understanding of emotional representations.

Emotional context remains one of the most neglected dimensions in previous ERC studies. Examining the emotional states of interlocutors not only maps the emotional trajectory of the dialogue but also enhances the model’s ability to comprehend and differentiate emotional labels, revealing latent affective relationships embedded in context. Our proposed TraceERC implicitly broadens the temporal and emotional context scope and captures emotional transitions, thereby deepening the model’s insight into the emotions conveyed within each utterance. The detailed procedure for generating and integrating emotional context is seamlessly integrated into the TraceERC training procedure, which is elaborated in the subsequent section.

5.3.5 Learning Scenarios

To evaluate the effectiveness of our TraceERC framework compared to the baseline method, we implement two widely adopted learning paradigms for LLMs, including *Instruction fine-tuning* and *In-context learning*. In the instruction fine-tuning scenario [7, 22], the model is trained on the full manually annotated dataset, with its weights updated to specifically address the ERC task. In the in-context learning scenario [4], the model’s weights remain unchanged during the learning process; instead, it performs inference by conditioning on task-relevant information embedded directly within the input prompt. Evaluating our framework under both paradigms provides a robust measure of its generalizability, minimizing the influence of training-specific biases and demonstrating the model’s adaptability across varying deployment conditions.

Table 5.2: TraceERC prompting input template and example using speaker description and emotional context information.

<i>(TraceERC-prompting template)</i>	
<i>system</i>	<p>### You are an expert at analyzing the emotion of utterances among speakers in a conversation.</p> <p>### Given the characteristic of this speaker, {speaker name s_j}: {speaker description d_j}</p> <p>### Given the following conversation as a context {conversation \mathcal{C} and predicted emotional context $(\mathcal{E}^{prediction})$}</p>
<i>user</i>	Based on the above conversation and characteristics of the speakers, which emotional label of $\{s_j\}$ in the utterance {utterance u_i } ?
<i>assistant</i>	{emotional label of u_i in text: e_i }
<i>(example)</i>	
<i>system</i>	<p>### You are an expert at analyzing the emotion of utterances among speakers in a conversation.</p> <p>### Given the characteristic of this speaker, SPEAKER_0: SPEAKER_0 seems to be a bit of a grump in this conversation. They don't seem to be in the mood for jokes or playful banter, as evidenced by their terse responses to SPEAKER_1's attempts at humor.</p> <p>### Given the following conversation as a context</p> <p>SPEAKER_0: Thanks. (predicted emotion: Joyful)</p> <p>SPEAKER_1: Yep! Looks like it's gonna be a leeeean Christmas at the Dragon house this year. (predicted emotion: Joyful)</p> <p>SPEAKER_0: Enough!</p> <p>SPEAKER_1: Lean-lean-lean! (predicted emotion: Joyful)</p>
<i>user</i>	Based on above conversation and characteristic of the speakers, which emotional label of SPEAKER_0 in the utterance "Enough!" ?
<i>assistant</i>	Mad

Instruction Fine-tuning

Similar to the baseline system, our approach also adopts the *causal language modeling* loss during training. To optimize training efficiency, we utilize LoRA [22], a lightweight fine-tuning technique that significantly reduces the number of trainable parameters. To offer a clear and structured overview of our methodology, the TraceERC training procedure is detailed in Algorithm 1. In the first stage (Lines 1–7), for each conversation in the training dataset, all speaker personality descriptions are generated and stored in a set \mathcal{B} , with each entry corresponding to a distinct speaker identity. Subsequently, in the second stage (Lines 9–19), these speaker descriptions are then incorporated with the conversational content to train an emotional context estimation model (\mathcal{M}^{ec}) over ep^{ec} epochs, where ep^{ec} is a hyper-parameter chosen during the optimization process (e.g., 3 epoch). After that, the predicted emotional context likelihoods (\mathcal{E}^{ec}) generated by \mathcal{M}^{ec} are then used to train the final model, TraceERC. Ultimately, we obtain two checkpoints containing the emotional context model \mathcal{M}^{ec} and the final model \mathcal{M} , which is utilized for evaluation on the test set.

The label refinement strategy employed in our framework gradually transitions emotion predictions from coarse-grained to fine-grained representations. During the first fine-tuning phase, the model produces an initial coarse-grained prediction. This is followed by a second fine-tuning phase, where the model refines the labels by correcting earlier prediction errors, thereby enabling more accurate emotion classification. In comparison to the InstructERC [32], a two-step fine-tuning method, which first trains a speaker identification model to boost the model’s ability to distinguish between speakers before applying it to the final ERC task. Meanwhile, our method instead uses explicit speaker descriptions to support the model’s understanding of the speaker’s personality. Additionally, while MPLP [84] also employs a label refinement approach, it enhances the model’s semantic understanding of emotion labels by providing explanations for each label. However, these descriptions are static and fail to adapt to the nuances of specific conversational contexts. In contrast, our method provides contextual utterances along with their cor-

Algorithm 1 *TraceERC* training procedure

Input: a conversation content \mathcal{C} , annotated data \mathcal{D} , limitation of epoch number for predicting emotional context ep^{ec}

Output: fine-tuned TraceERC model \mathcal{M} , emotional context model \mathcal{M}^{ec}

```
1: for each  $\mathcal{C} \in \mathcal{D}$  do                                ▷ generate speaker descriptions for whole dataset
2:    $\mathcal{S} \leftarrow$  all the speakers in  $\mathcal{C}$ 
3:   for each  $s_j \in \mathcal{S}$  do
4:      $d_j \leftarrow LLM_{spdesc-gen}(prompting(\mathcal{C}, s_j))$     ▷ generate speaker descriptions
5:   end for
6:    $\mathcal{B} \leftarrow \mathcal{B}.append(d_j)$ 
7: end for
8:
9:  $\mathcal{M} \leftarrow$  initialize from pre-trained LLM
10:  $\mathcal{E}^{ec} \leftarrow \emptyset$                                 ▷ emotional contextual is initialized by empty set
11: while  $ep < max\_epochs$  do
12:    $\mathcal{M} \leftarrow \mathcal{M}.train(prompting(\mathcal{E}^{ec}, \mathcal{B}, \mathcal{D}_{train}))$     ▷ call model train procedure
13:   if  $ep = ep^{ec}$  &&  $\mathcal{E}^{ec} = \emptyset$  then
14:      $\mathcal{E}^{ec} \leftarrow \mathcal{M}.predict(\mathcal{D}_{train})$                 ▷ emotional contextual is updated
15:      $\mathcal{M}^{ec} \leftarrow \mathcal{M}$                                 ▷ save the checkpoint of emotional context model
16:      $\mathcal{M} \leftarrow$  initialize from pre-trained LLM
17:   end if
18:    $ep \leftarrow ep + 1$ 
19: end while
20:
21: return  $\mathcal{M}, \mathcal{M}^{ec}$ 
```

responding emotion predictions as instances to serve as explanations for emotion labels. Unlike static descriptions, our method generates dynamic descriptions that are tailored to specific conversational contexts. This strategy not only enhances the semantic richness of the emotional representations but also strengthens the model’s ability to comprehend emotions by incorporating diverse conversational scenarios and speaker-specific characteristics.

In-context Learning

Within the *in-context learning* scenario, where model parameters remain unchanged, the primary challenge lies in effectively leveraging prompts to maximize the reasoning capabilities of LLMs in the ERC task. To handle this issue, we adopt the few-shot prompting technique [4], which lets LLM deeply understand the ERC tasks and produce the desired outputs. Specifically, we first retrieve the most similar conversations from training data to the query conversation, constructing high-quality contextual examples. Our approach extends beyond the traditional approach employed in [32], which primarily focuses on utterance-level similarity. Instead, we incorporate a concatenation of contextual windows and speaker personality traits to identify the most comparable dialogues. We hypothesize that conversations sharing both similar contextual content and speaker characteristics possess greater referential value for the model. These retrieved dialogues are then embedded as few-shot examples provided in the input prompting to guide LLM inference.

To achieve this, we construct a domain base, \mathcal{D}^{domain} , from the training data, in which each focal utterance is accompanied by its surrounding utterances, forming a contextual window. Formally, given a target utterance u_i and the corresponding speaker profile d_j , we define the context c_i is the contextual as the concatenation of neighboring utterances within a window of size w : $c_i = [u_{i-w}, \dots, u_{i+w}]$. We then generate enriched contextual representations for each utterance in \mathcal{D}^{domain} using Sentence BERT [60]. Each contextual vector is further enhanced by concatenating it with the corresponding speaker personality vector. For a given query utterance u_i^{query} , we similarly combine its speaker profile and contextual information to retrieve the most relevant utterance c^{sim} from \mathcal{D}^{domain} based

on *cosine* similarity between the concatenated vectors. This retrieval strategy allows us to identify utterances that are most similar to the target in terms of both speaker-specific traits and contextual relevance.

$$v^{domain} = \text{SBERT}(c_i^{domain}) \oplus \text{SBERT}(d_j^{domain}) \quad (5.5)$$

$$\mathcal{D}^{domain} = \{v^{domain}\}_{\text{for all utterances in training data}} \quad (5.6)$$

$$v^{query} = \text{SBERT}(c_i^{query}) \oplus \text{SBERT}(d_j^{query}) \quad (5.7)$$

$$c^{sim} = \arg \max_{v_t \in \mathcal{D}^{domain}} \text{cosine}(v^{query}, v_t) \quad (5.8)$$

Finally, the contextual information and emotion labels of these similar conversations are provided as few-shot examples in the prompt, guiding the LLM during inference under the in-context learning framework.

5.4 Experiment

Implementation Details. Since the recent successful applications and advancing capabilities of pre-trained LLMs, we leverage the open-source **LLama-2** model [66] to extract personality descriptions for each participant in the conversation. For hyperparameters tuning in the instruction fine-tuning setup, the main large language models employed are **LLama-2-7b** and **LLama-2-13b**. The number of training epochs is fixed at 6, with the first three epochs ($ep^{ec} = 3$) dedicated to emotional context prediction. The learning rate is set to 3×10^{-4} , the batch size is 4, and the local context window size (w) is selected from $\{3, 5\}$. Reported results are averaged scores obtained across 10 independent runs to ensure robustness. For both the speaker personality generation phase and the in-context learning experiments, we conducted experiments using **LLaMA-2-70B**. In alignment with [84], we acknowledge that some prior works in the ERC field [88, 62] reported results optimized directly on the test set, which may lead to inconsistencies between reproduced and officially published results. For hyperparameter tuning, we selected the best model based on performance on the development set, which is essential

to ensure fair and generalizable model evaluation.

5.5 Result Analysis

In this section, we first present the performance of TraceERC in comparison to recent SOTA methods across three well-known benchmark datasets. We then conduct ablation studies to examine the contributions of each individual module. Following this, we provide case studies that illustrate how TraceERC leverages speaker-specific information and emotional context to rectify the misclassified samples. Lastly, we provide an in-depth error analysis to highlight and remaining challenges of our TraceERC framework.

5.5.1 Comparison with Previous Works

To evaluate the generalization of our proposed method, we compare the TraceERC framework against existing SOTA models under two learning scenarios: *instruction fine-tuning* (Table 7.1) and *in-context learning* (Table 5.4). All reported values represent the averages of ten independent runs with different random seeds.

As shown in Table 7.1, TraceERC demonstrates superior performance compared to the baseline and achieves results highly competitive with recent SOTA methods. Notably, it performs best on the MELD dataset. Regarding the comparison between *Baseline* method (described in the section 5.3.2) incorporates conversational context and demonstrates solid performance, TraceERC achieves consistently outstanding and significant improvements across all three datasets. In contrast to previous work BiosERC, which also utilized speaker descriptions, TraceERC achieves superior performance on the IEMOCAP and MELD datasets. While TraceERC delivers competitive results relative to BiosERC on the EmoryNLP dataset. This phenomenon may be attributed to the unique characteristics of the EmoryNLP dataset, which features an average of 3.34 speakers per conversation and relatively shorter dialogue lengths. Under such conditions, speaker personality traits play a more critical role than emotional context in enabling effective emotion recognition. Another contributing factor is the generally lower emotion prediction accuracy

Table 5.3: Performance comparison between our proposed method and previous works on the test sets. Column $\#T.Params.$ refers to the number of trainable parameters. The marked star (*) indicates the preprints (unofficial works). The abbreviation “w/o EC” stands for the setting without using emotional context information. The notation ‡ indicates the significant difference (t-test) with the baseline in levels $p < 0.05$.

Methods	$\#T.Params.$	IEMOCAP	EmoryNLP	MELD
HiTrans [33]		64.50	36.75	61.94
DAG [62]		68.03	39.02	63.65
DialogXL [63]		65.94	34.73	62.14
DialogueEIN [43]		68.93	38.92	65.37
SGED + DAG-ERC [2]		68.53	40.24	65.46
S+PAGE [39]		68.93	40.05	64.67
MPLP [84]		66.65	-	66.65
MFAM [21]		70.16	-	66.65
ESIHGNNa [82]		68.53	39.56	63.92
EACL [81]		70.41	40.24	67.12
<i>(LLM-based method)</i>				
MKFM [69]		68.88	39.76	65.66
*DialogueLLM w/o video [86]		-	-	59.75
*InstructERC +(ft LLM) [32]		71.39	41.39	69.15
<i>(Implemented in this works)</i>				
Baseline $Llama-2-7b$ (ours)	80×10^6	68.15	40.32	68.01
TRACEERC $_{Llama-2-7b}$ w/o EC (\sim BiosERC) [77]	80×10^6	69.02 _(\blacktriangle0.87)	41.44 _(\blacktriangle1.12)	68.72 _(\blacktriangle0.71)
TRACEERC $_{Llama-2-7b}$ (ours)	80×10^6	69.41 [‡] _(\blacktriangle1.26)	41.34 [‡] _(\blacktriangle1.02)	69.04 [‡] _(\blacktriangle1.03)
TRACEERC $_{Llama-2-13b}$ (ours)	125×10^6	71.32 [‡] _(\blacktriangle3.17)	41.18 [‡] _(\blacktriangle0.86)	69.91 [‡] _(\blacktriangle1.90)

observed for the EmoryNLP dataset compared to other benchmark datasets in the ERC domain. This lower accuracy increases the noise level in the provided emotional context, thereby impacting model performance.

Compared to recent prominent methods based on GNNs, Transformers, or PLMs (as shown in Table 7.1), TraceERC exhibits superior performance by harnessing LLMs to effectively integrate multi-contextual information, encompassing dialogue context, speaker personality context, and emotional context. These results underscore the capability of LLMs to provide diverse, complementary information for ERC tasks, emphasizing their robustness and significant potential for broader applications.

Additionally, when evaluated against LLM-based approaches such as MKFM [69], which incorporates multiple external knowledge components, including topical relevance and emotion causes, TraceERC achieves outstanding performance across all three benchmark datasets due to its integration of more distinct and effective speaker-specific attributes and emotional tones. TraceERC also performs competitively with InstructERC, while

Table 5.4: Performance comparison between our proposed method and previous works on the test sets in the *in-context learning* paradigm. The marked star (*) indicates the preprints (unofficial works).

Methods	IEMOCAP	EmoryNLP	MELD
*InstructERC _{Llama2} [32]	2.774	8.36	16.28
*InstructERC _{ChatGPT3.5} [32]	53.38	37.00	65.07
*ChatGPT 3-shot [89]	48.58	35.92	58.35
Baseline _{Llama2} (ours)	49.11	20.77	32.96
TRACEERC _{Llama2} (ours)	52.63 [‡] ($\blacktriangle 3.52$)	23.36 [‡] ($\blacktriangle 2.59$)	34.62 [‡] ($\blacktriangle 1.66$)

bypassing the latter’s two-step procedure that relies on pre-trained LLMs to sequentially identify speaker information for each utterance. Instead, TraceERC directly extracts speaker personality traits by employing prompt-based methodologies, offering improved efficiency and enhanced adaptability for practical applications.

In terms of in-context learning performance (Table 5.4), TraceERC demonstrated clear advantages over the baseline model, achieving F1 scores improvements of 3.52%, 2.59%, and 1.66% on three datasets, respectively. Moreover, our approach significantly outperforms the InstructERC method when leveraging the open-source Llama-2, further validating the effectiveness of our proposed framework. Notably, our results on IEMOCAP are competitive even with proprietary LLMs such as ChatGPT. To the best of our knowledge, we are among the first to effectively evaluate the capabilities of open-source LLMs for in-context learning in the ERC task.

5.5.2 Ablation Study

We conducted an ablation study to assess the contributions of two key components, *speaker descriptions* and *emotional context*, to the overall performance of TraceERC. First, we designed three variants of speaker descriptions to investigate the impact of varying speaker personalities on the ERC tasks. These descriptions were supplemented with a conversational profile as auxiliary information (as detailed in Table 5.5) and evaluated them individually. We then conducted experiments within the TraceERC framework to assess the role of emotional context, evaluating its effectiveness both in isolation and

Table 5.5: Prompting template for generating different types of speaker descriptions and their corresponding output examples.

DetailDesc (<i>Prompting content</i>)	Given this conversation between speakers: $\{\text{conversation content } \mathcal{C}\}$ In overall above conversation, what do you think about the <u>characteristics of speaker $\{\text{speaker identification } s_j\}$</u> ? (Note: provide an answer within 250 words)
(<i>Output example</i>)	SPEAKER_0 seems to be a friendly and playful person. In the conversation, they use the phrase “little village people” to refer to the recipient of the gift, which suggests a sense of affection and familiarity. The use of the word sweet to describe the gift also implies a positive and cheerful attitude. Additionally, SPEAKER_0 seems to be enthusiastic and optimistic, as they express their confidence that the recipient will love the gift. Overall, SPEAKER_0 appears to be a warm and upbeat person who values building positive relationships with others.
EmoTransDesc (<i>Prompting content</i>)	Given this conversation between speakers: $\{\text{conversation content } \mathcal{C}\}$ In overall above conversation, what is speaker $\{\text{speaker identification } s_j\}$ ’s <u>emotional transformations and personality traits</u> ? (Note: provide an answer within 250 words)
(<i>Output example</i>)	SPEAKER_0 starts with a friendly greeting, showing enthusiasm and energy, then expresses confusion and surprise when mistakenly calling the person. This mistake and the quick correction show that the speaker is not paying close attention to the conversation or is easily distracted. Overall, the speaker’s emotional transformation throughout the conversation is minimal, as they maintain a lighthearted and playful demeanor. SPEAKER_0 personality traits include being friendly, playful, and possibly easily distracted or lacking attention to detail.
ConciseDesc (<i>Prompting content</i>)	Given this conversation between speakers: $\{\text{conversation content } \mathcal{C}\}$ In overall above conversation, what do you think about the <u>characteristics of speaker $\{\text{speaker identification } s_j\}$</u> ? (Note: only 3 sentences)
(<i>Output example</i>)	SPEAKER_0 seems to be friendly and curious, as they greet the other speakers and ask about their activities. They also appear to be interested in the holiday candy making process, as they ask about it. Overall, SPEAKER_0 seems to be a sociable and engaged person.

incorporating with the optimal description to assess the role of emotional context (Table 5.6).

Speaker Description.

As shown in Table 5.5, each variant of speaker description offers unique advantages. To generate these personality descriptions, the LLMs are directed to restrict the output by incorporating a “note” regarding the desired output length. These length constraints are designed to guide LLMs toward generating content that is both concise and informative. As illustrated in Figure 5.4, a comparison of the lengths for the three different speaker personalities reveals that the average word count for **ConciseDesc** is the short-

est, followed by DetailDesc, with EmoTransDesc being the longest. DetailDesc refers

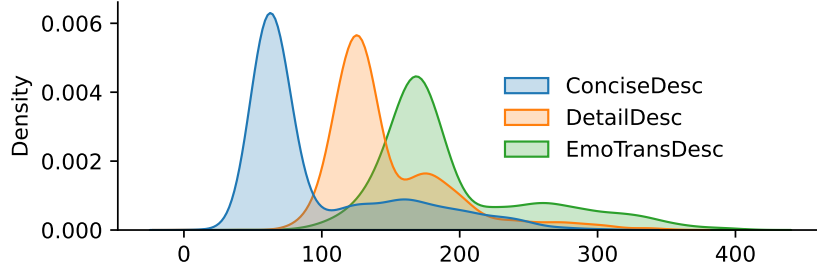


Figure 5.4: Length distribution comparison of different speaker description variants on the MELD dataset.

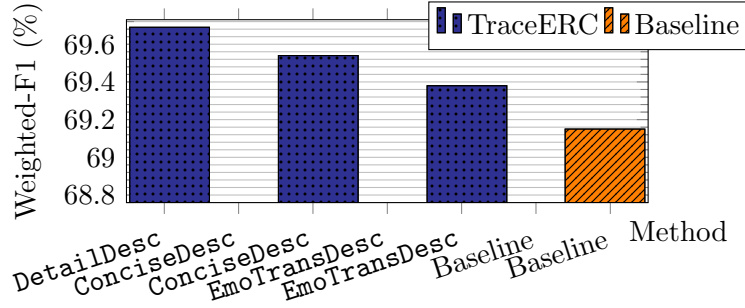


Figure 5.5: Performance comparison of our TraceERC (excluding emotional context information) with the baseline method across four types of speaker descriptions on the MELD development set.

to the extraction of each speaker’s personality traits throughout the dialogue, providing highly informative speaker-specific information within a limit of 250 words. As exemplified in Table 5.5, this description analyzes the speaker’s behavior and emotional tone, individual personality characteristics, and linguistic habits, enabling a deeper discourse understanding of the underlying implicit meanings (e.g., interpreting phrases such as “*little village people*”). Moreover, it is capable of detecting complex human emotions, such as sarcasm, in certain cases, thereby mitigating potential misinterpretations arising from habitual speaking patterns. In contrast, **ConciseDesc** incorporates personality extraction with stringent length constraints. Specifically, the output is limited to a maximum of three sentences, aiming to preserve essential speaker traits while omitting redundant information, thereby achieving a more concise and efficient representation. To enhance understanding of emotional dynamics, we developed **EmoTransDesc**, which prioritizes emo-

tional conveyance in speaker descriptions. This method captures the emotional nuance, speaker’s emotional inertia and transitions, ultimately producing personality descriptions enriched with emotional context. In summary, by employing various prompt designs, we extract diverse facets of speaker-specific information, providing critical insights that fill gaps in existing datasets and enhance the understanding of the conversational context.

As shown in Figure 5.5, our TraceERC consistently outperforms the baseline system when incorporating any type of speaker description. In contrast to prior studies [68], which have demonstrated the importance of various speaker-specific external knowledge sources provided by closed-source LLMs (e.g., ChatGPT), our work highlights the insufficient attention given to the critical factor of speaker-specific personalities. Among the three variants, the detailed speaker description (**DetailDesc**) and its concise version (**ConciseDesc**) achieved the highest performance gains, which we attribute to their emphasis on speaker modeling information. Therefore, these two description types are selected for further experiments within the TraceERC framework, as discussed in the subsequent sections.

To examine the generalization effectiveness of speaker descriptions across various learning scenarios, we conducted ablation experiments in both in-context learning and instruction fine-tuning using two distinct forms of speaker descriptions. As shown in Table 5.6, the inclusion of speaker personality descriptions (without emotional context) consistently improves performance across all datasets compared to the baseline in the instruction fine-tuning setting. Similarly, in the in-context learning, compared to the baseline in this setting, integrating speaker personality descriptions (without emotional context) leads to consistent performance improvements across all datasets. with a notable improvement on IEMOCAP from 49.33% to 53.51% (+4.18%). These findings underscore the robustness and effectiveness of leveraging explicit speaker personality information, demonstrating its robustness across various approaches in consistently improving performance.

Table 5.6: Performance comparison between our proposed method and the baseline model on the development set under the *in-context learning* and *instruction fine-tuning* paradigms. The notation \ddagger denotes the significant difference (t-test) from the baseline in levels $p < 0.05$.

Methods	Desc. type	Emo. context	IEMOCAP	MELD
<i>(In-context learning)</i>				
TRACEERC	ConciseDesc	yes	53.58\ddagger	33.67\ddagger
	ConciseDesc	no	53.51 \ddagger (\blacktriangledown 0.07)	32.07 \ddagger (\blacktriangledown 1.60)
	DetailDesc	no	51.17 (\blacktriangledown 2.41)	32.17 \ddagger (\blacktriangledown 1.50)
Baseline	(none)	no	49.33 (\blacktriangledown 4.25)	30.20 (\blacktriangledown 3.47)
<i>(Instruction fine-tuning)</i>				
TRACEERC	ConciseDesc	yes	69.78\ddagger	70.20\ddagger
	DetailDesc	yes	69.30 \ddagger (\blacktriangledown 0.48)	69.79 (\blacktriangledown 0.41)
	(none)	yes	67.73 \ddagger (\blacktriangledown 2.05)	69.27 (\blacktriangledown 0.93)
	ConciseDesc	no	66.42 (\blacktriangledown 3.36)	69.54 (\blacktriangledown 0.66)
	DetailDesc	no	65.96 (\blacktriangledown 3.82)	69.69 (\blacktriangledown 0.51)
Baseline	(none)	no	65.33 (\blacktriangledown 4.45)	69.15 (\blacktriangledown 1.05)

Emotional Context.

In this section, we conduct ablation studies to rigorously examine the contribution of emotional context information in TraceERC under both fine-tuning and in-context learning settings. As shown in Table 5.6, we performed ablation experiments using two versions of personality descriptions across two datasets. In the instruction fine-tuning scenario, integrating emotional context information consistently led to significant performance improvements across both datasets compared to the baseline. Notably, in the IEMOCAP dataset, incorporating emotional context alone—without any speaker-specific information—resulted in a substantial 2.4% performance gain.

Our findings reveal that integrating the speaker’s personality with emotional context information yields better performance than using either component individually across both datasets. This demonstrates that TraceERC effectively leverages emotional context to establish an overarching emotional atmosphere, and when integrated with speaker personality characteristics and contextual semantics, forms a more comprehensive foundation for emotion recognition. The impact of emotional context is particularly pronounced in the IEMOCAP dataset. When emotional context was incorporated alongside both the

concise (**ConciseDesc**) and detailed (**DetailDesc**) speaker descriptions, performance improved by 3.36% and 3.34%, respectively, compared to using speaker information alone. These enhancements translated to overall gains of 4.45% and 3.97% over the baseline. Similarly, in the MELD dataset, integrating speaker information with emotional context resulted in a 1.05% performance improvement. This further validated the effectiveness of emotional context information and highlighted its strong synergistic effect when combined with speaker-specific features.

In the in-context learning setting, both speaker descriptions and emotional context demonstrated their effectiveness. When integrated, they led to improvements of 4.25% on IEMOCAP and 3.47% on MELD compared to the baseline. These results highlight two key benefits: enhanced the quality of retrieving similar demonstrations (speaker-based few-shot examples), and effective emotional guidance provided by incorporating dynamic emotional context.

To demonstrate the importance of emotional context and further examine its effect on overall performance, we conducted experiments by injecting the *low quality* emotional context predictions (\mathcal{E}^{ec} in Algorithm 1) into the TraceERC framework. To achieve this, we used the predictions (\mathcal{E}^{ec}) generated by earlier checkpoints of the ERC system as modules for emotional context generation and re-evaluated the fine-tuned TraceERC with these inputs (Figure 5.6). The results align well with expectations, as shown by the red line in Figure 5.6, incorporating extracted emotional context (illustrated by the blue line) into the TraceERC consistently and reliably enhances performance.

These results demonstrate three key insights: (i) As the quality of the emotion context improves, TraceERC’s performance continuously increases and remains consistently superior, providing clear evidence of the effectiveness of incorporating emotion context. (ii) The performance gains of TraceERC are more pronounced in the early stages of emotion context enhancement than in later stages, indicating an upper limit to the benefits of emotion context integration. Notably, TraceERC effectively maximizes the potential of incorporating emotional context. (iii) Since the emotion context can be extracted using various models that integrate diverse types of information, our approach is highly flex-

ible, ease of implementation and broadly applicability, as it is compatible with various emotional context extraction models.

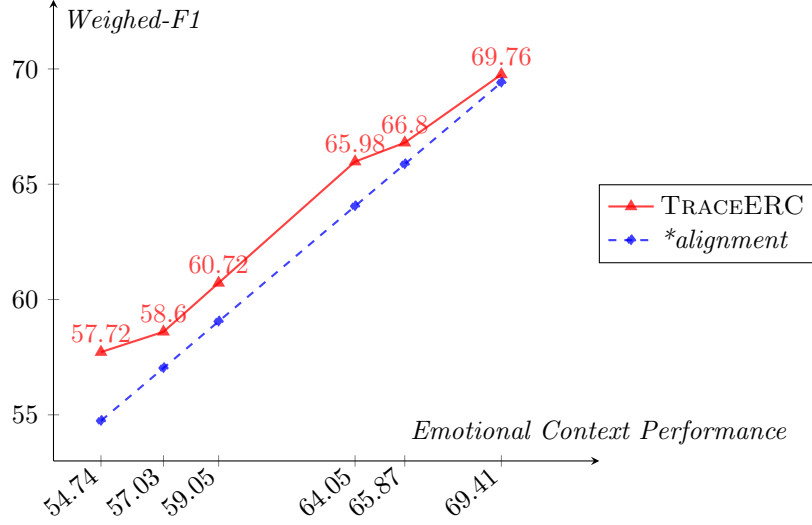


Figure 5.6: Dependence of performance between the emotional context model and our TRACEERC model. The blue dashed line shows the equivalent alignment between the x - and y -axes of the graph.

5.5.3 Analysis of Emotional Context Quality and Its Impact on Performance

We argue that incorporating predicted emotional context serves two main purposes: (1) it helps the model grasp the emotional atmosphere of the conversation, enhancing its adaptability and preparedness for subsequent emotional transitions; and (2) it provides personalized emotional guidance tailored to each target utterance within its dialogue context.

Furthermore, by using predicted emotion labels to construct the emotional context, the model is encouraged to learn from its predictions and allowing for self-correction, and then enhances robustness. In contrast, using golden emotion labels would introduce strong bias and eliminate the model’s opportunity to learn from its mistakes. To validate our approach, we provide results under three settings as shown in Table (5.7): (a) using only predicted labels in the training process, (b) combining predicted and golden labels in the training process, and (c) using only golden labels in the training process. These

comparisons are presented in the table (Table) and have been updated both in the thesis and the corresponding slides.

Table 5.7: Performance comparison among variants of BiosERC on the MELD development set.

Methods	Models	Weighted-F1
a. PREDEC	TraceERC(Llama-2-7B)	69.78
b. GOLDEC	TraceERC(Llama-2-7B)	68.13
c. GOLDEC + PREDEC	Llama-2-chat-70b	68.97

To further examine the role of emotional context, we conducted supplementary experiments to analyze how model performance varies under different levels of emotional context quality. As shown in Fig. 5.7, the x-axis indicates the quality score of emotional context within the local context (ranging from 0 to 1), and the y-axis represents the change in model performance after incorporating emotional context.

The results clearly show that, even when the emotional context is of only moderate quality, its inclusion consistently leads to performance gains over the baseline without emotional context. This demonstrates that emotional context serves as a robust and valuable signal for improving emotional understanding in dialogue. Notably, the positive impact holds across varying quality levels, emphasizing its generalizability and resilience.

5.5.4 Error Analysis

In this section, to thoroughly inspect the current limitations of TraceERC and compare its performance against the baseline model, we depict confusion matrices to visually evaluate the predictions on two settings: emotional confusion (Figures 5.8, 5.9), and bipolar sentiment confusion (Figures 5.10, 5.11). To obtain sentiment confusion matrices, we map predicted emotional labels to bipolar sentiment categories using heuristic rules. For instance, *happy* and *excited* are mapped to *positive*, while *frustrated*, *angry*, and *sad* are categorized as *negative*.

Figure 5.8 and Figure 5.9 clearly illustrate that: (i) TraceERC outperforms the baseline in generalizing emotion recognition accuracy across all emotion labels within both

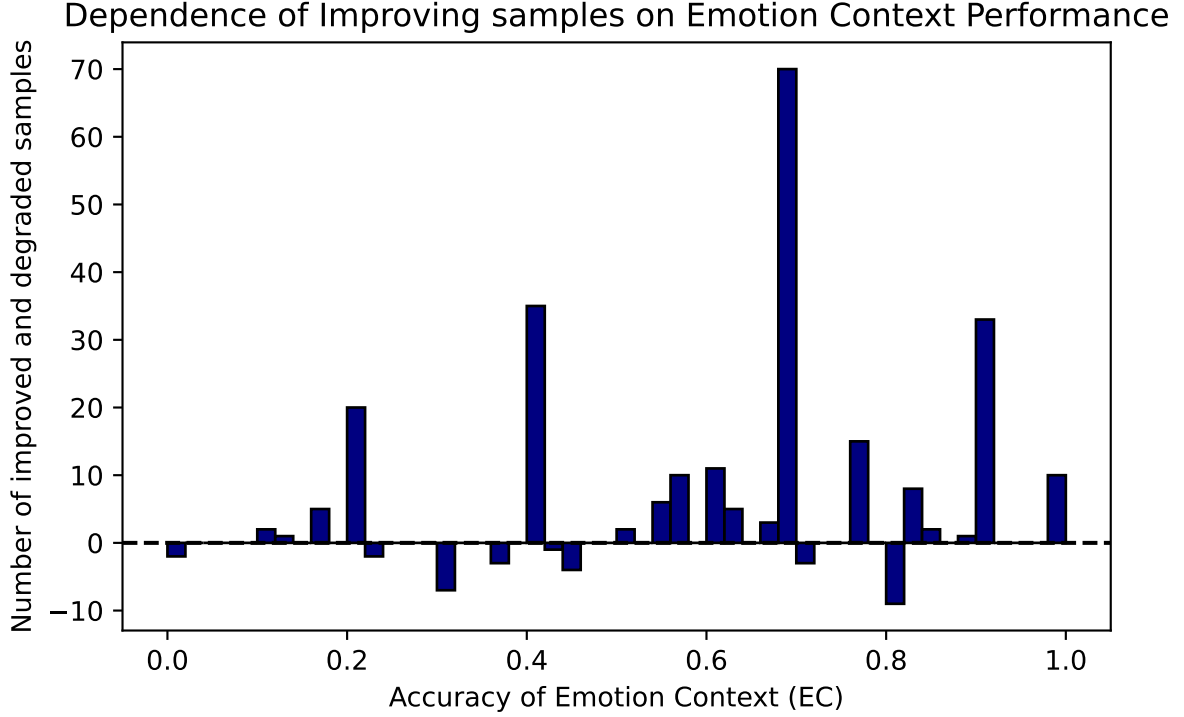


Figure 5.7: Impact of emotional context quality on model performance in the local context setting

IEMOCAP and MELD dataset. (ii) Notably, the model shows significant improvements for highly salient emotions. On the IEMOCAP dataset, accuracy for happy, angry, and excited increases by 22.6%, 4%, and 3.6%, respectively. On the MELD dataset, recognition accuracy for fear, sadness, and disgust sees substantial improvements of 8.2%, 8.1%, and 9.2%, respectively (in diagonal of Figure 5.8). (iii) It substantially mitigates confusion between similar emotions. The most challenging misclassification in the baseline is mistaking happy for excited and is notably reduced from 41.3% to 21.3%, while the misclassification of angry as frustrated decreases from 19.4% to 16.9%. Similarly, in the MELD dataset, the confusion between angry and frustrated drops from 15.6% to 8.4%. These findings substantiate that by incorporating the static personality traits of speakers and dynamic emotional context, TraceERC not only captures speaker emotional fluctuations with greater precision but also leverages emotional context to better understand the relationship between labels and their corresponding utterances. This enables the model to distinguish subtle differences between emotion labels, thereby enhancing its ability to

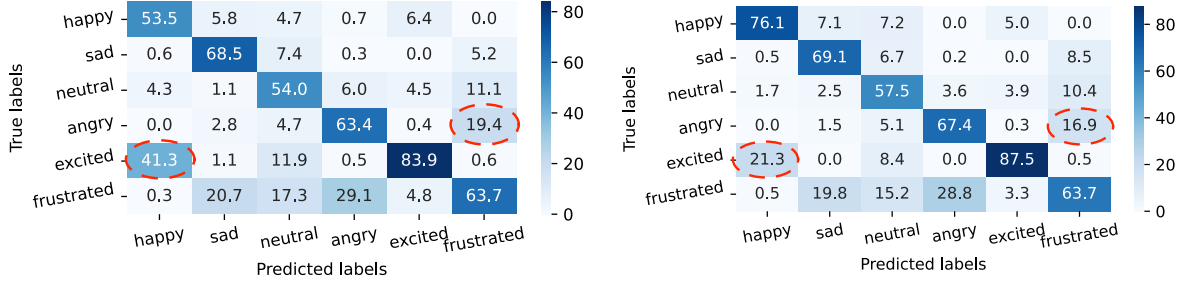


Figure 5.8: Comparison of confusion matrices between our TRACEERC (right) and the baseline system (left) on the IEMOCAP dataset.

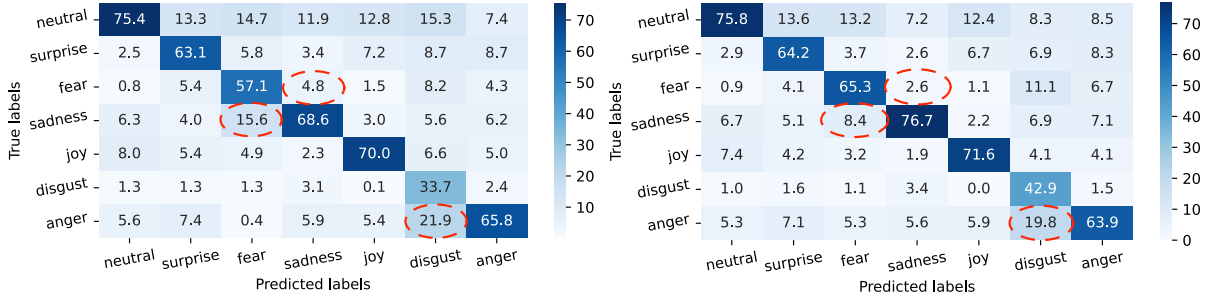


Figure 5.9: Comparison of confusion matrices between our TRACEERC (right) and the baseline system (left) on the MELD dataset.

capture emotional nuances effectively.

In terms of bipolar sentiment analysis, our approach significantly improves the model’s ability to distinguish between positive and negative emotions. On the IEMOCAP dataset (as shown in Figure 5.10), the accuracy of correctly identifying positive and negative emotions increased by 1.7% and 1%, respectively, while the probability of completely confusing positive and negative emotions decreased by 1.2%. Similarly, on the MELD dataset (as shown in Figure 5.11), the accuracy of identifying positive and negative emotions improved by 0.6% and 2.9%, respectively, with a 2.0% reduction in the probability of total confusion between these emotions. These results highlight the tangible benefits of incorporating speaker-specific descriptions and emotional context information. Even in cases where our approach may not fully capture certain specific emotions within individual sentences, it deepens the model’s understanding of the conversational context, enabling it to recognize the emotional tone of the sentence and mitigate the issue of confusion between positive and negative classifications. This helps reduce sentiment-level misclassification

and demonstrates the practical utility of our approach for real-world emotion recognition tasks.

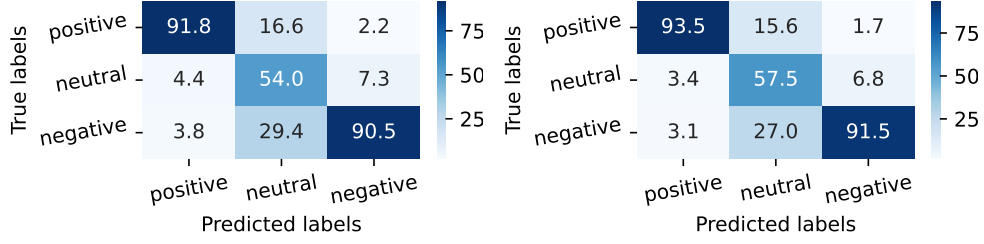


Figure 5.10: Comparison of confusion matrices between our TRACEERC (right) and the baseline system (left) after mapping to sentiment labels on the IEMOCAP dataset.

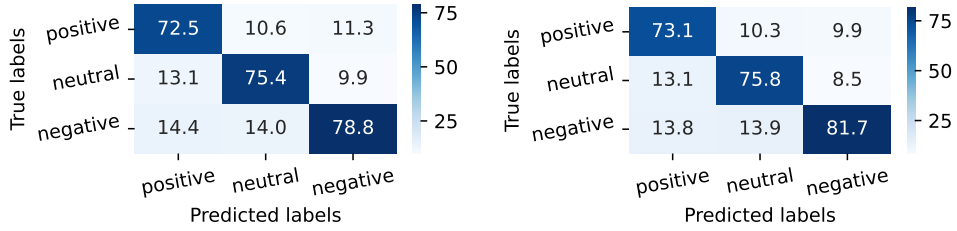


Figure 5.11: Comparison of confusion matrices between our TRACEERC (right) and the baseline system (left) after mapping to sentiment labels on the MELD dataset.

5.5.5 Case Study

This section presents a detailed analysis of representative cases from the MELD and IEMOCAP datasets, highlighting the significance of static speaker personality and the impact of emotional context. As illustrated in Tables 5.8, 5.9, 5.10, we compare the performance of TraceERC, TraceERC_{noEC} (without emotional context), and the baseline system. First, speaker personality offers crucial speaker-specific information, enabling the model to interpret utterances accurately despite surface semantics. In the first example *Ses03M_impro08*, our model effectively extracts that “Robert” starts the conversation *calmness and professionalism*, and attempts to remain *polite*. Embedding speaker-specific descriptions allows our method to recognize that Robert’s professional role necessitates polite interactions with clients. This enables the model to correctly interpret Robert’s

Table 5.8: Case study of improvement examples collected from two datasets. The red and green labels refer to the incorrect and correct prediction by the models, respectively.

Idx	IEMOCAP: Conversation Ses03M_impro08		Label	Baseline	TraceERC _{noEC}	TraceERC
	ROBERT	JENNIFER				
d ₀	ROBERT starts the conversation calmly and professionally . He attempts to remain polite , but his tone ... Ultimately, ROBERT's frustration boils over, and he makes a veiled threat to JENNIFER...					
d ₁	JENNIFER's emotional transformation during the conversation is quite drastic , starting off frustrated and gradually becoming more aggressive and hostile towards ROBERT. Her personality comes across as assertive and demanding , as she is determined to get her phone service restored and her frustration mounts. Despite ROBERT's attempts to placate her, JENNIFER remains unsatisfied and ultimately ends the conversation in a confrontational manner.					
u41	I'm really sorry ma'am. There's not—a lot we can do for it.		neutral	frustrated	neutral	neutral
u42		That you better be sorry.	angry	angry	angry	angry
u43		Well you're not helping me at all. God this is so frustrating.	frustrated	angry	angry	angry
u44	What I can do for you, if you want to make a call right now, I could contact you through our computer...		neutral	frustrated	neutral	neutral
u45		What? No. That's way too complicated.	frustrated	angry	angry	frustrated
u46		Look—you need to fix my phone. I don't care what you have to do.	frustrated	angry	angry	frustrated
u47	I'm sorry it has to do with an electrical power outage and stuff.		neutral	neutral	neutral	neutral
u50		No, no, no. I didn't choose to live here. I work there. Okay? Yeah. work work work	frustrated	frustrated	frustrated	frustrated
u51	Well I'm sure maybe there's a phone at work you could use.		neutral	neutral	neutral	neutral

utterances (e.g., #41 and #44) as expressing a *neutral* emotion, even though their surface semantics may appear negative. Second, the integration of emotional context and speaker personality descriptions effectively alleviates confusion between similar emotions. In the first example *Ses03M_impro08*, utterances #45 and #46 were initially ambiguous between “angry” and “frustrated”. Although Jennifer’s personality is identified as *quite drastic, frustrated and gradually becoming more aggressive and hostile towards ROBERT*, making her overall emotion clearly negative, distinguishing similar emotions remains challenging. However, the emotional context provides dynamic examples of frustration and anger (e.g., frustrated in utterance #50 and angry in utterance #43), offering the model concrete references aligned with both Jennifer’s personality and the conversation’s actual emotional states. The third example (#5 in MELD conversation 1143) further demonstrates this. This facilitates a nuanced analysis of the emotional undertones specific to each speaker, enabling a deeper understanding of the semantic implications of emotion labels in varying conversational contexts and different speakers. Third, handling incomplete and multi-speaker dialogues, in the second example (MELD dialogue 1146), the characteristics of the MELD dataset become apparent, including incomplete dialogue content,

Table 5.9: Case study of improvement examples collected from two datasets. Red and green labels indicate incorrect and correct model predictions, respectively.

Idx	MELD Conversation 1146			Label	Baseline	TraceERC _{noEC}	TraceERC
	Speaker_0	Speaker_1	Speaker_2				
d ₀	SPEAKER_0 begins the conversation with a confident statement about their identity , ...SPEAKER_0 remains engaged in the conversation and open to feedback, indicating a willingness to learn and grow.						
d ₁	SPEAKER_1's emotional transformation during the conversation is quite drastic , starting from a playful and lighthearted tone to a more serious and frustrated one. This transformation is evident in the way their language changes, ... SPEAKER_1's personality comes across as confident and assertive, as they express their strong opinions and emotions without hesitation.						
d ₂	SPEAKER_2's emotional transformation during the conversation is minimal, as they remain uncertain and insecure about their acting abilities. Their personality is revealed to be humble and self-doubting , as they question their own talent despite praise from others. Throughout the conversation, SPEAKER_2's language use suggests a desire for validation and reassurance, indicating a reliance on external approval for their self-worth.						
u ₀	I am.			neutral	neutral	neutral	neutral
u ₁		Joey, what are you talking about? You're a terrific actor.		neutral	surprise	neutral	surprise
u ₂		You really think so?		surprise	neutral	surprise	surprise
u ₃		Ugh, how can you even ask that question?!		surprise	neutral	surprise	surprise

Table 5.10: Case study of improvement examples collected from two datasets. Red and green labels refer to the incorrect and correct model predictions, respectively.

Idx	MELD Conversation 1143		Label	Baseline	TraceERC _{noEC}	TraceERC
	Speaker_0	Speaker_1				
d ₀	SPEAKER_0 emotional transformation during the conversation is quite drastic , starting from a playful and happy tone to a more serious and defensive one,..., but also insecure and defensive when faced with the possibility of expressing their true feelings. Despite their attempts to downplay their emotions, SPEAKER_0's genuine feelings are revealed through their actions and reactions to SPEAKER_1's words.					
d ₁	SPEAKER_1 transforms from a playful and teasing tone at the beginning to a more serious and emotional tone towards the end, showing a sensitive side. Throughout the conversation, SPEAKER_1 displays a persistent and determined personality...					
u ₃	Look, Monica		neutral	neutral	neutral	neutral
u ₄		Look!	surprise	neutral	anger	joy
u ₅	This is not going to work.		anger	neutral	sadness	anger
u ₆		I bet this will work!	joy	joy	joy	joy
u ₇	You are so great! I love you!		joy	joy	joy	joy
u ₈		What?	surprise	neutral	surprise	surprise
u ₉	Nothing! I said, I said "You're so great" and then I just, I just stopped talking!		fear	anger	joy	fear

multiple speakers, and simplistic interactions, as it is derived from TV show excerpts. Sentiment recognition in text-only settings is particularly challenging in such contexts due to the absence of visual and speech cues. However, our speaker description and emotional context effectively extract additional relevant information from the available context, enhancing recognition efficiency for short dialogues with limited contextual information. Fourth, discerning emotional polarity in ambiguous utterances. In the final example (MELD dialogue 1143), it is evident that incorporating speaker personality information and emotional context enables the model to discern the negative or positive inclination of conversational utterances, as demonstrated in sentences #5, #8, and #9. Fifth, recognizing ultra-short utterances. Our approach proves particularly effective in

recognizing ultra-short utterances within dialogues, as seen in sentences #4 (Look!) and #8 (What?), where traditional methods struggle due to the lack of explicit emotional cues. The inclusion of the speaker’s personality and emotional context provides essential interpretative cues, improving emotion classification in such cases.

5.6 Chapter Summary

In this section, we proposed TraceERC, a novel contextual emotion recognition framework that holistically integrates utterance content, explicit speaker personality, and emotional state context. By pioneering the extraction of previously overlooked speaker-specific information through Large Language Models (LLMs), and by providing emotion-labeled contextual exemplars for atmosphere modeling and dynamic emotional guidance, our approach significantly enhances emotion recognition accuracy across diverse conversational contexts. Experimental results on three well-known datasets demonstrate that our proposed method significantly improves baseline models and achieves competitive results with SOTA methods. The framework is not only effective in modeling speaker personality and emotional transitions within conversations but also simple and flexible, making it readily applicable to a broad range of user modeling and sentiment analysis tasks in natural language processing.

Chapter 6

Cross-Lingual Multi-Label Emotion Detection Using Generative Models

6.1 Motivation and Objectives

With the rapid expansion of social media and the increasing prevalence of global digital communication, online platforms have become a dominant channel for information exchange and opinion sharing [51]. With users from diverse linguistic and cultural backgrounds increasingly expressing their opinions through online comments, there is a growing need for robust cross-lingual sentiment detection systems [51]. Consequently, multi-lingual sentence-level sentiment analysis has emerged as a critical task for tracking public sentiment across different languages and regions [74]. Sentiment analysis remains one of the most extensively researched areas in NLP. In text-based emotion recognition, it is common for a single sentence to convey multiple emotions with varying intensities [12]. This challenge is further compounded by the limited availability of annotated training data, particularly for low-resource languages, and the insufficient representation of these languages in existing pre-trained language models.

To address these challenges, this study explores the integration of PLMs such as RoBERTa and LLMs for multi-label sentiment analysis across both high-resource languages (e.g., English and Chinese) and low-resource languages (e.g., various African languages). We

formulated the multi-label emotion recognition as a text generation task and leverage the capabilities of multilingual PLMs models to enhance both semantic understanding and emotional tone recognition, especially in low-resource languages. Furthermore, to tackle the multi-label classification challenge, we propose two complementary methods: the *pair-wise* method and the *base* method. Our findings further demonstrate that training the model on a combined multilingual dataset yields better performance than training on monolingual data, indicating the benefits of cross-lingual knowledge transfer.

We participated in both tracks of SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection. Extensive evaluations, including ablation studies, confirm the effectiveness of our proposed approach in addressing both sub-tasks: multi-label emotion classification (Track A) and emotion intensity detection (Track B). In Track A, our system ranked among the top four across ten languages, achieving first place in Hindi. In Track B, it placed within the top five for seven languages, further demonstrating the robustness and generalizability of our method. Overall, our approach exhibits strong cross-task performance and offers a scalable and practical solution for emotion detection in low-resource language settings, where such capabilities are particularly critical.

6.2 Related Works

Sentence-level sentiment analysis (SLSA) has undergone substantial advancements propelled by deep learning and the emergence of multilingual sentiment detection techniques. Early research in this area primarily focused on extracting handcrafted sentiment features such as n-grams [67], sentiment lexicons, and rule-based heuristics [6] to enhance SVM-based classifiers [29] and feed-forward into convolutional or recurrent neural networks [6, 49]. These methods depended on static word embeddings, which constrained their capacity to capture complex linguistic phenomena, including long-range dependencies and cross-lingual variations.

To overcome these limitations, researchers adopted Transformer-based PLMs such as BERT and T5 [59], which more effectively capture fine-grained emotional representations

[93, 38] by encoding richer semantic information. In multilingual sentiment analysis domain, architectures like mT5 [79] and XLM-RoBERTa [9] learn cross-lingual embeddings that have become the de facto standard for sentiment analysis across multiple languages [24]. The field has witnessed significant progress with the advent of LLMs, such as LLaMA 2 [66], which have driven major breakthroughs across a wide range of NLP tasks [70, 61]. These models demonstrate strong zero-shot and few-shot learning capabilities, making them highly adaptable to novel sentiment analysis tasks [46]. In the domain of ERC, LLMs have been effectively employed with prompt-based techniques to extract latent supplementary knowledge from text, which is then injected to support more accurate emotion recognition [78]. In the broader NLP landscape, a range of techniques—such as fine-tuning, prompting, transfer learning, and domain adaptation—have proven essential for adapting LLMs to sentiment analysis tasks across various domains and languages.

Despite these advancements, most PLMs are primarily trained on English corpora, limiting their effectiveness in multilingual sentiment analysis unless additional fine-tuning is applied to optimize performance across diverse linguistic settings [85]. To address this limitation, numerous studies have explored the use of LLM embeddings for sentiment classification in low-resource languages, aiming to assess their cross-lingual adaptability [10, 50]. In this work, we leverage BERT-based multilingual models to extend multi-label sentiment classification tasks, facilitating more effective knowledge transfer across languages. Furthermore, by incorporating LLMs, we introduce a *pairwise emotional recognition* strategy that focuses on one label at a time, enabling the model to more accurately capture both emotional intensity and sentiment polarity within each sentence. Additionally, we reformulate the multi-label classification task as a text generation problem, thereby enhancing the model’s flexibility and generalization capabilities across a wider range of NLP tasks.

6.3 Methodology

In this work, the target task focuses on multilingual emotion perception, which aims to identify the emotion that a majority of people would associate with the speaker based on a given sentence or short text snippet. Given a text input x , the objective is twofold: (i) to retrieve all relevant emotional labels expressed (y_e) in the text (Track A: multi-label emotion classification), and (ii) to estimate the intensity (y_i) of each emotion class (Track B: emotion intensity regression).

6.3.1 System Overview

To address the challenges of cross-lingual emotion recognition, we leverage the strengths of pre-trained multilingual models. Our system is built upon two core architectures: (1) fine-tuned BERT-based classification models [14], and (2) instruction-tuned generative LLMs, inspired by recent state-of-the-art approaches in emotion recognition [78]. To effectively handle multi-label classification, we propose two complementary strategies: (1) the *base* strategy, which directly maps the input text to a set of associated emotion labels; and (2) the *pairwise* strategy, which independently evaluates the relationship between the input and each candidate label, enabling more fine-grained emotion detection.

$$\text{base: } \mathbb{P}_A(\{y_e\} \mid x) \quad \mathbb{P}_B(\{\langle y_e, y_i \rangle\} \mid x) \quad (6.1)$$

$$\text{pairwise: } \mathbb{P}_A(\{0, 1\} \mid x, y_e) \quad \mathbb{P}_B(y_i \mid x, y_e) \quad (6.2)$$

Here, x denotes the input text; \mathbb{P}_A and \mathbb{P}_B represent the probability models used for Track A and Track B, respectively; and y_e and y_i refer to the emotional label and its corresponding emotional intensity, both drawn from a predefined label set.

6.3.2 Methods

BERT-based method. As a baseline, we implement a BERT-based multi-label classification model. Specifically, we append fully connected layers with non-linear activa-

tion functions—sigmoid (σ) and tanh—on top of the BERT architecture. These layers transform the representation of the [CLS] token [14] from the hidden representation to an output space aligned with the number of emotion labels. Formally, the computation is defined as:

$$h^{CLS}, h^{words} = \text{BERT}(x) \quad (6.3)$$

$$h^{out} = \sigma(\tanh(h^{CLS} \cdot W^h) \cdot W^o) \quad (6.4)$$

During fine-tuning, the learnable weights (W^*) are optimized using a cross-entropy loss function to maximize the log-likelihood of the correct label assignments based on the annotated training data.

LLM-based method. To leverage the advanced language understanding capabilities of LLMs [66], we adopt an instruction prompting approach (highlighted in blue in Table 6.1) to explicitly convey the task objective. Our methodology follows the instruction fine-tuning paradigm, optimizing a *causal language modeling* objective to train the LLM to generate emotional label text, which is highlighted in red in Table 6.1.

$$s = \text{instruction-prompting}(x, y) \quad (6.5)$$

$$\mathbb{P}(s) = \prod_{z=1}^{|s|} \mathbb{P}(s_z | s_0, s_1, \dots, s_{z-1}) \quad (6.6)$$

Here, s and x denote sequences of tokens, while z represents the token indices within the prompting input (see Table 6.1). To improve training efficiency, we employ LoRA [22], a lightweight fine-tuning technique that reduces the number of trainable parameters. The fine-tuned LLM is trained to model the conditional distribution of emotional labels (or emotional intensities) given the prompt s . During inference, the emotional label y , which is excluded from the input prompt, is generated by the LLM.

As described in the overview section, we implement two modeling strategies to solve this task: *base* and *pairwise* methods. Both approaches employ the same training techniques across tracks A and B. Examples of generated outputs for both tracks are presented in

<i>system</i>
You are an expert in analyzing the emotions expressed in a natural sentence. The emotional label set includes {anger, fear, joy, sadness, surprise}. Each sentence may have one or more emotional labels, or none at all.
<i>user</i>
Given the sentence: “{input text: x ”, which emotions are expressed in it?
<i>assistant</i>
{emotional label in text: y_e or $\langle y_e, y_i \rangle$ }

Table 6.1: Instruction prompting using the *base* template (track A).

Task	Strategy	Input	Output	Output example
Track A	base	x	$\{y_e\}$	“disgust, sadness”
Track B	base	x	$\{\langle y_e, y_i \rangle\}$	“moderate degree of anger, low degree of sadness”
Track A	pairwise	x, y_e	$\{0, 1\}$	“yes”
Track B	pairwise	x, y_e	y_i	“moderate”

Table 6.2: Examples of output format for text generation.

Table 6.2. Detailed examples for each track are provided in Table 6.3 and Table 6.4.

<i>system</i>
You are an expert in analyzing the emotions expressed in a natural sentence. The emotional label set includes {anger, fear, joy, sadness, surprise}. Each sentence may have one or more emotional labels, or none at all.
<i>user</i>
Given the sentence: “bro dont do this to us”, which emotions and their corresponding intensities are expressed in it?
<i>assistant</i>
fear

<i>system</i>
You are an expert in analyzing the emotions expressed in a natural sentence. The emotional label set includes {anger, fear, joy, sadness, surprise}. Each sentence may have one or more emotional labels, or none at all.
<i>user</i>
Given the sentence: “I could not unbend my knees.”, is the emotion anger expressed in it?
<i>assistant</i>
No

Table 6.3: Instruction prompting templates for the *base* (top) and *pairwise* (bottom) strategies in Track A.

6.4 Experiment

Dataset. To evaluate the effectiveness of our proposed methods, we employ the original emotional datasets released by the SemEval Task 11 organizers. This dataset com-

<i>system</i>
You are an expert in analyzing the emotions expressed in a natural sentence. The emotional label set includes {anger, fear, joy, sadness, surprise}. Each sentence may have one or more emotional labels, or none at all.
<i>user</i>
Given the sentence: "A penny hit me square in the face.", which emotions and their corresponding intensities are expressed in it?
<i>assistant</i>
moderate degree of anger, low degree of sadness

<i>system</i>
You are an expert in analyzing the emotions expressed in a natural sentence. The emotional label set includes {anger, fear, joy, sadness, surprise}, with three levels of intensity: low, moderate, and high. Each sentence may have one or more emotional labels, or none at all.
<i>user</i>
Given the sentence: "Totally creeped me out.", what is the intensity of the emotion fear expressed in it?
<i>assistant</i>
high

Table 6.4: Instruction prompting template for the *base* (top) and *pairwise* (bottom) strategies in Track B.

prises three subsets—training, development, and test—distributed across two competition phases: development and test. For improved generalization, we consistently reserve 10% of the training data from each language as an internal development set. This held-out portion is used for hyper-parameter tuning and model checkpoint selection. To accommodate the multilingual nature of the task, we define two training configurations: (1) *separatedspacelangs*, where an independent model is trained for eachspace language, and (2) *mixedspacelangs*, where a unified model is trained jointly on all languages.

Evaluation Metric. Following the competition guidelines, the evaluation metric for Track A is the macro-averaged F1 score, while for Track B, it is the Pearson correlation coefficient between the predicted and gold-standard labels.

Experimental Setup. All experiments are implemented using widely adopted frameworks such as PyTorch and HuggingFace. For pretrained LLMs, we primarily experiment with XLM-RoBERTa-Large, Llama2 (7B-13B), and Qwen2.5 (14B–32B). Hyperparameters are set with a learning rate of 3×10^{-4} , optimized using the AdamW algorithm, and training proceeds for 5–6 epochs.

Table 6.5: Results of Sub-task A. For a fair comparison, the *average* column is computed based on all languages except for *orm*, *ron*, *ptbr*, and *ptmz*, as these languages are missing in some settings. The red-highlighted values indicates the best-performing configurations used for the final test submission. The asterisk (*) marks results obtained during the post-evaluation phase.

Model	Strategy	Data	afz	amh	arq	ary	chn	deu	eng	esp	hau	hin	ibo	kin	mar	orm
<i>(development)</i>																
Qwen 32b	pairwise	separated langs	0.5143	0.5049	0.6574	0.5242	0.6909	0.7187	0.8189	0.8366	0.5724	0.8694	0.5049	0.4274	0.9507	-
Qwen 32b	base	separated langs	0.4610	-	-	-	-	-	0.8054	-	-	-	-	-	-	-
Qwen 32b	base	mixed langs	0.5140	0.557	0.64	0.537	0.732	0.677	0.751	0.839	0.57	0.899	0.509	0.477	0.959	0.478
Qwen 14b	base	mixed langs	0.4320	0.594	0.588	0.567	0.643	0.691	0.743	0.835	0.606	0.887	0.498	0.454	0.924	0.503
xml-roberta	base	mixed langs	0.5070	0.66	0.607	0.548	0.623	0.654	0.703	0.786	0.687	0.855	0.488	0.328	0.948	0.513
JNLP (<i>test</i>)			0.5925	0.6767	0.6407	0.609	0.6805	0.6990*	0.8036	0.8303	0.6504	<i>0.9257</i>	0.5404	0.4289	0.878	0.573
<i>(development)</i>																
Model	Strategy	Data	pcm	ptbr	ptmz	ron	rus	som	sun	swa	swe	tat	tir	ukr	vmw	yor
Qwen 32b	pairwise	separated langs	0.6202	0.6407	0.5161	0.7548	0.8809	0.3571	0.5307	0.2658	0.5915	0.6282	0.4581	0.6761	0.1265	0.4554
Qwen 32b	base	separated langs	0.611	-	-	0.7230	-	-	-	-	-	-	-	-	-	-
Qwen 32b	base	mixed langs	0.638	0.546	0.571	-	0.902	0.416	0.557	0.332	0.509	0.72	0.429	0.639	0.114	0.355
Qwen 14b	base	mixed langs	0.622	0.576	0.553	-	0.895	0.394	0.51	0.319	0.494	0.764	0.485	0.64	0.19	0.348
xml-roberta	base	mixed langs	0.574	0.502	0.579	-	0.876	0.499	0.539	0.348	0.501	0.692	0.5	0.594	0.074	0.198
JNLP (<i>test</i>)			0.6343	0.6184	0.4535	0.7787	0.8912	0.4965	0.4596	0.2949	0.6186	0.7223	0.4849	0.6873*	0.2261	0.3608

Table 6.6: Results of Sub-task B. The meanings of the denotations and colors are the same as in Table 6.5.

Model	Strategy	Data	amh	arq	chn	deu	eng	esp	hau	ptbr	ron	rus	ukr	Average
<i>(development)</i>														
Llama2-13b	pairwise	separated langs	-	0.4411	0.73857	0.6197	0.8207	0.7221	0.5691	0.4938	0.691	0.8719	0.6229	0.6757
Qwen-32b	pairwise	separated langs	0.5433	0.6147	0.75	0.6793	0.8101	0.7715	0.6143	-	0.7245	0.9051	0.6428	0.7234
Qwen-32b	base	mixed langs	0.542	0.566	0.711	0.658	0.802	0.761	0.595	0.718	-	0.898	0.659	0.7063
Qwen-32b	pairwise	mixed langs	0.563	0.627	0.727	0.705	0.787	0.779	0.665	0.6	-	0.906	0.694	0.7363
JNLP (<i>test</i>)			0.6038	0.5873	0.6589	0.725	0.8129	0.7747	0.6496	0.6512	0.7055	0.9074	0.6719	0.7044

6.5 Result Analysis

We evaluate our proposed methods and their variants on the development set to identify the best-performing configurations for each language. The selected configurations, which are used for the final test submissions, are highlighted in red in Tables 6.5 and 6.6.

6.5.1 Track A: Multi-label Emotion Detection.

Development Result. As shown in Table 6.5, we evaluate both the *base* and *pairwise* methods on Qwen-32B, Qwen-14B, and RoBERTa models. The results indicate that the Qwen models employing the *pairwise* strategy consistently achieve the highest overall performance. However, in datasets where most samples contain zero or only one emotion label, the *base* method outperforms the *pairwise* approach. This suggests that while the *pairwise* strategy is inherently better suited for multi-label emotion recognition, it may be less effective in scenarios characterized by sparse label distributions. Furthermore, for low-resource languages, LLMs tend to exhibit inferior performance, whereas the RoBERTa-based models demonstrate greater robustness.

Test Result. Our approach achieves competitive performance across languages: 4th place for Chinese (CHN), Spanish (ESP), Nigerian Pidgin (PCM), and Brazilian Portuguese (PT-BR); 3rd place for Algerian Arabic (ARQ), Egyptian Arabic (ARY), Romanian (RON), and Russian (RUS); and top ranks of 2nd and 1st for Swedish (SWE) and Hindi (HIN), respectively. These results demonstrate the generalizability and efficiency of our methods.

6.5.2 Track B: Emotion Intensity.

The results of Track A highlight the superior performance of LLMs compared to the XLM-RoBERTa model, motivating us to focus primarily on LLM-based experiments rather than XLM-RoBERTa in this track.

Development Result. As shown in Table 6.6, we conducted experiments employing both the *base* and *pairwise* strategies on the LLaMA 2 and Qwen-32B models. The results

demonstrate that the Qwen-32B model consistently outperforms LLaMA 2, and that the *pairwise* strategy achieves superior performance compared to the *base* method.

Test Result. In the final evaluation of Track B across 11 languages, our system ranks 3rd for Ukrainian (UKR) and Algerian Arabic (ARQ), 4th for Romanian (RON), and 5th for Russian (RUS), Brazilian Portuguese (PT-BR), English (ENG), and German (DEU). With top-five rankings in seven languages, these results confirm the effectiveness and cross-lingual generalizability of our approach.

6.5.3 Result Analyses

Strategies Comparison: Base vs. Pairwise Strategies. To gain a comprehensive understanding of the two modeling strategies, we analyze the distribution of improved samples—measured by the F1 score at the instance level—across four languages: English, Swedish, Chinese, and Kinyarwanda. The results are illustrated in Figure 6.1 for Track A and Figure 6.2 for Track B. Our findings reveal that the *pairwise* strategy predominantly enhances performance on samples from languages that frequently express multiple emotions per sentence, notably English and Swedish. In contrast, for languages or datasets characterized by a limited number of emotional labels per instance—such as Chinese and Kinyarwanda, where samples typically contain zero to two emotions—the *base* strategy exhibits a clear advantage. We attribute this phenomenon to the fact that the *pairwise* strategy independently evaluates each emotion label, rendering it more sensitive to label imbalance. This characteristic can hinder its performance in low-variance settings. On the other hand, the *base* strategy jointly predicts all emotion labels for a given input, which appears to be advantageous in contexts where only a few emotions are typically present.

Emotional Type. To assess the model’s effectiveness in capturing emotional intensity across different emotions, we plotted the distribution of emotional labels along with their corresponding intensity values, as shown in Figure 6.3. These experiments were conducted by aggregating data from all languages, enabling us to examine the corre-

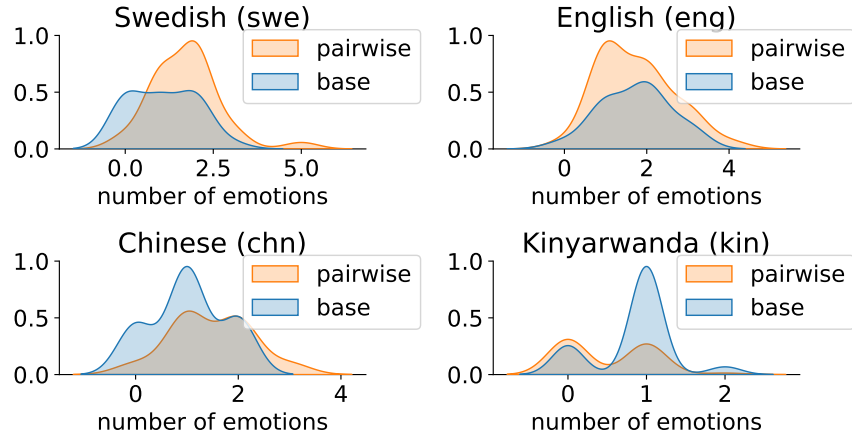


Figure 6.1: Distribution of improved samples between *base* and *pairwise* strategies, categorized by the number of emotions (track A).

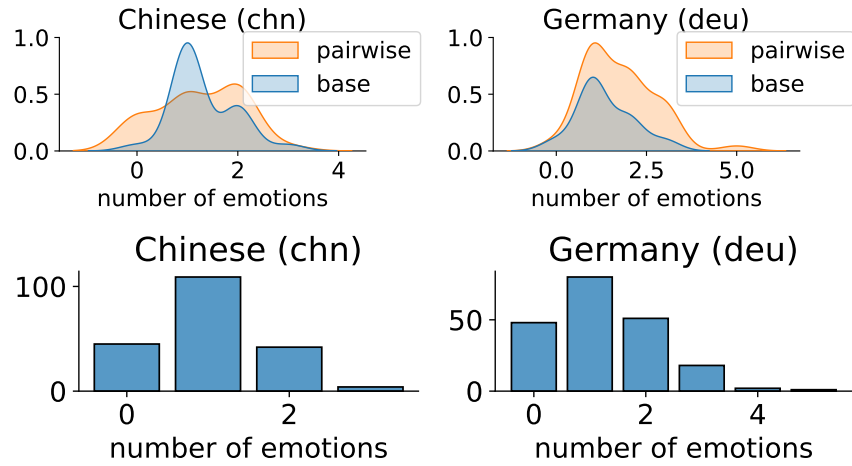


Figure 6.2: Distribution of improved samples between *base* and *pairwise* strategies, categorized by the number of emotions (track B).

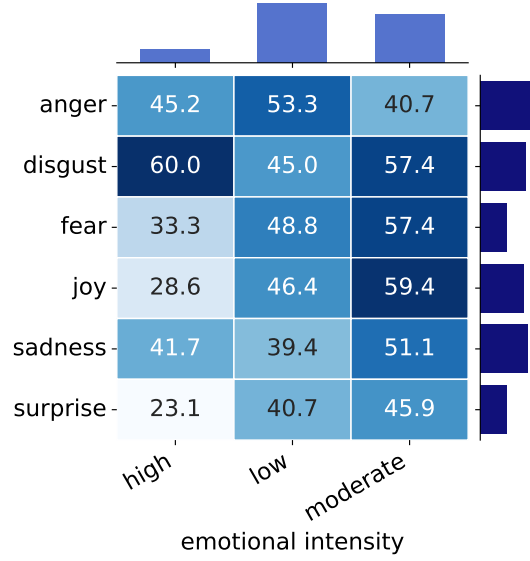


Figure 6.3: Overall performance of the pairwise strategy across all emotional labels and languages (Track B).

lation between emotion type, label frequency, intensity, and model performance. Our analysis revealed that classes with limited training data, such as high-surprise and high-joy, generally exhibited lower performance in our system. Conversely, the predominant class, *disgust*, achieved the highest performance compared to other emotional categories, including surprise—especially high-surprise.

Mixed languages. In both Sub-tasks A and B, the mixed-language setting—where a single model is fine-tuned jointly across all languages—demonstrates superior performance compared to training separate models for each language (Tables 6.5, 6.6). We attribute this improvement to a more balanced distribution of emotion categories within the multilingual training corpus, as well as the model’s enhanced ability to generalize across diverse linguistic contexts. These findings underscore the advantages of leveraging cross-lingual transfer learning in multilingual emotion recognition tasks.

6.6 Chapter Summary

In this work, we presented a multilingual emotion recognition system developed for SemEval-2025 Task 11. Our system demonstrates competitive performance, ranking among the top

teams across multiple languages. To address the challenges of multilingual and multi-label emotion recognition, we designed two complementary architectures—BERT-based and LLM-based—and explored two decoding strategies: *base* and *pairwise*. Through extensive experimentation, we analyzed the strengths and limitations of each approach across different settings, including language-specific and mixed-language training, as well as variations in emotional intensity and class distribution. The empirical results validate the simplicity and effectiveness of our proposed methods. They also underscore the importance of strategy selection based on language and label characteristics. Our findings provide valuable insights for future research in multilingual emotion recognition and offer practical guidance for extending these techniques to other cross-lingual and low-resource tasks.

Chapter 7

Conclusion and Future Work

7.1 Conclusion and Future Work

7.1.1 Conclusion

Our work primarily focuses on two key areas: (1) Emotion Recognition in Conversations, and (2) Multilingual multi-label emotion analysis.

In the ERC domain, we explore several under-investigated yet crucial aspects and list the main ERC models in recent years and their comparison with various models proposed by us in the table [7.1](#):

- Emotional Context Modeling: We propose the *Long-range dependencY emotionS Model (LYSM)*, which captures long-range emotional dependencies across entire dialogues. LYSM demonstrates strong and competitive performance on four benchmark datasets, despite its simple and modular architecture. The modularity makes it easily integrable into other emotion recognition frameworks.
- Utterance Modeling: We introduce *Accumulating Word Representations in Multi-level Context Integration for ERC Task (AccWR)*, which accumulates word-level representations within utterances and integrates them with the surrounding context. This sequence is then input into a PLM (e.g., RoBERTa), fully leveraging its power. Experimental results across four widely-used benchmarks show that our

model not only surpasses many existing approaches but also rivals SOTA systems. Furthermore, this method holds broader potential for context-aware sentence representation tasks such as response generation, multi-turn QA, and dialogue-based semantic parsing.

- **Speaker Modeling:** We propose *BiosERC: Integrating of Biography Speakers Supported by Large Language Model for ERC Task*, which utilizes explicit speaker attributes as external knowledge. This allows the model to capture speaker-specific linguistic styles and personality traits, leading to enhanced performance. Consistently achieving SOTA or highly competitive results on three benchmark datasets, BiosERC demonstrates both its effectiveness and adaptability to various dialogue analysis tasks.
- **Multi-context Modeling:** We present TraceERC: **T**racking **R**elational **A**wareness of Contextual, **C**haracter, and **E**motional States in **ERC** tasks, a comprehensive framework that integrates semantic context, static speaker context, and dynamic emotional context. Leveraging LLMs, TraceERC captures speaker-specific personality traits and emotion dynamics to generate rich, context-aware, emotionally grounded representations. As one of the pioneering efforts in LLM-based in-context learning and instructional fine-tuning for ERC, we conduct rigorous component-wise evaluations. The model consistently achieves competitive results and sets a new SOTA on the MELD dataset.
- **Bridging the Gap in Text-Based Emotion Detection:** In our multilingual sentiment recognition study, we explore combining pre-trained models such as RoBERTa with LLMs for multi-label emotion detection across high-resource (e.g., English, Chinese) and low-resource languages (e.g., African languages). We formulate multi-label sentiment recognition as a text generation task and leverage multilingual PLMs for enhanced semantic understanding and emotion tone detection, especially in resource-scarce settings. Our team, JNLP, participated in SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection, achieving top results in both Track A

Table 7.1: Performance Comparison of Proposed and Existing Methods. The columns EmoDept, SpkrMd, LocalCtx, and GlobalCtx refer to the Emotional dependencies, Speaker, Local context, and Global context modeling.

Methods	EmoDept	SpkrMd	LocalCtx	GlobalCtx	IEMOCAP	EmoryNLP	MELD
CESTa ([73])	✓	✓		✓	65.47	-	58.36
COSMIC* ([15])	✓	✓		✓	65.28	38.11	65.21
EmotionFlow ([65])	✓		✓	✓	65.05	-	-
LYSM (ours 2023 [76])	✓	✓		✓	65.47	38.34	65.40
HiTrans ([33])		✓	✓	✓	64.50	36.75	61.94
EmoBerta ([28])		✓	✓		67.42	-	65.61
SKAIG ([34])		✓	✓	✓	66.96	38.88	65.18
DialogXL ([62])		✓	✓	✓	65.94	34.73	62.14
AccWR-MLP (ours 2023 [26])		✓	✓	✓	67.65	39.33	64.58
MPLP ([84])		✓		✓	66.65	-	66.65
ESIHGNNa ([82])		✓		✓	68.53	39.56	63.92
EACL ([81])		✓	✓		70.41	40.24	67.12
MKFM ([69])		✓		✓	68.88	39.76	65.66
*DialogueLLM ([86])			✓		-	-	59.75
*InstructERC ([32])		✓	✓		71.39	41.39	69.15
BiosERC_{Llama-2} (ours 2024 [77])		✓	✓		71.19	41.68	69.83
TraceERC_{Llama-2} (ours 2025)	✓	✓	✓		71.32 [‡]	41.18 [‡]	69.91[‡]

(multi-label emotion classification) and Track B (emotion intensity detection).

7.2 Future Work

In future research on ERC, we plan to explore the following directions:

- **Multimodal Emotion Recognition:** Text-based emotion recognition suffers from inherent limitations due to the lack of paralinguistic and contextual cues. To overcome this bottleneck, we aim to incorporate multimodal signals—such as visual and acoustic features—to enhance the accuracy of emotion understanding in conversations.
- **Explicit Causes of Emotion Shifts:** We intend to investigate the explicit reasons behind emotional changes, focusing on identifying both the cause of each emotional expression and the specific utterance it refers to. By uncovering such explicit causal links, the model can avoid being biased toward irrelevant or misleading features and instead form more accurate associations with the truly informative causal reasons.
- **Speaker Relationship Modeling:** Although speaker descriptions provide some relational cues, our current work does not fully explore the deeper interpersonal dy-

namics between speakers, such as social roles, hierarchies, or mutual familiarity. We aim to model these nuanced speaker relationships to better capture emotional flow.

We also plan to further explore multilingual emotion recognition by addressing key challenges such as label imbalance and context sparsity. To mitigate label imbalance, we aim to develop emotion-aware data augmentation strategies; to alleviate context sparsity, we intend to generate synthetic dialogue contexts. These approaches hold the potential to improve model generalizability and robustness across diverse languages and cultures, paving the way for scalable, cross-cultural emotion understanding.

Publications and Awards

Submitted Journals

- [1] XUE Jieying, Phuong Minh Nguyen, and Le Minh Nguyen. TraceERC: Tracking Relational Awareness of Contextual, Character, and Emotional States in Emotion Recognition in Conversations, In *Neurocomputing* (Revised).

Conference papers

- [2] Jieying XUE, Phuong Nguyen, Nguyen Minh Le, and Xin Liu. JNLP at SemEval-2025 Task 11: Cross-Lingual Multi-Label Emotion Detection Using Generative Models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.
- [3] Xue, J., Nguyen, M.P., Matheny, B., Nguyen, L.M. BiosERC: Integrating biography speakers supported by LLMs for ERC tasks. In *International Conference on Artificial Neural Networks (pp. 277-292)*. Cham: Springer Nature Switzerland. Cham. DOI: 10.1007/978-3-031-72344-5_19
- [4] Xue Jieying, Phuong Nguyen, Matheny Blake and Nguyen Minh Le. Accumulating Word Representations in Multi-level Context Integration for ERC Task. In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, Hanoi, Vietnam, 2023, pp. 1-6, DOI: 10.1109/KSE59128.2023.10299463.

- [5] Xue, J.; Nguyen, M. and Nguyen, L. Emotions Relationship Modeling in the Conversation-Level Sentiment Analysis In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence* Volume 3: ICAART, ISBN 978-989-758-623-1; ISSN 2184-433X, SciTePress, pages 276-284. DOI: 10.5220/0011658200003393.

Awards

- ENNS Student Award by 33rd International Conference on Artificial Neural Networks ICANN2024, European Neural Network Society (ENNS) 9/2024, Lugano, Switzerland — for the paper: *BiosERC: Integrating Biographical Speakers Supported by LLMs for ERC Tasks*.
- Best System Description Award by 2025 Annual Conference of the Association for Computational Linguistics (ACL), the 19th International Workshop on Semantic Evaluation, 8/2025, Vienna, Austria — for the paper: *JNLP at SemEval-2025 Task 11: Cross-Lingual Multi-Label Emotion Detection Using Generative Models*.

Bibliography

- [1] K. Bansal, H. Agarwal, A. Joshi, and A. Modi. Shapes of emotions: Multi-modal emotion recognition in conversations via emotion shifts. In *Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models*, pages 44–56, Virtual, Oct. 2022. International Conference on Computational Linguistics.
- [2] Y. Bao, Q. Ma, L. Wei, W. Zhou, and S. Hu. Speaker-guided encoder-decoder framework for emotion recognition in conversation. In L. D. Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4051–4057. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.
- [3] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy, July 2019. Association for Computational Linguistics.
- [4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [5] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang,

- S. Lee, and S. S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, Dec 2008.
- [6] P. Chikersal, S. Poria, and E. Cambria. Sentu: sentiment analysis of tweets by combining a rule-based classifier with supervised learning. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 647–651, 2015.
- [7] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models, 2022.
- [8] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.
- [9] A. Conneau, s. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.
- [10] P. Dadure, A. Dixit, K. Tewatia, N. Paliwal, and A. Malla. Sentiment analysis of Arabic tweets using large language models. In M. El-Haj, editor, *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 88–94, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics.
- [11] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In D. Lee, M. Sugiyama,

- U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
- [12] J. Deng and F. Ren. Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. *IEEE Transactions on Affective Computing*, 14(1):475–486, 2020.
- [13] J. Deng and F. Ren. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1):49–67, 2021.
- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [15] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria. COSMIC: COMmonSense knowledge for eMotion identification in conversations. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online, Nov. 2020. Association for Computational Linguistics.
- [16] D. Ghosal, N. Majumder, R. Mihalcea, and S. Poria. Utterance-level dialogue understanding: an empirical study. *CoRR*, abs/2009.13902, 2020.
- [17] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh. DialogueGCN: a graph convolutional neural network for emotion recognition in conversation. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

- [18] G. Guibon, M. Labeau, H. Flamein, L. Lefeuvre, and C. Clavel. Few-shot emotion recognition in conversation with sequential prototypical networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6858–6870, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.
- [19] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [20] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.
- [21] G. Hou, Y. Shen, W. Zhang, W. Xue, and W. Lu. Enhancing emotion recognition in conversation via multi-view feature alignment and memorization. In *Conference on Empirical Methods in Natural Language Processing*, 2023.
- [22] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- [23] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li. UniMSE: towards unified multimodal sentiment analysis and emotion recognition. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.
- [24] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. Xtreme:

- A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR, 2020.
- [25] Z. Jian, A. Wang, J. Su, J. Yao, M. Wang, and Q. Wu. EmoTrans: emotional transition-based model for emotion recognition in conversation. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5723–5733, Torino, Italia, May 2024. ELRA and ICCL.
- [26] X. Jieying, N. Phuong, M. Blake, and N. Le Minh. Accumulating word representations in multi-level context integration for erc task. In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–6, 2023.
- [27] L. Kang, J. Liu, L. Liu, Z. Zhou, and D. Ye. Semi-supervised emotion recognition in textual conversation via a context-augmented auxiliary training task. *Information Processing & Management*, 58(6):102717, 2021.
- [28] T. Kim and P. Vossen. Emoberta: speaker-aware emotion recognition in conversation with roberta. *CoRR*, abs/2108.12009, 2021.
- [29] U. Kumari, A. K. Sharma, and D. Soni. Sentiment analysis of smart phone product review using svm classification technique. In *2017 International conference on energy, communication, data analytics and soft computing (ICECDS)*, pages 1469–1474. IEEE, 2017.
- [30] B. Lee and Y. S. Choi. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

- [31] J. Lee and W. Lee. CoMPM: context modeling with speaker’s pre-trained memory tracking for emotion recognition in conversation. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5669–5679, Seattle, United States, July 2022. Association for Computational Linguistics.
- [32] S. Lei, G. Dong, X. Wang, K. Wang, and S. Wang. Instructerc: reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*, 2023.
- [33] J. Li, D. Ji, F. Li, M. Zhang, and 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*, pages 4190–4200, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics.
- [34] J. Li, Z. Lin, P. Fu, and W. Wang. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1204–1214, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.
- [35] R. Li, Z. Wu, J. Jia, Y. Bu, S. Zhao, and H. M. Meng. Towards discriminative representation learning for speech emotion recognition. In *International Joint Conference on Artificial Intelligence*, 2019.
- [36] W. Li, L. Zhu, R. Mao, and E. Cambria. Skier: A symboli knowledge integrated model for conversational emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13121–13129, Jun. 2023.

- [37] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and 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)*, pages 986–995, Taipei, Taiwan, Nov. 2017. Asian Federation of Natural Language Processing.
- [38] Z. Li, Y. Wei, Y. Zhang, and Q. Yang. Hierarchical attention transfer network for cross-domain sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [39] C. Liang, J. Xu, Y. Lin, C. Yang, and Y. Wang. S+PAGE: A speaker and position-aware graph neural network model for emotion recognition in conversation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 148–157, Online only, Nov. 2022. Association for Computational Linguistics.
- [40] C. M. Liapis, A. Karanikola, and S. Kotsiantis. Enhancing sentiment analysis with distributional emotion embeddings. *Neurocomputing*, 634:129822, 2025.
- [41] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *ArXiv*, abs/1703.03130, 2017.
- [42] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: a robustly optimized bert pretraining approach, 2019.
- [43] Y. Liu, J. Zhao, J. Hu, R. Li, and Q. Jin. DialogueEIN: emotion interaction network for dialogue affective analysis. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, editors, *Proceedings of the 29th International Conference on Compu-*

- tational Linguistics*, pages 684–693, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics.
- [44] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.
- [45] A. Maazallahi, M. Asadpour, and P. Bazmi. Advancing emotion recognition in social media: A novel integration of heterogeneous neural networks with fine-tuned language models. *Information Processing & Management*, 62(2):103974, 2025.
- [46] L. L. Maceda, J. L. Llovido, M. B. Artiaga, and M. B. Abisado. Classifying sentiments on social media texts: A gpt-4 preliminary study. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval, NLPPIR '23*, page 19–24, New York, NY, USA, 2024. Association for Computing Machinery.
- [47] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria. Dialoguernn: an attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825, 2019.
- [48] Y. Mao, G. Liu, X. Wang, W. Gao, and X. Li. DialogueTRM: exploring multi-modal emotional dynamics in a conversation. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2694–2704, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.
- [49] S. Minaee, E. Azimi, and A. Abdolrashidi. Deep-sentiment: sentiment analysis using ensemble of cnn and bi-lstm models. *arXiv preprint arXiv:1904.04206*, 2019.

- [50] M. Mujahid, K. Kanwal, F. Rustam, W. Aljedaani, and I. Ashraf. Arabic chatgpt tweets classification using roberta and bert ensemble model. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(8), Aug. 2023.
- [51] P. Nandwani and R. Verma. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81, 2021.
- [52] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, page 79–86, USA, 2002. Association for Computational Linguistics.
- [53] T. S. Polzin and A. Waibel. Emotion-sensitive human-computer interfaces. In *Proc. ITRW on Speech and Emotion*, pages 201–206, 2000.
- [54] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and 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)*, pages 873–883, Vancouver, Canada, July 2017. Association for Computational Linguistics.
- [55] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics.
- [56] S. Poria, N. Majumder, D. Hazarika, D. Ghosal, R. Bhardwaj, S. Y. B. Jian, P. Hong, R. Ghosh, A. Roy, N. Chhaya, et al. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332, 2021.

- [57] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.
- [58] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.
- [59] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- [60] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
- [61] S. Sharma, R. Aggarwal, and M. Kumar. Mining twitter for insights into chatgpt sentiment: A machine learning approach. *2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, pages 1–6, 2023.
- [62] W. Shen, J. Chen, X. Quan, and Z. Xie. Dialogxl: all-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797, 2021.
- [63] W. Shen, S. Wu, Y. Yang, and X. Quan. Directed acyclic graph network for conversational emotion recognition. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560, Online, Aug. 2021. Association for Computational Linguistics.

- [64] X. Shi, X. Li, and T. Toda. Emotion Awareness in Multi-utterance Turn for Improving Emotion Prediction in Multi-Speaker Conversation. In *Proc. INTERSPEECH 2023*, pages 765–769, 2023.
- [65] X. Song, L. Zang, R. Zhang, S. Hu, and L. Huang. Emotion-flow: capture the dialogue level emotion transitions. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8542–8546, 2022.
- [66] H. Touvron, L. Martin, K. R. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. M. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. S. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. M. Kloumann, A. V. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. H. M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023.
- [67] A. Tripathy, A. Agrawal, and S. K. Rath. Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, 57:117–126, 2016.
- [68] G. Tu, B. Liang, B. Qin, K.-F. Wong, and R. Xu. An empirical study on multiple knowledge from chatgpt for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12160–12173, 2023.
- [69] G. Tu, B. Liang, B. Qin, K.-F. Wong, and R. Xu. An empirical study on multiple knowledge from ChatGPT for emotion recognition in conversations, Dec. 2023.

- [70] A. Upadhye. Sentiment analysis using large language models: Methodologies, applications, and challenges. *Int. J. Comput. Appl*, 186:30–34, 2024.
- [71] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention neural information is all you neural information need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Neural Information Associates, Inc., 2017.
- [72] R. Wang, Q. Yang, S. Tian, L. Yu, X. He, and B. Wang. Transformer-based correlation mining network with self-supervised label generation for multi-modal sentiment analysis. *Neurocomputing*, 618:129163, 2025.
- [73] Y. Wang, J. Zhang, J. Ma, S. Wang, and J. Xiao. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195, 1st virtual meeting, July 2020. Association for Computational Linguistics.
- [74] M. Wankhade, A. C. S. Rao, and C. Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.
- [75] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- [76] J. Xue, M.-P. Nguyen, and L. M. N. 0001. Emotions relationship modeling in the conversation-level sentiment analysis. In *ICAART (3)*, pages 276–284, 2023.
- [77] J. Xue, M.-P. Nguyen, B. Matheny, and L.-M. Nguyen. Bioserc: Integrating biography speakers supported by llms for erc tasks. In M. Wand, K. Malinovská,

- J. Schmidhuber, and I. V. Tetko, editors, *Artificial Neural Networks and Machine Learning – ICANN 2024*, pages 277–292, Cham, 2024. Springer Nature Switzerland.
- [78] J. Xue, M.-P. Nguyen, B. Matheny, and L.-M. Nguyen. Bioserc: Integrating biography speakers supported by llms for erc tasks. In *International Conference on Artificial Neural Networks*, pages 277–292. Springer, 2024.
- [79] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics.
- [80] L. Yang, Z. Wang, Z. Li, J.-C. Na, and J. Yu. An empirical study of multimodal entity-based sentiment analysis with chatgpt: improving in-context learning via entity-aware contrastive learning. *Information Processing & Management*, 61(4):103724, 2024.
- [81] F. Yu, J. Guo, Z. Wu, and X. Dai. Emotion-anchored contrastive learning framework for emotion recognition in conversation. In K. Duh, H. Gomez, and S. Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4521–4534, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
- [82] X. Zha, H. Zhao, and Z. Zhang. Esihgnn: event-state interactions infused heterogeneous graph neural network for conversational emotion recognition. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11136–11140, 2024.
- [83] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou. Model-

- ing both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI*, pages 5415–5421, 2019.
- [84] T. Zhang, Z. Chen, M. Zhong, and T. Qian. Mimicking the thinking process for emotion recognition in conversation with prompts and paraphrasing. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI '23*, 2023.
- [85] X. Zhang, S. Li, B. Hauer, N. Shi, and G. Kondrak. Don’t trust ChatGPT when your question is not in English: a study of multilingual abilities and types of LLMs. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore, Dec. 2023. Association for Computational Linguistics.
- [86] Y. Zhang, M. Wang, P. Tiwari, Q. Li, B. Wang, and J. Qin. Dialoguellm: context and emotion knowledge-tuned llama models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*, 2023.
- [87] W. Zhao, M. Geva, B. Y. Lin, M. Yasunaga, A. Madaan, and T. Yu. Complex reasoning in natural language. In Y.-N. V. Chen, M. Margot, and S. Reddy, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 11–20, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [88] W. Zhao, Y. Zhao, and X. Lu. Cauain: causal aware interaction network for emotion recognition in conversations. In L. D. Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4524–4530. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.
- [89] W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*, 2023.

- [90] W. Zhao, Y. Zhao, and B. Qin. MuCDN: mutual conversational detachment network for emotion recognition in multi-party conversations. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7020–7030, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics.
- [91] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023.
- [92] P. Zhong, D. Wang, and C. Miao. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
- [93] X. Zhou, X. Wan, and J. Xiao. Attention-based lstm network for cross-lingual sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 247–256, 2016.