PFA-ERC: Psuedo-Future Augmented Dynamic Emotion Recognition in Conversations

Tanmay Khule*
University of Western Ontario
London Ontario
tkhule@uwo.ca

Rishabh Agrawal*
University of Western Ontario
London Ontario
ragrawa9@uwo.ca

Apurva Narayan
University of Western Ontario
London Ontario
apurva.narayan@uwo.ca

Abstract

AI systems' ability to interpret human emotions and adapt to variations is becoming more crucial as AI gets embedded into everyone's daily lives. Emotion Recognition in Conversations (ERC) is based on this fundamental challenge. Current state-of-the-art technologies in ERC are limited due to the need for future information. We introduce High-Dimensional Temporal Fusion Transformer (HiTFT), a time-series forecasting transformer that predicts pseudo-future information to overcome this constraint. This retains the models' dynamic nature and provides future information more efficiently than other methods. Our proposed method combines pseudo future embeddings with an encoder that models the speaker's emotional state using past and pseudo-future information as well as inter and intra speaker interactions; these speaker states are then passed through a decoder block that predicts the inferred emotion of that utterance. We further evaluate our method and show that it achieves state of the art performance on three ERC datasets - MELD, EmoryNLP, & IEMOCap.

1 Introduction

Directly or indirectly, chatbots and language models have become a crucial component of cyber infrastructure. With the high penetration of Large Language Models (LLMs) in various domains, such as healthcare, law, and many other safety-critical systems, any unexpected output generated by these models could lead to catastrophic consequences. The goal of Affective Computing Systems is to correctly identify users' emotions based on their responses and direct future conversations toward a desired emotion such as happiness or neutrality. Emotion Recognition in Conversations (ERC) is a field of active research in Affective Computing. The ERC task goes beyond the conventional recognition of emotions at the sentence level; it considers the complex interplay of conversational context and speaker's emotional states across multiple utterances. An example of our proposed ERC system in action is described in Figure 1.

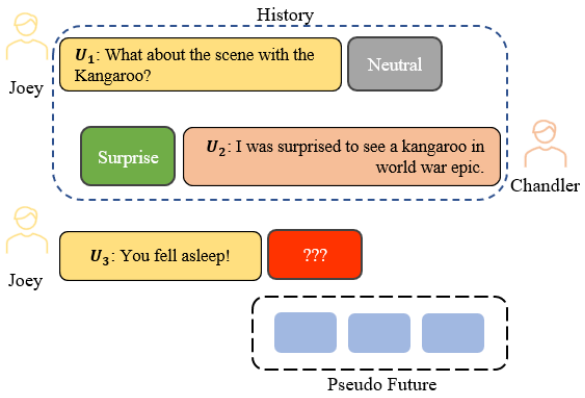

Figure 1: In this conversation, the speaker is at turn U_3 and U_1 & U_2 are past dialogues. We predict future context as pseudo-future, improving emotion inference for U_3.

Researchers have developed several methods to effectively exploit conversational features, from prompt engineering (Lei et al., 2023) to sophisticated speaker interaction tracking systems (Song et al., 2023; Bao et al., 2022; Guo et al., 2024). These systems depend on future dialogue discourse to accurately predict the emotion of the current utterance. This reliance limits their applicability in real-time scenarios where such information is unavailable. Wei et al. have explored the use of pseudo-future contexts but with significant computational overhead, as they rely on Pre-trained LLMs with billions of parameters for extracting these features. While pre-trained language models offer extensive linguistic knowledge, their computational overhead limits system usability in real-time edge systems equipped with low-power hardware. This highlights the need for models that can operate efficiently without sacrificing the accuracy of

* Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 16196–16207
November 12-16, 2024

ERC systems in real-time use cases.

This paper introduces a novel dynamic forecasting method that significantly reduces computational overhead while enhancing real-time responsiveness for ERC tasks. Unlike existing methods that rely heavily on pre-trained LLMs for future context prediction, our method leverages a Temporal Fusion Transformer (TFT)-based architecture to generate pseudo-future embeddings. This time-series approach to ERC, which forecasts emotional contexts rather than relying on computationally expensive LLMs, not only achieves comparable or superior accuracy but also improves inference speed and memory efficiency. Our contributions are particularly focused on the following key areas:

1. **Exploiting Pseudo-Future as a Time-Series Forecasting Problem:** We propose a novel approach to model future context as a time-series forecasting problem. Unlike previous work that directly generates future utterances using pre-trained LLMs, our method focuses on forecasting future token embeddings based on known current and past contexts. This approach not only reduces the need for large-scale models but also significantly improves computational efficiency, achieving real-time performance with an average inference time of 0.5s per utterance.

2. **Speaker Modelling for Emotion Prediction:** Our Speaker State Encoder dynamically captures both intra-speaker and inter-speaker dependencies, which are critical for emotion recognition in conversations. By integrating this with our pseudo-future context prediction, we improve the system's ability to track emotional trajectories across multiple speakers. This detailed speaker modeling allows us to outperform previous state-of-the-art methods in handling longer, more complex dialogues, where emotional dynamics are harder to predict.

Our approach represents a significant step forward in balancing the trade-offs between efficiency and accuracy in ERC systems. By reformulating future context prediction as a time-series problem and leveraging TFT's capability to handle multi-dimensional data, we create an ERC model that not only achieves state-of-the-art performance on challenging datasets but also operates efficiently on real-time, low-power hardware.

2 Related Methods

Traditional approaches in Emotion Recognition in Conversations (ERC) have primarily focused on analyzing discrete emotions from static texts, often overlooking the dynamic nature of conversations and speaker interactions. The evolution towards text-based ERC has seen the adoption of advanced methodologies, particularly in modeling conversation context.

Use of RNNs and Transformers: Early ERC studies relied on Transformer models and Recurrent Neural Networks (RNNs). These methods captured the continuous and inter-related nature of dialogues. DialogXL (Shen et al., 2020) uses a memory-augmented Transformer XL to exploit the hierarchical structure of conversations. DialogueCRN (Hu et al., 2021) incorporates a reasoning module using RNNs for contextual features. Advanced designs like SGED (Bao et al., 2022) include a speaker-guided decoder network with an attention-based speaker state encoding system. However, RNNs suffer from vanishing gradient issues, limiting modelling long-term context dependencies, while transformers can be computationally expensive, posing challenges for real-time applications.

Use of External Knowledge: External heterogeneous data has been used to improve ERC's contextual relations. DialogueRNN (Majumder et al., 2019) combines global representation embeddings with RNNs. COSMIC (Ghosal et al., 2020) utilizes COMET (Bosselut et al., 2019) embeddings to address emotion changes and misclassification issues. TODKAT (Zhu et al., 2021) integrates topic-driven context modeling with COMET features. Recent work by InstructERC (Lei et al., 2023) leverages pre-trained LLMs like LLama2 (Touvron et al., 2023) and GPT-3(Brown et al., 2020) for context modelling. Despite these approaches' impressive performance, they have significant drawbacks. Integrating external knowledge sources can cause domain adaptation issues and increase memory and storage requirements, limiting practical deployment.

Use of Pseudo-Future Knowledge: Use of pseudo-future context is a relatively unexplored domain. To the best of our knowledge, only one work exploiting the pseudo-future context (Wei et al., 2023b) achieves convincing results. This work is also limited in its use as they leverage a GPT-2(Sanh et al., 2019) model for future predic-

tion, which limits its usability on low-power hardware.

Contrastive Learning Solutions: Song et al. introduces a method that uses contrastive learning to improve emotional representation in dialogues, leading to better performance. Similarly, Yu et al. (2024) employs contrastive learning techniques to capture dynamic conversational contexts more effectively. CoG-BART(Li et al., 2022) adapts supervised contrastive learning to make different emotions mutually exclusive to identify similar emotions better. These methods need careful sample design and struggle with generalization across conversational scenarios, impacting model effectiveness and adaptability.

3 Methodology

The primary goal of ERC systems is to identify the underlying emotion for each utterance in a dialogue. Formally, a dialogue in our context is defined as a sequence of utterances, denoted as $D = \{u_1, u_2, ..., u_N\}$ and corresponding emotions $E = \{e_1, e_2, ..., e_N\}$, where N represents the total number of utterances in the conversation. $e_i \in C$ where $C = \{C_1, C_2, ...C_m\}$ for m emotion classes. Each utterance u_i consists of multiple tokens(words), and each utterance is linked to a unique speaker s_i. All speakers involved in dialogue D are represented by S_D = $\{s_1$,s_2,s_3,..$s_k\}$ where $k \geq 1$. For each utterance, we store a list of speaker and utterance pairs in order of occurrence, which helps us to map the speakers with their utterances. This mapping is used later to decode inter and intra-speaker dependency vectors.

3.1 Feature Extraction

For utterance-level feature extraction, we format our input as suggested by (Kim and Vossen, 2021). For each utterance u_i, we prepend the speaker name s_i linked to the utterance, enclosing this with EOS tokens on each side. Before this EOS token, we append the past eight utterances from u_{i-1} to u_{i-8}, the joint string is then enclosed by EOS tokens. This formatted input is then passed to the pre-trained RoBERTa Large model(Liu et al., 2020). For any given utterance u_i, the input array is formed by appending a [CLS] token at the beginning: [CLS], $x_{i1}, x_{i2}, ..., x_{in_i}$, where x_i is byte pair encoded token(Devlin et al., 2019). This [CLS] token serves as a context token, containing the representation of that utterance. We extract the [CLS] token in the final layer for each u_i; we call it h_i. For a dialogue $D = \{u_1, u_2, ..., u_N\}$, we extract the corresponding context tokens represented by $H_d = [h_1, h_2, ..., h_n]$.

3.2 Pseudo Future Extraction

Traditional sequence models like RNNs struggle with extended temporal relationships. Our novel High-Dimensional Temporal Fusion Transformer (HiTFT) addresses this by accurately predicting high-dimensional pseudo-future utterance embeddings to enhance emotion recognition. The standard Temporal Fusion Transformer (TFT) (Lim et al., 2021) leverages transformer architecture to implement a temporal attention mechanism, dynamically adjusting attention scores across previous time steps. TFTs, designed for multivariate time series forecasting, integrate known, unknown, and static variables—combining time-variant features with established correlations and time-dependent features without direct correlations with time-invariant attributes.

A key aspect of TFT is the Variable Selection Network (VSN), which dynamically identifies relevant variables at each time step. This allows the model to adaptively concentrate on the most informative features. Additionally, the TFT architecture incorporates attention encoders for known, unknown, and static covariates. These encoders capture the influence of both static and historical features on the observed time series, enriching the context for future time-step predictions. Lim et al. describe the VSN network as follows:

$$\text{GLU}(\mathbf{x}) = \sigma(\mathbf{W}_x.x + \mathbf{b}_x) \odot (\mathbf{W}_c.x + \mathbf{b}_c) \quad (1)$$

$$\begin{aligned} \text{GRN}(v, u) &= \text{Norm}(v + \mathbf{GLU}(\mathbf{W}_v.\mu + \mathbf{b}_v)) \\ \mu &= \text{ELU}(\mathbf{W}_a.v + \mathbf{W}_b.u + \mathbf{b}_a)) \end{aligned} \quad (2)$$

$$\begin{aligned} \mathbf{VSN}(\lambda, c) = &\sum_{j=1}^{d} \text{Softmax}(\mathbf{GRN}(\lambda_t, c))^{(j)} \\ &\odot \mathbf{GRN}^{(j)}(\lambda_t^{(j)}, c) \end{aligned} \quad (3)$$

Where, $\mathbf{x}$ represents the input at time step t to the Gated Linear Unit (GLU). $\mathbf{W}_{x,c}$, $\mathbf{b}_{x,c}$ are learnable parameters for the GLU with σ representing the sigmoid activation. GLUs provide the flexibility to suppress any parts of the architecture that are not required for a given dataset. Gated Residual Network (GRN) takes in a primary input v and an optional context vector u

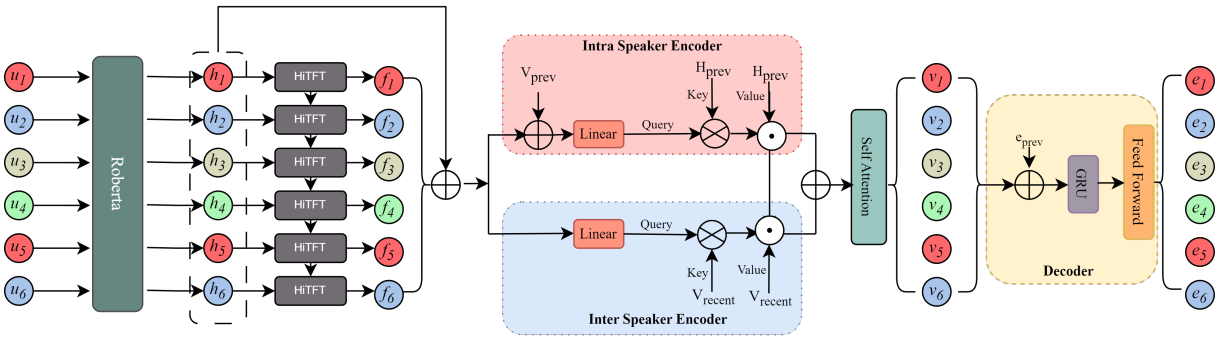

Figure 2: A detailed overview of the proposed method, where u_i are input utterances formatted as described in Section 3.1, h_i is the feature vector extracted from Roberta. HiTFT predicts the pseudo future embeddings f_i. After that we pass it through our encoder and decoder to get our predicted emotion e_i

and gives model the flexibility to apply non-linear transformations only where needed. VSN does a weighted addition of the overall input $\lambda \in R^d$ and optional static context c at each time-step. Per-feature $\mathbf{GRN}^{(j)}, j \in \{1, 2, ...d\}$ blocks for both static covariates and time-dependent covariates provide instance-wise variable selection. All static, past and future inputs make use of separate VSNs and are used to augment the context vector with temporal and static features at time step t as H_t, described in (Lim et al., 2021) as:

$$\mathbf{H}_t = \mathbf{GRN}(L_{x,t} \oplus A_t, E_s) \quad (4)$$

$$\hat{y}_t = \text{FC}(\text{Norm}(\mathbf{GLU}(\mathbf{GRN}(\mathbf{H}_t)) + L_{x,t})) \quad (5)$$

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 \quad (6)$$

Here, $\mathbf{E}_s \in R^{d'_s}$ represents the encoded representation of static covariates, $\mathbf{L}_{x,t} \in R^{T \times d}$ are the locally enhanced features and $\mathbf{A}_t$ is the output of the sequence model at time step t. The context enrichment network is represented by another GRN (Lim et al., 2021). Equation 5 represents the forecasting step where, $\hat{y}_t$ and $_t$ are the predicted pseudo-future values and ground truth future values. $\mathcal{L}$ is the loss function, representing the mean squared error between predictions and observed future time-series variables.

Despite its strengths, TFT faces certain limitations in natural language processing (NLP) applications: **1. Dependence on Known Future Variables:** Predicting an observed variable at time-step $\hat{t}$ necessitates the corresponding known variable at $\hat{t}$. This dependency constrains its application in real-time NLP scenarios, where only past and current variables (text transcripts of conversations) and static features (participating speakers) are available. **2. Scalability Issues:** The original TFT architecture struggles to handle the high dimensionality of inputs and varying sequence lengths inherent in NLP tasks during training. For instance, RoBERTa embeddings are high-dimensional (1024), and conversation sequences can vary significantly in length, posing challenges for direct application of the standard TFT model.

We address the first issue by making the future observations independent from the known variables during prediction. We modify the context enrichment step to only include the static and historical information and combine this with a multi head self-attention block. This enables us to directly parameterize the output on past context and also predict multiple time steps in the future without the need of known future context. To tackle the second issue, we redesign the network to incorporate variable sequence lengths by introducing input batching and sequence padding operations. This modified HiTFT algorithm is given in Algorithm 1.

3.3 Speaker Emotion Encoder

To properly model context in Emotion Recognition in Conversations (ERC), we developed an encoder network that aims to capture both inter- and intra-speaker dependencies. This section details the formulation of these dependencies.

Intra-Speaker Dependency Formulation

For a given utterance u_i linked to a speaker $s_{(u_i)}$, we concatenate the most recent state vector v_i of the same speaker alongside the entire previous context $\mathcal{H}_p = [h_1, h_2, ..., h_i] \in R^{h \times i}$. This approach models the intra-speaker dependency by considering the flow of dialogue.

Algorithm 1 Hi-TFT

Input: Observed time series $\mathbf{y}$, known covariates $\mathbf{x}_k$, unknown covariates $\mathbf{x}_u$, static features $\mathbf{S}$
Output: Predicted future values $\hat{\mathbf{y}}$

$\tau \leftarrow 5$, $b \leftarrow 128$
$x_k \leftarrow \text{Pad}(x_k, \tau)$, $x_u \leftarrow \text{Pad}(x_u, \tau)$
$X_k \leftarrow \text{Stack}([x_{k,1}, x_{k,2}...x_{k,b}])$
$X_u \leftarrow \text{Stack}([x_{u,1}, x_{u,2}...x_{u,b}])$

$\hat{X}_k \leftarrow \text{VSN}(X_k)$, $\hat{X}_u \leftarrow \text{VSN}(X_u)$
$\hat{S} \leftarrow \text{VSN}(S)$

$A \leftarrow \text{LSTM}(\hat{X}_k \oplus \hat{X}_u)$
$E_S \leftarrow \text{GRN}(\hat{S})$
$L_X \leftarrow \text{GRN}(\hat{X}_k \oplus \hat{X}_u, E_S)$

$L_X \leftarrow \text{LayerNorm}(\text{GLU}(L_X) + A)$
$H \leftarrow \text{GRN}(L_X, E_S)$
$M \leftarrow \text{AttentionMask}(H)$

$\hat{H} \leftarrow \text{SelfAttention(H, M)}$
$\hat{H} \leftarrow \text{LayerNorm}(\text{GLU}(\hat{H}) + H)$
$\hat{H} \leftarrow \text{LayerNorm}(\text{GLU}(\hat{H}) + L_X)$
$\hat{y} \leftarrow \text{Linear}(\text{SelfAttention}(\hat{H}, M))$

Initially, the process starts with the previous speaker state vector of v_i combined with the context vector h_i. We compute the intra-speaker query vector q_i^{intra} as follows:

$$q_i^{intra} = W_q^{intra} \cdot [v_i \oplus h_i] + b_q^{intra}, \quad (7)$$

An attention mechanism is then introduced to model the intra-speaker context, with the query vector q_i^{intra} acting as the query and the previous context c_i serving as the key and value in the attention framework. This mechanism generates the intra-speaker state vector v_i^{intra} as follows:

$$\alpha_i^{intra} = \text{softmax}(W_1^{intra} \cdot q_i^{intra} + b_1), \quad (8)$$

$$v_i^{intra} = \alpha_i^{intra} \circ c_i, \quad (9)$$

Inter-Speaker Dependency Formulation

To model inter-speaker dependencies, we use the context vector h_i. The key vector k_i for the attention mechanism is calculated prior local information $\{v_j | j < i\}$ linked to speaker $s_{(u_i)}$. This approach captures the latent inter-speaker dependency. To derive the inter-speaker state vector for u_i, we apply the following procedure:

$$q_i^{inter} = W_q^{inter} h_i + b_q^{inter}, \quad (10)$$

$$\alpha_i^{inter} = \text{softmax}(W_a^{inter}(q_i^{inter} \otimes k_i) + b_2), \quad (11)$$

$$v_i^{inter} = \alpha_i^{inter} \circ k_i, \quad (12)$$

Combining Speaker Vectors with Self-Attention

After obtaining the intra-speaker and inter-speaker state vectors, v_i^{intra} and v_i^{inter}, we combine them into a unified representation $v_i^{combined}$:

$$v_i^{combined} = v_i^{intra} \oplus v_i^{inter},$$

Next, we apply a self-attention mechanism to the combined vectors $v_i^{combined}$ across all utterances in the dialogue.

$$V_i = \text{SelfAttention}(V_i^{combined})$$

This final speaker state vector captures the complete emotional context of the conversation, integrating both intra-speaker and inter-speaker dependencies, allowing the ERC system to make more accurate predictions about the emotional states conveyed in each utterance.

3.4 Utilizing Future for ERC

To effectively utilize pseudo-future embeddings in our ERC model, we experiment with various architectural and simple ways. We evaluated several hypothesis like creating a seperate branch to model future and current context interplay, simple concatenation with current context. We decided to take a straightforward approach, i.e. to concatenate the pseudo-future context f_i with the current context h_i. This new context vector $h_i^{'}$ replaces h_i in the model. This method's simplicity and low computational cost make it easy to implement, allowing seamless integration of future information. Any context modelling done on this concatenated data will ensure that future information is not lost.

4 Decoder

Our decoder network employs a Gated Recurrent Unit (GRU) architecture to maintain simplicity and efficiency. When processing an utterance u_i, we align the speaker state vector with the utterance's representational vector:

$$m_i = \text{ReLU}(v_i \odot (W_m h_i + b_m)^T), \quad (13)$$

where W_m and b_m are model parameters.

The match vector m_i is then concatenated with the emotional embedding of the previously predicted emotion and passed to the GRU as suggested

by (Wang et al., 2020) enabling the decoder to track the dynamic interplay of emotions:

$$o_i = \text{GRU}([m_i \oplus e_{i-1}], o_{i-1}), \quad (14)$$

Finally, we combine the utterance representation with the GRU output and pass this through a feedforward neural network to predict the emotion:

$$z_i = \text{ReLU}(W_o[h_i \oplus o_i] + b_o), \quad (15)$$

$$P_i = \text{softmax}(W_z z_i + b_z), \quad (16)$$

yielding the predicted emotion $\hat{y}_i$.

For model training, we minimize the cross-entropy loss:

$$\mathcal{L}(\theta) = -\sum_{j=1}^{M}\sum_{t=1}^{N_j} \log P_{j,t}[y_{j,t}], \quad (17)$$

5 Implementation Details

In the pre-training step, our High-Dimensional Temporal Fusion Transformer (HiTFT) was trained for 100 epochs with initial learning rate of 1e-4 and weight decay of 1e-3. Once the pre-training is complete, the HiTFT model parameters are frozen during further fine-tuning of the speaker modeling encoder-decoder network. Additionally, the speaker modeling network was fine-tuned on past, current and pseudo-future predicted TFT embeddings, with an initial learning rate of 5e-5, followed by a linear decay after a 20% epoch warm-up period. We use the PyTorch framework for implementing our method and use the AdamW optimizer. The trained model processes dialogue data in real-time with an average inference time of 0.5 milliseconds per conversation.

5.1 Datasets

Our research uses MELD (Poria et al., 2019), IEMOCAP (Bansal et al., 2022), and EmoryNLP (Zahiri and Choi, 2017), which offer unique speaker modeling and emotion recognition characteristics. Note that we only use text modality from these datasets.

MELD: This dataset adds 1,400 dialogues and 13,000 utterances from "Friends." to EmotionLines. It has auditory, visual, and textual data labeled with *anger, disgust, fear, joy, neutral, sadness, and surprise*. Its multimodality and extensive annotations make it perfect for multimodal emotion identification model training.

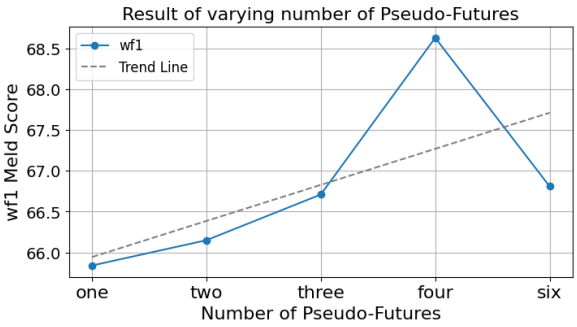

Figure 3: The graph shows the impact of varying the number of predicted future time steps (τ) for pseudo-future utterances on wF1 as τ increases from one to six. wF1 increases from 65.84 to 68.63 utterances at $\tau = 4$, then drops to 66.81.

IEMOCAP:IEMOCAP contains 12 hours of audiovisual data, including professional actors' scripted and unscripted dialogues. It lists *anger, happiness, sadness, and neutral*. The dataset's detailed emotional annotations and various interactions are necessary for robust emotion recognition algorithms.

EmoryNLP: This dataset analyzes multi-party discussions from "Friends," annotated with six emotions: *sad, mad, terrified, powerful, serene, and cheerful*. Textual and auditory data provide complete emotion analysis. EmoryNLP's multi-party focus illuminates complicated emotional relationships.

6 Results

To determine the optimal number of future time steps for pseudo-future embeddings, we increment the number of predicted future time steps τ from one to six as shown in Figure 3 and evaluate model performance on MELD dataset. Initially, wF1 scores rose exponentially due to valuable additional information. However, further increases led to a decline in wF1 scores as excessive data points introduced noise, making it hard for the model to decode anything from current utterance. We fix $\tau = 4$ for all further analysis.

6.1 Baselines

Table 1 provides a comparative analysis of various ERC models across MELD(Poria et al., 2019), IEMOCAP(Bansal et al., 2022), and EmoryNLP(Zahiri and Choi, 2017) datasets, using weighted F1 scores. Our method establishes a new state-of-the-art on the IEMOCAP dataset, outperforming InstructERC (Lei et al., 2023), which

Model Name	MELD weighted F1	IEMOCAP weighted F1	EmoryNLP weighted F1	Type
SGED 2022	65.46	68.53	40.24	RNN based
InstructERC 2023	**69.15**	71.39	41.39	PLM based
EmoBERTA 2021	64.55	68.57	–	Transformer based
CoG-BART 2021b	64.81	66.18	39.04	PLM based
SPCL 2022	67.25	69.24	40.94	Contrastive Learning based
SKAIG 2021a	65.18	66.98	38.88	GNN-based
EACL 2024	67.12	70.41	40.24	Contrastive Learning based
CoMPM 2022	66.52	66.33	38.93	Transformer Memory based
ERCMC 2023a	65.43	66.51	38.90	Transformer Memory based
Our: Enc+Dec+Real	66.34	66.24	40.23	RNN based
Our: Enc+Dec+HiTFT	68.63	**77.34**	**42.94**	Pseudo Future based

Table 1: Comparison of Our Model with various ERC Models on Three Datasets

uses a 7B parameter pre-trained LLM. Crucially, our model achieves this with significantly fewer parameters and resources.

Our approach surpasses models that incorporate future dialogue information, such as COG-BART (Li et al., 2021a), EmoBERTa-large (Kim and Vossen, 2021), and InstructERC (Lei et al., 2023), by generating pseudo-future CLS token embeddings instead of full future utterances. This reduces computational overhead and accelerates inference without sacrificing accuracy, making it more suitable for real-time applications compared to COG-BART, which relies on generating future utterances.

Additionally, our method establishes new benchmarks on all three datasets, outperforming real-time systems like SPCL (Song et al., 2022), EACL (Yu et al., 2024), and SGED (Bao et al., 2022), while maintaining computational efficiency. Our model processes each utterance in 0.5 seconds, providing a 3x speedup over InstructERC's 1.5 seconds. It also requires just 12GB VRAM during inference, compared to InstructERC's 28GB, representing a 50% reduction in memory usage, underscoring its suitability for resource-constrained, real-time environments.

HiTFT excels in leveraging the temporal dynamics of dialogues, evidenced by its enhanced performance on datasets with longer dialogue sequences. On the IEMOCAP dataset, where the average dialogue length is 52 utterances, integrating HiTFT-generated pseudo-future significantly outperforms other models due to its superior ability to capture and predict long-term trends through pseudo-future contexts. In contrast, on the MELD dataset with an average dialogue length of just 9 utterances, the pseudo-future integration shows only a marginal performance increase, as the shorter sequences provide limited scope for trend modeling. However, the model regains strong performance on the EmoryNLP dataset, which features medium-length dialogues averaging 16 utterances, aligning well with HiTFT's capabilities.

Parts	wF1 MELD	wF1 IEMO
Enc + Dec + HiTFT	**68.63**	**77.34**
Dec + HiTFT	67.60	74.63
Enc + HiTFT	66.62	73.85
Enc + Dec	66.26	66.24

Table 2: Ablation Study, analysing the impact of different PFA-ERC components on overall performance for two datasets

6.2 Ablation Study

We systematically dismantled our ERC model architecture to assess the contributions of the Encoder, Decoder, and HiTFT components, using weighted F1 scores on MELD and IEMOCap datasets. The full configuration achieved optimal scores of 68.63 on MELD and 77.34 on IEMOCap, establishing a strong baseline. Removing HiTFT, responsible for integrating pseudo-future contexts, led to a significant performance drop to 66.26 and 66.24, highlighting its role in enhancing the model's anticipation of emotional shifts. Interestingly, removing the Encoder improved results on MELD but not IEMOCap, suggesting

Utt.No.	Text	Labels	V_1	V_2
Dialogue 1				
1	S_1: Why did he invite her here?	Angry	Frustrated	Angry
2	S_2: Oh, God, line.	Neutral	Frustrated	Neutral
3	S_1: Why does that bother you?	Neutral	Neutral	Neutral
4	S_1: She's been in New York three and a half years. Why, all of a sudden–	Angry	Angry	Angry
Dialogue 2				
1	S_1: Hi, how can I help you?	Neutral	Neutral	Neutral
2	S_2:lost my luggage. Your airline–	Frustrated	Neutral	Frustrated
3	S_1: I'm sorry.	Neutral	Neutral	Neutral
Dialogue 3				
1	S_1: Thank you for calling Sprint. We care about everybody. How can I help you?	Frustrated	Neutral	Frustrated
2	S_2: Hi. I've been on the phone for an hour trying to get a little discrepancy on my bill fixed. I was charged for two hundred dollars worth of calls that I didn't make.	Frustrated	Frustrated	Frustrated
3	S_1: Are you sure you didn't make them?	Neutral	Frustrated	Frustrated
4	S_2: I'm positive. They came from like another state.	Frustrated	Neutral	Frustrated

Table 3: Case Study to demonstrate how Pseudo-Future V_2 embeddings are more effective at capturing emotion transitions when compared to Real V_1 embeddings. Emotions with blue and red color represent correct and incorrect predictions respectively.

that MELD's shorter dialogues benefit less from modeling speaker dependencies. Removing the Decoder consistently reduced performance across both datasets. These ablation studies confirm that while each component contributes to performance, integrating all three is critical for optimal ERC, with HiTFT having the most substantial impact.

6.3 Case Study

To demonstrate the effectiveness of our HiTFT, we compared two versions of our model: V_1 uses real future embeddings, and V_2 uses pseudo-future embeddings. As shown in the results in Table 1, V_2 consistently outperforms V_1 by a significant margin, achieving superior emotion detection across all datasets, with performance improvements up to 7%.

To further investigate the reasons behind V_2's improved performance, we analyze their predictions on selected utterances as a case study, as shown in Table 3. In the examples, it is clear that V_2 always accurately predicts the utterance emotions at the initial time steps. The gain in performance of V_2 is a result of V_1 inability to effectively decipher relevant information from future events, which leads to the miss-classification of certain emotions at the start of a dialogue . Another intriguing observation is that V_2 demonstrates the ability to effectively decode emotions from ambiguous texts and easily adapt to changes in emotions, this can be seen in Dialogue 2. This strong adaptation to dialogue flow can be attributed to the VSN which dynamically weighs the importance of each feature, helping the encoder and decoder to focus on relevant parts of the context, thus yielding a better overall performance.

To analyze this further, we conducted two experiments focusing on the spectral properties and temporal smoothness of the two embeddings.

Experiment 1: Power Spectral Density Analysis While text embeddings are static representations at individual time steps, they can be viewed as a multivariate time series when analyzed over a sequence, such as in a dialogue. By treating this sequence of embeddings as a temporal signal, we applied Power Spectral Density (PSD) analysis to explore

the distribution of frequency components. We computed the PSD for both real future embeddings and Hi-TFT pseudo-future embeddings to examine and compare their temporal frequency characteristics.

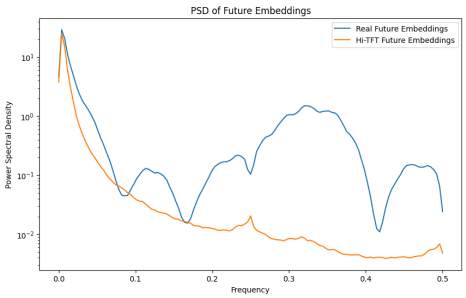

Figure 4: Power Spectral Density comparison between real future embeddings and Hi-TFT embeddings.

As depicted in Figure 4, the PSD curve for real future embeddings exhibits significantly higher power across most frequency bands, particularly in the mid-frequency range. This elevated power suggests the presence of substantial high-frequency noise or rapid fluctuations within the embeddings. This noise can introduce instability in tasks like emotion recognition in conversations. For effective modeling of emotional cues, consistent temporal patterns are crucial, as they provide a more reliable signal than transient or erratic fluctuations.

In contrast, the PSD curve for Hi-TFT future embeddings remains consistently lower across all frequency bands, indicating a smoother spectral profile with reduced high-frequency content. This decrease in temporal noise suggests that Hi-TFT embeddings exhibit fewer fluctuations and greater temporal stability, which is beneficial for the emotion recognition in conversations task that relies on modelling stable temporal dynamics.

Experiment 2: Mean Square Derivative Analysis To further assess temporal smoothness, we analyzed the Mean Square Derivative (MSD) of the embeddings, which measures the variance in their temporal derivatives.

Figure 5 illustrates that real future embeddings exhibit a wide range of MSD values, spanning from 0.1 to 0.7. This considerable variance indicates significant fluctuations over time, reflecting less smooth transitions and inconsistencies that could impair model performance.

In contrast, Hi-TFT future embeddings demonstrate extremely low MSD values, close to zero, signifying minimal variance in their temporal derivatives. This finding confirms that Hi-TFT embeddings maintain a high degree of temporal smoothness, ensuring consistent transitions that are beneficial for downstream applications requiring stable input representations.

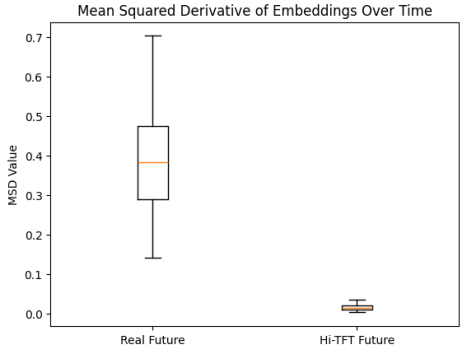

Figure 5: Mean Square Derivative comparison between real future embeddings and Hi-TFT embeddings.

7 Conclusion

In this study, we present a novel approach to ERC using a cutting-edge Hi-Dimensional Temporal Fusion Transformer (HiTFT) model built to forecast pseudo-future utterance embeddings. This methodology improves the ERC task by combining dynamic temporal dependencies and speaker context, making our proposed method better for modelling long range conversational interactions. Our method efficiently manages the complexities and varying lengths of dialogues. It also adeptly addresses the challenges of real-time processing and class imbalance inherent in traditional ERC systems. We perform rigorous evaluations of the proposed method using IEMOCAP, MELD, and EmoryNLP datasets and our model achieves state-of-the-art performance on IEMOCAP and EmoryNLP datasets.

This research uses advanced approaches to bring a fresh perspective to the ERC domain, creating a new standard for future research and applications. It will shape real-time, efficient ERC systems, enabling robust context-aware computing systems in everyday interactions.

8 Limitations

Extensive testing has highlighted a few limitations of our HiTFT-generated pseudo future embeddings, particularly their modest impact on the MELD dataset. This dataset's short dialogue sequences challenge HiTFT's ability to model dialogue flow effectively, as detailed in Section 6.3. HiTFT requires a number of past known inputs equal to the

future time steps it predicts. This condition hampers its performance in scenarios involving short dialogues, as it struggles to capture short-term sequence interactions.

To achieve minimal inference time for real-time system applicability, we employed a frozen HiTFT during training. This approach ensures that the encoder and HiTFT do not learn a shared embedding space. Replacing the frozen HiTFT model with a trainable version during the encoder and Hi-TFT's joint fine-tuning could promote a cohesive learning environment and further enhance the prediction accuracy.

Additionally, our current framework is designed to predict only four pseudo-future embeddings. To predict more embeddings, we would have to retrain the entire TFT, which poses scalability issues. Addressing this limitation could involve developing a more flexible architecture that allows for adjusting the number of embeddings without full retraining.

Lastly, training with a diverse dataset that includes a broad spectrum of dialogue lengths could improve the robustness and generalizability of HiTFT, making it more effective across different real-world scenarios. These refinements could further optimize HiTFT for varied applications by enhancing its architecture and training process, promising better performance and adaptability in environments with fluctuating dialogue lengths.

References

Keshav Bansal, Harsh Agarwal, Abhinav Joshi, and Ashutosh Modi. 2022. Shapes of emotions: Multimodal 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. International Conference on Computational Linguistics.

Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu. 2022. Speaker-guided encoder-decoder framework for emotion recognition in conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4051–4057. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. 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. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Lili Guo, Yikang Song, and Shifei Ding. 2024. Speaker-aware cognitive network with cross-modal attention for multimodal emotion recognition in conversation. *Knowledge-Based Systems*, 296:111969.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7042–7052, Online. Association for Computational Linguistics.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *ArXiv*, abs/2108.12009.

Joosung Lee and Wooin Lee. 2022. CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *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. Association for Computational Linguistics.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *Preprint*, arXiv:2309.11911.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021a. 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. Association for Computational Linguistics.

Shimin Li, Hang Yan, and Xipeng Qiu. 2021b. Contrast and generation make bart a good dialogue emotion recognizer. *ArXiv*, abs/2112.11202.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and generation make bart a good dialogue emotion recognizer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:11002–11010.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6818–6825.

Soujanya Poria, Navonil Majumderd, Rada Mihalceae, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, PP:1–1.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC^{2}Workshop*.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixiang Xie. 2020. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *AAAI Conference on Artificial Intelligence*.

Rui Song, Fausto Giunchiglia, Lida Shi, Qiang Shen, and Hao Xu. 2023. Sunet: Speaker-utterance interaction graph neural network for emotion recognition in conversations. *Engineering Applications of Artificial Intelligence*, 123:106315.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. 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. Association for Computational Linguistics.

Yinyi Wei, Shuaipeng Liu, Hailei Yan, Wei Ye, Tong Mo, and Guanglu Wan. 2023a. Exploiting pseudo future contexts for emotion recognition in conversations. *Preprint*, arXiv:2306.15376.

Yinyi Wei, Shuaipeng Liu, Hailei Yan, Wei Ye, Tong Mo, and Guanglu Wan. 2023b. Exploiting pseudo future contexts for emotion recognition in conversations. In *Advanced Data Mining and Applications*, pages 309–323, Cham. Springer Nature Switzerland.

Fangxu Yu, Junjie Guo, Zhen Wu, and Xinyu Dai. 2024. Emotion-anchored contrastive learning framework for emotion recognition in conversation. *arXiv preprint arXiv:2403.20289*.

Sayyed M. Zahiri and Jinho D. Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *ArXiv*, abs/1708.04299.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, Online. Association for Computational Linguistics.

A Appendix

A.1 Risks

Deploying Emotion Recognition in Conversation (ERC) systems poses privacy risks, as they process sensitive data to analyze emotions, risking misuse or unauthorized access. ERC's accuracy varies by demographic factors, potentially leading to biases and discrimination in applications like hiring or law enforcement. Over-reliance on ERC technology may also degrade human judgment and empathy in interpersonal interactions. Mitigating these risks involves creating transparent, inclusive ERC systems with robust data protection and continuously monitoring to prevent biases. This ensures that ERC technology supports rather than undermines fair and empathetic human communication.

A.2 Training Environment

All TFT models were trained for 400 epochs for each of the analysed datasets. Speaker Encoder and Decoder Networks are trained for 20 epochs on MELD and 40 epochs on EmoryNLP and IEMOCAP, averaging results from the best-performing epoch across 5 seeds. Experiments were conducted on a server with 200GB RAM, 1x 50GB Nvidia A6000 GPU, and 1x Intel Xeon Gold 6226R processor. The model training and evaluation for emotion classification took around 12 hours each for the three datasets.

A.3 Licences

License details for different components used in the paper:

A.3.1 Datasets

We use three benchmark datasets MELD, IEMOCap, EmoryNLP. We have GPL 3.0 license for MELD and Custom (research-only, non-commercial) for the rest two.

A.3.2 Pre-trained models

Our paper uses a few pre-trained models, their links are as follows:

1. RoBERTa - MIT
2. EmoBERTa - MIT

A.3.3 Release

Upon acceptance, the PFA-ERC code will be released on github under a research only non-commercial liscence. The used datasets will not be released will not be released as a part of source code as it cannot be distributed by a third party. Readers are encouraged to seek permission from the original author for these datasets.