What Matters in Transformer-based Emotion Recognition in Conversations? A Systematic Empirical Study

Rufaida Kashif[1], **Benjamin Piwowarski**[1], **Helena Gómez Adorno**[2]
[1] Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, F-75005 Paris, France
[2] Universidad Nacional Autónoma de México, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Ciudad de México, México
{rufaida.kashif, benjamin.piwowarski}@sorbonne-universite.fr
helena.gomez@iimas.unam.mx

Abstract

Emotion Recognition in Conversations (ERC) requires modeling complex contextual dependencies across dialog turns. While transformer-based models achieve strong performance on ERC benchmarks, several key design choices including context construction, optimization strategies, and imbalance handling remain insufficiently examined. In this work, we conduct a systematic empirical study of transformer-based ERC models across three benchmark datasets. We analyze the impact of context length and directionality, layer freezing, learning rate scheduling, parameter-efficient fine-tuning, and class imbalance mitigation strategies. Our results show that short-to-medium conversational context and moderate layer freezing provide stable and strong performance, while very long context windows, aggressive freezing, and parameter-efficient adaptation offer limited gains. Furthermore, imbalance-aware losses and data augmentation do not consistently outperform standard cross-entropy training. Overall, our findings provide practical insights into effective and stable design choices for transformer-based conversational emotion recognition.

Keywords: Emotion Recognition in Conversations, Transformers, Context Modeling, Optimization, Class Imbalance

1. Introduction

Human emotions are central to human behavior, cognition, and social interaction, making their accurate identification an important goal in various disciplines. In natural language processing, text-based emotion recognition has gained increasing attention due to its relevance in applications such as mental health support (Machová et al., 2023), empathetic dialog systems (Tafreshi et al., 2021), and socially aware human-computer interaction (Erol et al., 2020).

Emotion recognition from textual data has been an active area of research for many years, particularly in affective computing (Pereira et al., 2022). Much of this work focuses on isolated text, such as tweets or short user-generated content, where each instance is treated independently for emotion classification. In contrast, conversational text introduces a more complex setting, as it consists of dialogs—sequences of utterances exchanged between two or more speakers. Unlike single-text inputs, conversational data exhibit strong contextual dependencies, where the emotional meaning of an utterance is often shaped by preceding discourse. Speakers continuously respond to one another, giving rise to emotional transitions and pragmatic cues that evolve over multiple turns. As a result, the same utterance may convey different emotions depending on prior context, speaker roles, and interaction dynamics. Emotion Recognition in Conversations (ERC) addresses this challenge by modeling dialogs as sequences of interdependent utterances, often involving multiple speakers whose emotional states evolve over time (Tu et al., 2022). Consider the utterance "*It doesn't matter*." In one dialog, following the exchange "Do you want to try setting a goal for next week?" $\rightarrow$ "I tried last time and failed anyway," $\rightarrow$ *It doesn't matter*; the same utterance may express *hopelessness*. In contrast, in another dialog following "Should we do this work meeting before or after the break?" $\rightarrow$ "Either's okay," $\rightarrow$ *It doesn't matter* it may instead be interpreted as *neutral*. This demonstrates that identical utterances can convey different emotions depending on the broader conversational context, highlighting the importance of context modeling in ERC.

Before the widespread adoption of transformer-based models, various approaches for modeling conversational context were proposed, including sequential architectures such as DialogueRNN (Majumder et al., 2019), which models conversational context by maintaining speaker-specific and global recurrent states across dialog turns. Hierarchical and transformer-based models such as DialogueXL (Majumder et al., 2020) and contextual encoding approaches (Zahiri and Choi, 2019) aim to represent longer conversational dependencies. More recently, transformer-based models have become the

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 93–104
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

dominant paradigm in ERC, as self-attention mechanisms enable flexible modeling of long-range dependencies. Pre-trained transformer models such as BERT (Devlin et al., 2019) and its variants have been widely adopted as utterance encoders in ERC systems due to their strong contextual representation capabilities (Dave et al., 2024). Handling conversational context remains a key challenge in ERC. (Hazarika et al., 2021) propose a transfer learning approach that employs BERT as an utterance encoder, while modeling conversational context using a bidirectional GRU and transfer learning. Another approach for the context was presented by (Li et al., 2020) where they stack transformer layers to capture conversational context. Recent work such as EmoBERTa (Kim and Vossen, 2021) effectively leverages pretrained language models, specifically RoBERTa, to model conversational context through structured input representations and token-level interactions. Similarly, BERT-ERC (Qin et al., 2023) extends this paradigm by incorporating conversational context within a BERT-based framework, employing techniques such as masking and teacher-student learning to enhance performance on standard ERC benchmarks. These findings further support the effectiveness of BERT-based encoders for emotion recognition tasks. Despite these advances, a key limitation of existing work is that many design choices in transformer-based ERC are treated as fixed implementation decisions rather than variables for systematic analysis. In particular, models differ in how conversational context is constructed, such as context length, directionality, and truncation strategy, yet the impact of these choices is rarely examined in a controlled manner. Similarly, training strategies including partial layer freezing, learning rate scheduling, and regularization are often adopted from general NLP practice without evaluating their task-specific effects. Furthermore, class imbalance, which is prevalent in ERC datasets, is typically addressed through predefined loss functions, with limited empirical analysis of their effectiveness.

In this work, we address these gaps through a systematic empirical study of transformer-based ERC. We focus on transformer-based encoders due to their ability to capture long-range dependencies and their strong empirical performance across ERC benchmarks. Rather than proposing a new model architecture, we adopt a controlled experimental framework to isolate and analyze key design choices commonly used in ERC pipelines. We first investigate how different formulations of conversational context—including context length, directionality, and fixed versus variable context windows—affect a model's ability to capture emotional dependencies across dialog turns. We then examine the role of training and optimization strategies, including partial layer freezing and hyperparameter configurations such as learning rate schedules, weight decay, and warmup. Finally, we analyze the impact of class imbalance by comparing alternative loss formulations and sampling strategies under consistent experimental conditions. We conduct experiments and evaluate our approach on three widely used ERC benchmarks: IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and EmoryNLP (Zahiri and Choi, 2018), which differ in conversational structure, annotation schemes, and domain characteristics. Since all datasets considered in this study consist of English-language conversations, our analysis is restricted to English ERC settings, and the findings may not directly generalize to other languages or cultural contexts.

In this study, we investigate the following research questions:

1. **RQ1:** How does conversational context influence transformer-based Emotion Recognition in Conversations?

2. **RQ2:** How do training and optimization strategies affect transformer performance in Emotion Recognition in Conversations?

3. **RQ3:** How does class imbalance affect transformer performance in ERC, and to what extent can alternative loss formulations mitigate its impact?

2. Related Work

2.1. Sequential and Hierarchical Models for ERC

Early work on Emotion Recognition in Conversations modeled dialog context using sequential neural architectures, primarily recurrent neural networks such as GRUs and LSTMs. These approaches encode utterances in temporal order to capture emotional dynamics across dialog turns, establishing the importance of conversational context for emotion recognition. To better reflect conversational structure, hierarchical and speaker-aware models were later introduced. DialogRNN (Majumder et al., 2019), for example, maintains separate recurrent states for individual speakers and global context, enabling explicit modeling of inter-speaker interactions. Subsequent extensions, including TL-ERC (Hazarika et al., 2021) and DialogCRN (Hu et al., 2021), further explored hierarchical and speaker-dependent representations using recurrent encoders. While effective for short to moderately long conversations, these models remain constrained by the sequential nature of recurrence, which limits their ability to capture long-range dependencies and flexibly adapt to varying

context lengths; an issue that becomes more pronounced in extended dialogs.

2.2. Graph-based Approaches to Conversational Emotion Modeling

To address the limitations of purely sequential modeling; several studies have proposed graph-based architectures for ERC that explicitly encode relationships among utterances and speakers. In these approaches, dialogs are represented as graphs, with nodes corresponding to utterances and edges capturing temporal, contextual, or speaker-related dependencies. DialogGCN (Ghosal et al., 2019) introduced context- and speaker-aware graphs that propagate emotional information across dialog turns using graph neural networks. Subsequent work, such as SKAIG (Li et al., 2021), refined this paradigm by incorporating structured interaction patterns and attention mechanisms to enhance relational reasoning. Some graph-based models further integrate external knowledge to enrich emotional representations. COSMIC (Ghosal et al., 2020), for instance, augments conversational graphs with commonsense knowledge to reason about implicit emotional and intentional states. Despite their expressiveness, graph-based methods typically rely on predefined graph structures or heuristic context selection, which constrains flexibility and complicates the systematic analysis of how conversational context should be defined or truncated. Recent work also explores integrating external knowledge and structured reasoning to enhance emotional understanding in conversations (Zhao et al., 2024), reflecting a growing interest in enhancing emotional understanding beyond surface-level text representations.

2.3. Transformer-based Approaches to ERC

Transformer architectures have become the dominant paradigm in Emotion Recognition in Conversations due to their ability to model long-range dependencies through self-attention. Unlike recurrent or graph-based models, transformers enable each utterance to attend directly to others within a context window, offering greater flexibility in capturing conversational dynamics. As a result, pretrained language models have been widely adopted as backbones for ERC. Several studies adapt pretrained transformers to conversational settings by incorporating utterance-level and dialog-level context during fine-tuning. EmoBERTa (Kim and Vossen, 2021) leverages RoBERTa representations with contextualized attention mechanisms to capture emotional dependencies across dialog turns, while BERT-ERC (Qin et al., 2023) integrates conversational context directly into transformer-based training. While these methods demonstrate strong empirical performance, they vary substantially in how conversational context is defined, truncated, and incorporated during training. Recent work has further extended transformer-based ERC by incorporating external knowledge, emotional reasoning, and large language model (LLM) capabilities. Knowledge-enhanced approaches such as SKIER (Li et al., 2023) and commonsense-aware frameworks (Wang et al., 2024) integrate structured semantic information to improve affective understanding across dialog turns. Other studies explicitly model emotional transitions and temporal dependencies (Jian et al., 2024; Tu et al., 2022), emphasizing the importance of contextual dynamics. In parallel, prompt-based and LLM-driven approaches have emerged as an alternative paradigm for ERC. Instruction-driven frameworks (Lei et al., 2023), contrast-enhanced prompt tuning methods (Gao et al., 2024), and empirical analyses leveraging ChatGPT (Tu et al., 2023) demonstrate the potential of LLMs for conversational emotion understanding under suitable prompting or retrieval strategies. Additional work explores multi-view representation learning (Hou et al., 2023) and affect-aware reasoning in complex conversational scenarios (Kumar et al., 2023). Most ERC research has primarily focused on English-language conversational datasets, such as IEMOCAP, MELD, and EmoryNLP, due to the availability of annotated resources. As a result, pretrained models used in ERC are typically evaluated in English settings, and their behavior in multilingual or cross-cultural contexts remains less explored. At the same time, recent advances in natural language processing have emphasized scaling transformer architectures toward larger models with increased parameter counts, such as Mistral (Jiang et al., 2023), alongside broader training data. While such models exhibit strong general language understanding capabilities, their effectiveness for fine-grained conversational emotion recognition remains an open question. In practice, widely used pretrained encoders such as RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021) remain competitive under supervised fine-tuning, particularly when contextual modeling strategies are carefully designed. Despite these advances, most prior work focuses on proposing new architectures or integrating additional knowledge sources. Despite these advances, most prior work focuses on proposing new architectures or incorporating additional knowledge sources, rather than systematically analyzing how fundamental design choices—such as context formulation, optimization strategies, or parameter-efficient fine-tuning—affect model performance. This gap motivates the empirical focus of the present study.

3. Methodology

In this section, we present a controlled experimental framework for systematically investigating the research questions introduced in Section 1. We study three dimensions of transformer-based ERC systems: (RQ1) conversational context construction, including context length and directionality; (RQ2) training and optimization strategies, including layer freezing, learning rate, warmup ratio, and parameter-efficient fine-tuning; and (RQ3) class imbalance mitigation through alternative loss functions and data-level augmentation. For each experiment, only the variable under investigation is changed while all other components are kept fixed. We report the mean and standard deviation of weighted F1 over five runs with different random seeds to assess both effectiveness and stability. Unlike prior work that primarily emphasizes new architectures or additional knowledge sources, our methodology focuses on the systematic evaluation of existing design choices, enabling a clearer understanding of the factors that influence transformer-based ERC performance. Full code and configuration details will be released upon publication.

3.1. Problem Formulation

Emotion Recognition in Conversations (ERC) aims to predict the emotional state of a target utterance within a dialog by leveraging its conversational context. A dialog is represented as an ordered sequence of utterances:

$$D = \big[(u_1, s_1), (u_2, s_2), \ldots, (u_T, s_T)\big], \qquad (1)$$

where u_t denotes the textual content of the t-th utterance and s_t denotes the corresponding speaker. Each utterance u_t is associated with an emotion label y_t, and the goal is to predict the label of a target utterance using its surrounding conversational context.

3.2. Context Construction

For a target utterance u_t, we construct a conversational context window as an ordered sequence:

$$C_t^{(k,m)} = \big[u_{t-k}, \ldots, u_{t-1}, u_t, u_{t+1}, \ldots, u_{t+m}\big], \quad (2)$$

where k and m denote the number of past and future context utterances, respectively. This formulation allows us to explicitly control the amount and direction of conversational context available to the model, which is central to our investigation in RQ1.

3.3. Input Representation and Encoding

Each utterance in the conversational context $C_t^{(k,m)}$ is prepended with an explicit speaker name to preserve speaker identity. The utterances are then concatenated into a single input sequence and tokenized according to the underlying pretrained transformer. All other components, including the encoder architecture, preprocessing pipeline, and evaluation protocol, are kept fixed across experiments to ensure fair comparison. For a target utterance u_t, the input is constructed by concatenating the past context, the target utterance, and the future context:

$$X_t = [C_t^- \texttt{ </s></s> } u_t \texttt{ </s></s> } C_t^+], \qquad (3)$$

where </s> denotes a special sentence boundary marker used to separate segments in transformer inputs (the exact token may vary depending on the specific pretrained model). $C_t^- = [u_{t-k}, \ldots, u_{t-1}]$ and $C_t^+ = [u_{t+1}, \ldots, u_{t+m}]$ denote the past and future context utterances, respectively. The resulting input sequence X_t is encoded using a pretrained transformer encoder $f_\theta(\cdot)$:

$$\mathbf{H}_t = f_\theta(X_t), \qquad (4)$$

where $\mathbf{H}_t$ denotes the contextualized token representations. The representation corresponding to the special classification token is used as the utterance-level representation for emotion prediction. We adopt identical preprocessing, tokenization, and training configurations across datasets to ensure a controlled comparison of design choices. This input formulation follows prior work in ERC (Kim and Vossen, 2021), where conversational context is encoded as a single sequence to leverage the self-attention mechanism of transformers.

3.4. Training Configuration

Unless otherwise specified, all experiments are conducted using RoBERTa-large as the encoder backbone. Models are trained for 20 epochs using the ADAM optimizer with a batch size of 16. For experiments involving learning rate and warmup analysis, the corresponding hyperparameters are varied as described in Section 8 and Section 9, while all other settings remain fixed. Validation performance is monitored using weighted F1 score on the validation split. The model checkpoint that achieves the highest validation weighted F1 is selected for final evaluation on the test set. Performance is evaluated using weighted F1 score ($F1_W$), which computes the F1 score for each class and averages them weighted by their support. This metric accounts for class imbalance and reflects overall performance across emotion categories. We adopt RoBERTa-large as the backbone encoder due to its strong performance in prior ERC studies and its ability to capture rich contextual representations.

Dataset	Train	Validation	Test
MELD	1,038 (9,989)	114 (1,109)	280 (2,610)
IEMOCAP	100 (4,778)	20 (980)	31 (1,622)
EmoryNLP	77 (9,934)	11 (1,344)	9 (1,328)

Table 1: Dataset statistics showing the number of dialogs and (utterances) in each split.

3.4.1. Loss Functions

To address class imbalance in ERC datasets, we evaluate standard cross-entropy loss alongside focal loss (Lin et al., 2017) which extends cross-entropy by down-weighting easy examples and focusing training on hard-to-classify instances:

$$\mathcal{L}_{FL} = -\sum_{c=1}^{C}(1-p_c)^{\gamma} y_c \log(p_c), \tag{5}$$

where γ is a focusing parameter controlling the strength of down-weighting. In our experiments, we set $\gamma = 2$.

4. Datasets

We evaluate our approach on three widely used ERC benchmark datasets: IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and EmoryNLP (Zahiri and Choi, 2018). These datasets cover diverse conversational settings, including dyadic and multiparty interactions, varying dialog lengths, and different emotion annotation schemes. All datasets provide utterance-level emotion labels within multi-turn conversations, making them suitable for analyzing the role of dialog context. Although IEMOCAP and MELD include multimodal signals, we restrict our experiments to the textual modality to ensure consistency across datasets. Dataset statistics are summarized in Table 1. Furthermore, all three corpora consist of English-language conversations. Consequently, the findings of this study primarily reflect model behavior in English ERC scenarios and may not directly generalize to other languages or cross-lingual settings. Each dataset is treated as an independent evaluation setting due to differences in annotation schemes and label spaces. No cross-dataset training or label-space harmonization is performed.

4.1. IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset (Busso et al., 2008) is a standard benchmark for conversational emotion recognition, consisting of 151 dyadic dialogs performed by professional actors across five sessions. Each session includes both scripted and improvised interactions, with emotion annotations obtained through human evaluation. As IEMOCAP does not provide a unified dialog-level textual representation, we reconstruct dialog sequences through a structured preprocessing pipeline. Transcriptions from both scripted and improvised interaction sessions are first aggregated and grouped using their dialog identifiers. Emotion annotations, which are distributed separately from the transcription files, are matched to utterances via their unique utterance IDs. To ensure correct conversational ordering, utterances within each dialog are sorted according to their start timestamps rather than relying solely on filename or identifier ordering, which can lead to inconsistencies in dialog flow. This process results in temporally coherent dialog sequences that preserve the original conversational structure. We adopt a six-class emotion subset *(happiness, sadness, anger, frustration, excited, and neutral)* as justified in section 4.4. Emotion annotations are derived from multiple annotators; we use the consensus ground-truth label for each utterance. Utterances with other labels are retained in the dialog context but excluded from loss computation. We adopt a session-based split (Sessions 1–4 for training/validation, Session 5 for testing) to prevent speaker overlap. The validation split follows the proportions used in EmoBERTa (Kim and Vossen, 2021). Since IEMOCAP does not provide explicit speaker names, we replace speaker tags (M/F) with consistent first names following the same preprocessing.

4.2. MELD

The Multimodal EmotionLines Dataset (MELD) (Poria et al., 2019) consists of multiparty dialogs extracted from the television series *Friends*, annotated at the utterance level with emotion and sentiment labels. We use the standard seven emotion categories provided in the dataset: *neutral, joy, anger, sadness, surprise, disgust, and fear*. MELD provides predefined train, validation, and test splits and requires minimal preprocessing. We retain utterance order and speaker annotations, use only the textual modality, and reformat inputs for consistency.

4.3. EmoryNLP

The EmoryNLP dataset (Zahiri and Choi, 2018) is another conversational emotion corpus derived from tv show *Friends*, featuring longer and more context-rich dialogs. Each utterance is annotated with one of seven emotion labels: *neutral, joyful, peaceful, powerful, scared, mad, and sad*. Like MELD, EmoryNLP provides predefined train, validation, and test splits. We retain dialog structure, utterance order, and speaker information, using only the textual modality. Its longer dialogs make it

Table 2: Emotion label distribution in the IEMOCAP dataset.

Emotion	# Utterances
No_Emotion	1587
Frustration	1149
Neutral	1167
Anger	711
Sadness	739
Happiness	392
Excited	620
Surprise	76
Fear	23
Disgust	2
Other	2

Table 3: Emotion label distribution in the MELD dataset.

Emotion	# Utterances
Neutral	4710
Joy	1743
Surprise	1205
Anger	1109
Sadness	683
Disgust	271
Fear	268

well-suited for analyzing extended conversational context in ERC.

4.4. Label Distribution and Dataset Characteristics

All three datasets exhibit notable class imbalance, with neutral or no-emotion dominating the label distributions. In MELD, *neutral* emotion utterances constitute the largest proportion of the data, followed by *joy* and *surprise*, while *fear* and *disgust* are comparatively underrepresented. EmoryNLP presents a more balanced distribution across seven emotion categories, although *neutral* and *joyful* utterances remain the most frequent. The imbalance is particularly pronounced in IEMOCAP, where several emotion categories such as *disgust*, *other*, and *fear* occur extremely rarely. In contrast, emotions such as *neutral*, *frustration*, and *anger* are substantially more prevalent. This skewed distribution motivates the adoption of a commonly used six-class evaluation subset (happiness, sadness, anger, frustration, excited, and neutral), as several other categories (e.g., disgust, fear, and other) are extremely underrepresented and do not provide sufficient samples for reliable model training or evaluation. This practice is widely adopted in prior ERC work to ensure comparability and statistical robustness.

Table 4: Emotion label distribution in the EmoryNLP dataset.

Emotion	# Utterances
Neutral	3034
Joyful	2184
Scared	1285
Mad	1076
Peaceful	900
Powerful	784
Sad	671

5. Results

5.1. Context Window Analysis

Tables 5 and 6 report the impact of conversational context formulation on emotion recognition performance across IEMOCAP, MELD, and EmoryNLP. Removing contextual information (k=0, m=0) leads to a substantial drop in performance on IEMOCAP (55.3 F1) compared to moderate context settings such as $(10, 0)$, which achieves 67.0 F1, corresponding to an absolute improvement of 11.7 points. This confirms that emotional interpretation in dialog cannot be reliably inferred from isolated utterances. On MELD and EmoryNLP, context removal yields 62.9 and 34.6 F1 respectively, indicating that contextual dependencies remain important even when performance degradation is less pronounced than on IEMOCAP. Introducing short symmetric context windows (e.g., $(3, 3)$) improves performance compared to no-context baselines, but increasing fixed context lengths beyond moderate ranges does not consistently yield further gains. For example, on IEMOCAP, $(3, 3)$ and $(10, 10)$ both achieve 66.5 F1, indicating diminishing returns from longer fixed context windows. Similarly, on MELD, extending context from $(1, 1)$ (61.5 F1) to $(10, 10)$ (64.5 F1) provides improvement, but does not consistently outperform past-focused configurations. Directional ablations further reveal that past-only context is generally more informative than future-only context. On IEMOCAP, $(10, 0)$ achieves 67.0 F1, compared to 61.9 F1 for $(0, 10)$, suggesting that emotional states are more strongly conditioned on preceding conversational history. While bidirectional context yields competitive results, its advantage over past-only context is inconsistent across datasets.

In addition to fixed context windows, we evaluate a variable context training strategy to increase robustness to differing conversational lengths. Instead of always appending a fixed number of past and future utterances (e.g., 3 before and 3 after), we define a maximum context size (either 3 or 10). During training, for each target utterance, the number of past and future utterances is independently sampled from a uniform distribution between 0 and

Past	Fut.	IEMOCAP	MELD	EmoryNLP
3	3	66.5	62.4	10.9
0	0	55.3	62.9	34.6
1	1	59.5	61.5	36.9
10	10	66.5	64.5	24.3
0	3	60.9	60.9	36.1
0	1	57.1	63.2	38.1
0	10	61.9	64.1	30.5
3	0	64.4	64.7	10.9
1	0	61.5	62.5	31.7
10	0	**67.0**	62.1	36.5

Table 5: Fixed context window ablations (RQ1). Results are reported as weighted F1 on IEMOCAP, MELD, and EmoryNLP.

Context Range	IEMOCAP	MELD	EmoryNLP
0–3	64.2 ± 1.62	64.0 ± 1.59	**37.1 ± 1.05**
0–10	**67.6 ± 1.21**	**65.1 ± 0.59**	36.9 ± 1.18

Table 6: Variable context window results (RQ1). Results are reported as mean ± standard deviation of weighted F1 over multiple random seeds.

the specified maximum, resulting in dynamically varying context sizes across training instances. At evaluation time, the full maximum context is used for consistency. Using a variable range of 0–10 yields the best overall performance on IEMOCAP (67.6 ± 1.21) and strong performance on MELD (65.1 ± 0.59). The smaller 0–3 range achieves 64.2 ± 1.62 on IEMOCAP and 64.0 ± 1.59 on MELD, while performing best on EmoryNLP (37.1 ± 1.05). These results indicate that exposure to varying context sizes during training improves robustness, without requiring long fixed context windows.

5.2. Training and Optimization Strategies

5.2.1. Layer Freezing

To analyze the effect of layer freezing on generalization and training stability, we evaluate models with different numbers of frozen encoder layers across five random seeds (Table 7). Fully fine-tuning the encoder (0 frozen layers) achieves 67.3 ± 1.61 on IEMOCAP and 65.1 ± 0.62 on MELD, but exhibits high variance on EmoryNLP (31.6 ± 11.63), indicating instability across runs. Freezing 12 or 18 lower layers—corresponding to freezing 50%–75% of the 24-layer RoBERTa-large encoder—maintains comparable or slightly improved average performance while substantially reducing variance. For example, freezing 12 layers yields 66.8 ± 0.63 on IEMOCAP and 65.7 ± 0.88 on MELD, while freezing 18 layers achieves 67.2 ± 1.08 on IEMOCAP and 37.5 ± 1.65 on EmoryNLP, dramatically stabilizing performance compared to full fine-tuning. These configurations reduce standard deviation by up to an order of magnitude on EmoryNLP (from 11.63 to below 2), indicating improved training stability. In contrast, aggressively freezing 22 layers leads to consistent degradation (63.3 on IEMOCAP, a drop of 4.0 points from full fine-tuning). Freezing all 24 layers causes a severe collapse in performance across datasets (18.1 on IEMOCAP and 11.8 on EmoryNLP), confirming that ERC requires substantial task-specific adaptation of pretrained representations. Overall, freezing 12–18 layers emerges as a stable and effective regime, balancing adaptation capacity and regularization, whereas excessive freezing restricts the model's ability to capture conversational emotional cues.

Frozen	IEMOCAP	MELD	EmoryNLP
None (0)	67.3 ± 1.61	65.1 ± 0.62	31.6 ± 11.63
12	66.8 ± 0.63	**65.7 ± 0.88**	37.5 ± 0.91
18	67.2 ± 1.08	64.5 ± 1.25	**37.5 ± 1.65**
22	63.3 ± 0.47	64.5 ± 0.62	35.9 ± 0.64
All (24)	18.1 ± 2.12	31.7 ± 0.46	11.8 ± 0.87

Table 7: Effect of layer freezing on emotion recognition performance (RQ2). Results are reported as mean ± standard deviation of weighted F1 over five random seeds.

Learning Rate	Best Val F1	Test F1
1×10^{-6}	57.6	57.4
3×10^{-6}	64.5	65.9
1×10^{-5}	**66.2**	65.3
3×10^{-5}	65.2	64.8
1×10^{-4}	15.9	8.9

Table 8: Effect of learning rate on model performance (RQ2). All runs use identical settings except for the learning rate.

5.2.2. Learning Rate and Warmup Ratio

We first analyze the effect of learning rate selection on model generalization while keeping all other optimization settings fixed (Table 8). Very small learning rates lead to slow convergence and underfitting, resulting in poor validation and test performance. Increasing the learning rate improves performance up to a moderate range, with 1×10^{-5} achieving the strongest validation performance and stable generalization. Further increasing the learning rate degrades performance and introduces training instability, suggesting overshooting during optimization. These results highlight the sensitivity of transformer-based ERC models to learning rate selection and indicate that moderate learning rates are critical for balancing convergence speed and generalization.

We next examine the impact of learning rate warmup on training stability and final performance,

Warmup Ratio	Best Val F1	Test F1
0.05	65.8	67.2
0.10	66.6	67.1
0.20	**67.1**	**67.4**

Table 9: Effect of warmup ratio on performance (RQ2). Learning rate is fixed to 1×10^{-5}.

fixing the learning rate to 1×10^{-5} (Table 9). Introducing a warmup phase consistently improves both validation and test performance compared to minimal warmup. Increasing the warmup ratio yields incremental but consistent gains, with a warmup ratio of 0.2 achieving the best overall test performance. These findings suggest that learning rate warmup plays a stabilizing role during early training for transformer-based ERC models, although performance gains saturate beyond moderate warmup durations.

5.2.3. Parameter-Efficient Fine-Tuning with LoRA

We further evaluate Low-Rank Adaptation (LoRA) (Hu et al., 2022) as a parameter-efficient alternative to full fine-tuning. LoRA injects trainable low-rank matrices into selected attention projections while keeping the pretrained backbone frozen, thereby reducing the number of trainable parameters. In our experiments, LoRA is applied to the query and value projection matrices of the transformer encoder with rank $r = 8$, scaling factor $\alpha = 32$, and dropout rate 0.1. Table 10 reports results averaged over five random seeds. Compared to full fine-tuning of RoBERTa-large under identical optimization settings, LoRA yields consistently lower performance across all datasets. On IEMOCAP, LoRA achieves 56.7 $F1_w$ compared to 67.6 under full fine-tuning (-10.9 points). On MELD, performance drops from 65.1 to 61.7 (-3.4 points), and on EmoryNLP from 37.5 to 31.1 (-6.4 points). These results suggest that, under the evaluated configuration, LoRA does not match the performance of full fine-tuning for ERC. However, we note that only a single LoRA configuration was explored in this study. Further experiments varying rank, target modules, or partial layer freezing could provide a more comprehensive assessment of parameter-efficient adaptation for conversational emotion recognition.

5.3. Class Imbalance Analysis

As shown in the dataset analysis (Section 4.4), all three corpora exhibit notable class imbalance, with some emotions dominating the label distribution and several minority classes occurring only sparsely. This motivates the investigation of imbalance-aware training objectives in ERC.

Model	IEMOCAP	MELD	EmoryNLP
RoBERTa-Large (LoRA)	56.7 $\pm$ 0.86	61.7 $\pm$ 0.86	31.1 $\pm$ 2.55

Table 10: Effect of LoRA-based parameter-efficient fine-tuning on ERC performance. Results are reported as weighted F1 ($F1_w$) averaged over five random seeds.

Loss	IEMOCAP	MELD	EmoryNLP
Cross-Entropy	66.8 $\pm$ 0.63	65.7 $\pm$ 0.88	37.5 $\pm$ 0.91
Focal	66.6 $\pm$ 1.81	65.2 $\pm$ 0.59	37.6 $\pm$ 0.44

Table 11: Loss function comparison for class imbalance analysis (RQ3). Results are reported as mean $\pm$ standard deviation of weighted F1 (F1(W)) over five random seeds.

5.3.1. Loss Function Analysis

To evaluate whether imbalance-aware objectives improve ERC performance (RQ3), we compare standard cross-entropy loss with focal loss across five random seeds (Table 11). On IEMOCAP, cross-entropy achieves 66.8 $\pm$ 0.63 $F1_w$, while focal loss obtains 66.6 $\pm$ 1.81. On MELD, performance decreases slightly from 65.7 $\pm$ 0.88 to 65.2 $\pm$ 0.59. On EmoryNLP, both losses yield nearly identical results (37.5 $\pm$ 0.91 vs. 37.6 $\pm$ 0.44). These results indicate that focal loss does not provide consistent gains over cross-entropy in ERC. The absence of systematic improvements suggests that label imbalance alone may not be the dominant bottleneck in transformer-based conversational emotion recognition. A possible explanation is that minority emotion classes are not only underrepresented but also exhibit substantial contextual ambiguity. By emphasizing hard examples, focal loss may increase sensitivity to label noise rather than improve discriminative learning. Overall, standard cross-entropy appears sufficiently robust under pretrained transformer fine-tuning in ERC.

5.3.2. Data Augmentation for Class Imbalance

To further examine whether data-level balancing improves ERC performance (RQ3), we apply a text-based oversampling strategy using a pretrained FLAN-T5 model (Chung et al., 2022). For each dataset, minority-class utterances in the training split are augmented via paraphrasing. Specifically, for each selected utterance, FLAN-T5 generates paraphrased variants conditioned on preserving the original semantic meaning and emotional label. The validation and test splits remain unchanged to ensure fair evaluation. The augmented samples

Setting	IEMOCAP	MELD	EmoryNLP
Baseline	66.8 ± 0.63	**65.7 ± 0.88**	**37.5 ± 0.91**
FLAN-T5	66.1 ± 1.18	62.0 ± 0.67	34.2 ± 0.79

Table 12: Effect of FLAN-T5-based oversampling on ERC performance (RQ3). Results are reported as mean ± standard deviation of weighted F1 over five random seeds.

are merged with the original training data while preserving dialog order and speaker annotations. Table 12 reports the performance after augmentation. Compared to the baseline without augmentation (66.8 on IEMOCAP, 65.7 on MELD, and 37.5 on EmoryNLP; Table 11), oversampling yields lower performance on all datasets: 66.1 on IEMOCAP, 62.0 on MELD, and 34.2 on EmoryNLP. These results indicate that synthetic oversampling does not improve—and may slightly degrade—performance in transformer-based ERC. A possible explanation is that paraphrased utterances introduce lexical variability without adding new contextual signals, potentially increasing noise in already ambiguous emotional classes. Overall, data-level balancing through paraphrasing does not appear sufficient to address the structural challenges of class imbalance in conversational emotion recognition.

5.4. Model Comparison

We further compare several pretrained transformer backbones under identical training, optimization, and evaluation settings to assess the impact of architectural design and model capacity on ERC performance. We consider RoBERTa-base (Liu et al., 2019), DistilRoBERTa, DeBERTa-v3-base (He et al., 2021), and NeoBERT (Le Breton et al., 2025). These models represent different encoder design philosophies: standard robust pretraining (RoBERTa), parameter compression (DistilRoBERTa), disentangled attention mechanisms (DeBERTa), and a modernized next-generation encoder architecture (NeoBERT). Table 13 reports weighted F1 scores across IEMOCAP, MELD, and EmoryNLP. DeBERTa-v3-base achieves the strongest overall performance on IEMOCAP (64.8) and EmoryNLP (37.5), suggesting that its disentangled attention and improved positional encoding enhance contextual modeling in conversational settings. RoBERTa-base performs competitively, particularly on MELD (64.1), and consistently outperforms DistilRoBERTa. DistilRoBERTa shows a systematic performance drop compared to RoBERTa-base (e.g., 62.1 to 59.7 on IEMOCAP; 64.1 to 61.6 on MELD), reflecting the trade-off between model compression and representational capacity. NeoBERT, despite its architectural improvements and extended context design, performs comparably

Model	IEMOCAP	MELD	EmoryNLP
RoBERTa-base	62.1 ± 0.81	64.1 ± 0.89	36.3 ± 0.37
DistilRoBERTa	59.7 ± 0.96	61.6 ± 0.24	34.3 ± 1.20
NeoBERT	60.5 ± 1.46	61.1 ± 0.79	34.6 ± 1.33
DeBERTa-v3-base	64.8 ± 1.39	63.0 ± 0.96	37.5 ± 1.05

Table 13: Comparison of transformer backbones on ERC benchmarks. Results are reported as weighted F1 ($F1_W$) under identical training and evaluation settings.

to DistilRoBERTa but does not surpass RoBERTa-base or DeBERTa-v3-base under our ERC setup. Overall, models with stronger contextual representation mechanisms (RoBERTa-base and DeBERTa-v3-base) provide more stable and competitive performance across datasets, highlighting the importance of expressive encoder architectures for modeling nuanced conversational affect.

6. Conclusion

In this work, we present a systematic empirical study of transformer-based models for Emotion Recognition in Conversations, examining the impact of contextual design, optimization strategies, and class imbalance handling across multiple benchmarks. Our results demonstrate that a short-to-medium conversational context is sufficient for effective emotion modeling, while very long fixed context windows yield diminishing returns. Past conversational history consistently proves to be more informative than future context, and variable context training enhances robustness without requiring full-dialog modeling. Regarding optimization strategies, moderate layer freezing improves training stability without sacrificing performance, whereas excessive freezing or fully parameter-efficient adaptation through LoRA limits task-specific adaptation. Learning rate selection and warmup scheduling further play a critical role in stabilizing transformer fine-tuning for ERC. Finally, imbalance-aware interventions such as focal loss and data-level oversampling do not consistently outperform standard cross-entropy loss, suggesting that ERC performance is more constrained by contextual ambiguity and semantic overlap than by label frequency alone.

Overall, our findings provide practical guidance for designing stable and effective transformer-based ERC systems and highlight the importance of controlled evaluation of architectural and optimization choices in conversational emotion modeling.

7. Ethics Statement

This work uses publicly available benchmark datasets for Emotion Recognition in Conversations (ERC), including MELD and EmoryNLP, as well as the IEMOCAP dataset, which was obtained through an official data access request process in accordance with its licensing requirements. All datasets consist of scripted or recorded conversational data collected under institutional review procedures by their original creators. Our study focuses exclusively on the textual modality of English-language conversations. While emotion recognition technologies may support beneficial applications such as mental health monitoring or socially aware dialog systems, they may also raise ethical concerns related to privacy, surveillance, and unintended profiling if deployed without appropriate safeguards. Automated inference of emotional states may be inaccurate or contextually misleading, particularly in real-world settings involving diverse populations. We emphasize that the models evaluated in this study are research prototypes trained on limited benchmark datasets and should not be interpreted as reliable indicators of psychological states in real-world decision-making contexts.

8. Limitations

This study has several limitations. First, all experiments are conducted on English-language datasets, and therefore the findings may not generalize to multilingual or low-resource conversational settings. Second, our analysis is restricted to the textual modality, although some of the evaluated datasets (e.g., IEMOCAP and MELD) are inherently multimodal. Emotional cues conveyed through acoustic or visual signals are not considered in this work. Additionally, benchmark datasets for ERC exhibit class imbalance and annotation ambiguity, which may affect both model training and evaluation. Emotional labels in conversational settings are inherently subjective and context-dependent, potentially limiting the ceiling performance of supervised approaches. Finally, while our study systematically evaluates several architectural and training design choices, the experimental space remains limited. For example, only a single LoRA configuration was explored, and alternative parameter-efficient settings may yield different conclusions. Similarly, although multiple backbone models were compared, we did not conduct an exhaustive exploration of all modern encoder architectures or extensive hyperparameter sweeps for each variant. Future work could investigate a broader range of adaptation strategies and model configurations to provide a more comprehensive assessment of optimization and architectural choices for conversational emotion recognition. A limitation of this work is the absence of a comparison with prompt-based or large language model (LLM) approaches, such as GPT- or Gemini-style models. While such methods have recently been explored for ERC, our study focuses on supervised transformer encoders under controlled experimental settings. Future work will extend this analysis to include prompt-based techniques in order to more comprehensively evaluate whether encoder-based models remain competitive in conversational emotion recognition.

9. Data and Code Availability

We will release the code and implementation details for this work on a public repository upon publication.

10. Acknowledgment

This project is co-funded by the European Union's Horizon Europe research and innovation programme Cofund SOUND.AI under the Marie Sklodowska-Curie Grant Agreement No 101081674. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim Kazemzadeh, Emily Mower Provost, Sungbok Kim, Jeannette N. Chang, Sung Lee, and Shrikanth S. Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Urva Dave, Om Desai, Nisarg Chaudhari, Dweepna Garg, Kashyap Patel, and Parth Goel. 2024. Emotion detection in text-based communication using fine-tuned bert model. In *Proceedings of the International Conference on Sustainable Communication Networks and Applications (ICSCNA)*, pages 875–880.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of

deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*.

B. A. Erol, A. Majumdar, P. Benavidez, P. Rad, K.-K. R. Choo, and M. Jamshidi. 2020. Toward artificial emotional intelligence for cooperative social human–machine interaction. *IEEE Transactions on Computational Social Systems*, 7(1):234–246.

Qingqing Gao, Jiuxin Cao, Biwei Cao, Xin Guan, and Bo Liu. 2024. Cept: A contrast-enhanced prompt-tuning framework for emotion recognition in conversation.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, and Erik Cambria. 2021. Tl-erc: Transfer learning for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *ICLR*.

Guiyang Hou, Yongliang Shen, Wenqi Zhang, Wei Xue, and Weiming Lu. 2023. Enhancing emotion recognition in conversation via multi-view feature alignment and memorization. In *Proceedings of ACL*.

Dou Hu, Lingwei Wei, Baoliang Huai, Yunlong Su, and Xianfeng Tao. 2021. Dialogcrn: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of ACL*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Zhongquan Jian, Ante Wang, Jinsong Su, Junfeng Yao, Meihong Wang, and Qingqiang Wu. 2024. Emo-trans: Emotional transition-based model for emotion recognition in conversation.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Pierre Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Elsie Hanna, Florian Bressand, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jihwan Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis (WASSA)*.

Shivani Kumar, Ishani Mondal, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Explaining (sarcastic) utterances to enhance affect understanding in multi-modal dialogues. In *Proceedings of ACL*.

Lola Le Breton, Quentin Fournier, Mariam El Mezouar, John X. Morris, and Sarath Chandar. 2025. Neobert: A next-generation bert. *arXiv preprint arXiv:2502.19587*.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llm framework. In *Proceedings of EMNLP*.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. 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). International Committee on Computational Linguistics.

Wei Li, Luyao Zhu, Rui Mao, and Erik Cambria. 2023. Skier: A symbolic knowledge integrated model for conversational emotion recognition. In *Proceedings of AAAI Conference on Artificial Intelligence*.

Zixiang Li, Zixiang Zhou, Qian Wang, Pengfei Cao, Bing Liu, and Ruifeng Xu. 2021. Skaig: A psychological knowledge-aware interaction graph for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of ICCV*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Katarína Machová, Mária Szabóová, Ján Paralič, and Jozef Mičko. 2023. Detection of emotion by text analysis using machine learning. *Frontiers in Psychology*, 14:1190326.

Navonil Majumder, Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2020. Dialoguexl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of AAAI*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of AAAI*.

Patrícia Pereira, Helena Moniz, and João Paulo Carvalho. 2022. Deep emotion recognition in textual conversations: A survey. *arXiv preprint arXiv:2211.09172*.

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

Lu Qin, Wen Chen, Zhiqiang Yu, Yong Xu, and Yuxuan Zhang. 2023. Bert-erc: Emotion recognition in conversations using pre-trained language models. *IEEE Transactions on Affective Computing*.

Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories. *arXiv preprint arXiv:2102.02000*.

Geng Tu, Bin Liang, Bing Qin, Kam-Fai Wong, and Ruifeng Xu. 2023. An empirical study on multiple knowledge from chatgpt for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics*.

Geng Tu, Jintao Wen, Cheng Liu, Dazhi Jiang, and Erik Cambria. 2022. Context-and sentiment-aware networks for emotion recognition in conversation. *arXiv preprint arXiv:2211.09172*.

Zhunheng Wang, Xiaoyi Liu, Mengting Hu, Rui Ying, Ming Jiang, Jianfeng Wu, Yalan Xie, Hang Gao, and Renhong Cheng. 2024. Ecok: Emotional commonsense knowledge graph for mining emotional gold.

Sayyed Zahiri and Jinho D. Choi. 2019. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Proceedings of ACL*.

Sayyed M. Zahiri and Jinho D. Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 141–147. Association for Computational Linguistics.

Huan Zhao, Xupeng Zha, and Zixing Zhang. 2024. EmoTransKG: An innovative emotion knowledge graph to reveal emotion transformation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12098–12110, Bangkok, Thailand. Association for Computational Linguistics.