

Article

EmoBERTa–CNN: Hybrid Deep Learning Approach Capturing Global Semantics and Local Features for Enhanced Emotion Recognition in Conversational Settings

Mingfeng Zhang ¹, Aihe Yu ², Xuanyu Sheng ¹, Jisun Park ³, Jongtae Rhee ⁴ and Kyungeun Cho ^{5,*}¹ Department of Computer Science and Artificial Intelligence, Dongguk University-Seoul, 30 Pildongro 1-gil, Jung-gu, Seoul 04620, Republic of Korea; mingfeng@dgu.ac.kr (M.Z.); shengxuanyu@dgu.ac.kr (X.S.)² Department of Autonomous Things Intelligence, Dongguk University-Seoul, 30 Pildongro 1-gil, Jung-gu, Seoul 04620, Republic of Korea; aihe@dgu.ac.kr³ NUI/NUX Platform Research Center, Dongguk University-Seoul, 30 Pildongro-1-gil, Jung-gu, Seoul 04620, Republic of Korea; jisun@dongguk.edu⁴ Industrial Artificial Intelligence Researcher Center, Dongguk University-Seoul, 30 Pildongro 1-gil, Jung-gu, Seoul 04620, Republic of Korea; jtrhee@dongguk.edu⁵ Department of Computer Science and Artificial Intelligence, College of Advanced Convergence Engineering, Dongguk University-Seoul, 30 Pildongro 1-gil, Jung-gu, Seoul 04620, Republic of Korea

* Correspondence: cke@dongguk.edu

Abstract

Emotion recognition in conversations is a key task in natural language processing that enhances the quality of human–computer interactions. Although existing deep learning and Transformer-based pretrained language models have shown remarkably enhanced performances, both approaches have inherent limitations. Deep learning models often fail to capture the global semantic context, whereas Transformer-based pretrained language models can overlook subtle, local emotional cues. To overcome these challenges, we developed EmoBERTa–CNN, a hybrid framework that combines EmoBERTa’s ability to capture global semantics with the capability of convolutional neural networks (CNNs) to extract local emotional features. Experiments on the SemEval-2019 Task 3 and Multimodal EmotionLines Dataset (MELD) demonstrated that the proposed EmoBERTa–CNN model achieved F1-scores of 96.0% and 79.45%, respectively, significantly outperforming existing methods and confirming its effectiveness for emotion recognition in conversations.

Keywords: pre-trained language model; deep learning; emotion recognition**MSC:** 68T07; 68T50; 68U15

Academic Editors: Grigoreta-Sofia Cojocar, Adriana-Mihaela Guran and Laura-Silvia Dioşan

Received: 2 July 2025

Revised: 23 July 2025

Accepted: 26 July 2025

Published: 29 July 2025

Citation: Zhang, M.; Yu, A.; Sheng, X.; Park, J.; Rhee, J.; Cho, K.

EmoBERTa–CNN: Hybrid Deep Learning Approach Capturing Global Semantics and Local Features for Enhanced Emotion Recognition in Conversational Settings. *Mathematics* **2025**, *13*, 2438. <https://doi.org/10.3390/math13152438>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Advancements in natural language processing (NLP) and artificial intelligence (AI) have positioned emotion recognition in conversations (ERC) as a critical capability for machines to interpret and respond to human emotions. Unlike textual emotion recognition (TER), which typically processes static text or isolated sentences, ERC is dedicated to analyzing multi-turn dialogues to assess the emotional content of utterances within a full conversational context. The ability to understand and process emotions within dialogues is paramount for advancing human–computer interaction (HCI) systems, e.g., virtual assistants and customer service bots. Consequently, integrating emotional intelligence into ERC systems can foster empathetic and human-like natural communication [1–3].

Traditional machine learning methods rely heavily on handcrafted features and shallow classifiers, which often suffer from limited generalization and poor adaptability to complex conversational structures. Consequently, to improve their performance, deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can now learn hierarchical representations from data. However, these models often struggle to capture long-range dependencies and global contextual semantics across multi-turn dialogues [4–6].

More recently, pretrained language models (PLMs) based on Transformer architectures, including bidirectional encoder representations from transformers (BERT), the robustly optimized BERT approach (RoBERTa), and their variants, have shown remarkable success in various NLP tasks, including ERC. These models excel at capturing global contextual information using self-attention mechanisms. Nonetheless, they have several limitations. PLMs, in particular, may overlook fine-grained local emotion cues, such as affective keywords or subtle tone changes, that are critical for accurately classifying emotion at the utterance level [7–9].

In ERC, global features encapsulate the overarching semantic context, while local features focus on fine-grained cues, including emotional keywords, short phrases, or sentiment shifts within individual utterances.

Given the strengths and limitations of these models, the development and application of hybrid approaches are becoming increasingly imperative for effectively capturing the interplay between broad semantic understanding and nuanced emotional signals. Compared to general pretrained models, such as BERT and RoBERTa, EmoBERTa is further pretrained on an emotion-aware corpus, enabling it to more effectively capture global emotional semantic features within conversations. Additionally, this study incorporates an attention-pooling layer, which dynamically assigns different weights to convolutional features, thereby capturing subtle and critical emotional cues more effectively. In contrast, traditional pooling methods, such as average pooling or max pooling, lack this flexibility and are less capable of emphasizing locally important emotional information.

The core contributions of this study are as follows:

- (1) This study developed EmoBERTa–CNN, a novel hybrid model integrating pretrained emotion-aware bidirectional encoder representations from transformers (EmoBERTa) with CNN layers to simultaneously capture global contextual semantics and local emotional features. Experimental results demonstrate that this hybrid architecture outperforms single-model approaches in emotion recognition tasks;
- (2) This study integrated an attention-pooling layer into the CNN, which dynamically reweights convolutional features according to their relevance. This enables the network to effectively identify and capture critical emotional cues;
- (3) Extensive experiments on benchmark ERC datasets demonstrated that EmoBERTa–CNN significantly outperformed existing models in terms of the F1-score.

The remainder of this paper is organized as follows. Section 2 provides a review of related works, detailing traditional approaches, deep learning architectures, PLMs, and the application of attention mechanisms to conversational emotion recognition. Section 3 details the proposed EmoBERTa–CNN model, including its architecture, embedding strategy, and attention-pooling mechanism. Section 4 presents the experimental settings, datasets, evaluation metrics, and performance comparisons with state-of-the-art (SOTA) models. Section 5 presents the conclusions drawn from this study.

2. Related Works

ERC, a task within NLP, focuses on identifying and classifying emotional states, such as joy, anger, sadness, and surprise, expressed through multi-turn dialogues. Traditional tex-

tual emotion recognition (TER) typically targets isolated sentences or short user-generated posts. By contrast, ERC focuses specifically on multi-turn dialogues, wherein emotions can shift and evolve with each utterance and must, therefore, consider the dynamic and context-dependent nature of emotional expressions in conversation.

This section traces ERC development from early machine learning methods to modern deep learning models and PLMs. It further highlights the advantages of hybrid architectures that combine global semantic understanding with fine-grained emotional cues.

2.1. Traditional Methods

Early emotion recognition research predominantly employed supervised machine learning models, including naïve Bayes classifiers, support vector machine (SVM) models, and decision tree algorithms. These methods generally rely on handcrafted feature engineering, which requires significant domain expertise to extract syntactic, lexical, and emotional cues from the text.

Asghar et al. [10] and Gaind et al. [11] evaluated various traditional classifiers—such as naïve Bayes, SVM, J48 decision tree, and sequential minimal optimization (SMO) algorithms—on manually or weakly annotated datasets comprising user-generated online content, including tweets and general textual posts. Their feature sets commonly included term frequency–inverse document frequency vectors, part-of-speech tags, dependency relations, emoticon patterns, and emotion lexicons. Among the compared methods, the SVM consistently performed well, demonstrating its robustness in handling sparse high-dimensional features.

While these traditional methods offer interpretability and efficiency for short single-turn texts, they generally struggle with the complexities of multi-turn dialogues owing to their limited contextual modeling capabilities and reliance on manually defined features.

2.2. Deep Learning Methods

2.2.1. Traditional Deep Learning Model

Deep learning approaches have significantly advanced ERC by facilitating the automatic extraction of context-aware and hierarchical emotional features. Various architectures, such as RNNs, CNNs, graph convolutional networks (GCNs), and hybrid models, have demonstrated improved performance over traditional methods.

Bharti et al. [12] proposed a model that integrates a bi-directional gated recurrent unit (Bi-GRU) and CNN to capture temporal and local emotional patterns in single-turn texts, achieving competitive performance on six-class emotion datasets. Similarly, Jiang et al. [13] introduced a long short-term memory (LSTM)–CNN attention model for aspect-level sentiment classification, which incorporates an attention mechanism to dynamically weight context words related to target aspects, demonstrating the effective identification of fine-grained sentiment cues.

For multi-turn dialogue emotion recognition, Majumder et al. [14] developed DialogueRNN, which models speaker-specific emotional states and contextual flows by using GRUs. The model maintains multiple recurrent chains to track the global context, speaker embeddings, and emotional evolution, achieving strong results on multi-turn datasets such as interactive emotional dyadic motion capture (IEMOCAP) and AVEC. Extending this idea, Ghosal et al. [15] introduced DialogueGCN, which constructs conversation graphs based on speaker interactions and utterance chronology, applying GCNs to propagate emotion signals across graph nodes. This structure enhances interspeaker emotion modeling and achieves improved accuracy on datasets such as the multimodal emotion lines dataset (MELD) and IEMOCAP.

Despite these advancements, deep learning models face challenges in long-range context modeling and maintaining stable performance in complex real-world conversational scenarios. These limitations underline the necessity for more flexible and global attentive architectures that can robustly capture nuanced emotional trajectories in multi-turn dialogues.

2.2.2. Pretrained Language Models—Single Model

PLMs have contributed significantly to ERC by leveraging Transformer architectures and large-scale pretraining corpora to extract semantically rich, context-sensitive representations. Recent studies have explored how fine-tuned PLMs can be adapted for multi-turn emotion recognition tasks, particularly by addressing inter-utterance dependencies and speaker dynamics.

Kim and Vossen [16], Shen et al. [17], and Song et al. [18] each developed specialized ERC models—EmoBERTa, DialogXL, and CASA, respectively. These models take entire dialogue sequences as input, with utterances tokenized and augmented with speaker or syntactic information. EmoBERTa [16] adds speaker tags to each utterance and uses the self-attention of RoBERTa to encode a dialogue-wide semantic context, while DialogXL [17] enriches XLNet with role-level, temporal, and global attention mechanisms that explicitly model inter-utterance dependencies and speaker-specific emotional transitions. In contrast, CASA [18] focuses on dialogue-level sentiment analysis by injecting syntactic and sentiment-aware features, such as dependency-based aspect terms and emotion knowledge graphs, into BERT representations to enhance the disambiguation of subjective expressions across turns.

Despite their strong performances, these single-model PLMs tend to emphasize global discourse-level understanding, often at the cost of local emotional cues such as subtle sentiment shifts, intensifiers, or emotionally charged words. This imbalance may hinder the accurate recognition of nuanced or context-specific emotions in conversations, suggesting the need for hybrid models that can jointly model global semantic flows and fine-grained emotional features.

2.2.3. Pre-Trained Language Models—Hybrid Model

To overcome the limitations of single PLMs in ERC, recent studies have proposed hybrid architectures that integrate PLMs with traditional deep learning components such as CNNs, LSTMs, and graph neural networks (GNNs). These models combine the global semantic understanding of PLMs with the fine-grained, localized feature extraction capabilities of CNNs and RNNs, thus effectively capturing emotional cues in dialogue.

For instance, several models have enhanced BERT-based representations using CNNs or RNNs to detect fine-grained emotional expressions. Abas et al. [19] proposed the BERT-CNN, which applies convolutional filters to BERT embeddings. This better captures local sentiment signals, achieving improved performance on the ISEAR and SemEval datasets, particularly for subtle emotional cues. Kumar and Raman [20] advanced this concept by using a dual-channel BERT-LSTM-CNN architecture, combining sequential LSTM-based modeling with rule-based emotional features and attention-enhanced CNN layers. This approach demonstrates higher interpretability and improved classification accuracy for ISEAR, particularly in detecting nuanced emotional variations.

To further enrich the contextual understanding, several studies have incorporated additional semantic encoders. Basile et al. [21] combined BERT, LSTM, and the universal sentence encoder (USE) in a multistream model (BERT-LSTM-USE) to handle short chatbot conversations. Their model leverages sentence-level semantic representations and contextual flow to achieve a strong performance on the SemEval dataset. Wu et al. [22] introduced DialoguePCN, a multimodule perception-cognition network that integrates

BERT with hierarchical encoding layers and speaker-level attention mechanisms. Evaluated using MELD and IEMOCAP, the model effectively captures both intra- and inter-speaker emotional transitions. Similarly, Song et al. [23] proposed EmotionFlow, which couples BERT with bidirectional GRUs and a graph-based reasoning module to model emotion dynamics across multi-turn dialogues. This model, tested on the MELD, excels in tracking temporal emotional shifts and in capturing the flow of affective content.

Similarly, Alqarni et al. [24] introduced Emotion-Aware RoBERTa, an enhanced version of RoBERTa that incorporates an emotion-specific attention mechanism and a TF-IDF-based gating layer. This configuration strengthens the model's ability to prioritize emotionally salient tokens while filtering out irrelevant information, resulting in improved robustness with noisy and imbalanced datasets. Yan et al. [25] proposed Emotion-RGCNet, which combines RoBERTa with relational graph convolutional networks to capture contextual emotional dependencies in social media texts. This model surpassed standard transformer baselines in multi-class emotion classification by leveraging inter-token relational dynamics. Najafi et al. [26] developed TurkishBERTweet, a transformer-based model tailored for emotion recognition in Turkish social media. By integrating domain-adapted encoders with CNN classification layers, the model achieved strong performance on noisy, morphologically rich, and low-resource datasets.

Collectively, these hybrid models demonstrate that integrating PLMs with localized modeling techniques, such as CNNs, LSTMs, and GNNs, yields substantially improved performance in emotion recognition tasks. However, the proposed EmoBERTa-CNN differs markedly from earlier hybrid PLM-CNN models such as BERT-CNN [19] and BERT-LSTM-CNN [20]. First, it employs an EmoBERTa-based global contextual encoder pretrained on emotion-rich corpora, making it more effective at capturing global affective context. Second, the CNN module integrates an attention-pooling mechanism that dynamically highlights emotionally salient local features, unlike the fixed or max-pooling layers used in prior models. Third, EmoBERTa-CNN is specifically designed for multi-turn conversational emotion recognition and is evaluated on diverse dialogue-based datasets, whereas previous models primarily focus on single-turn utterances. Collectively, these innovations position EmoBERTa-CNN as a more robust and context-aware hybrid framework for dialogue emotion recognition.

2.3. Attention Mechanism for NLP

Recently, attention mechanisms have demonstrated remarkable efficacy in the field of NLP, particularly in tasks such as text question-answering systems and summarization [27].

Du et al. [28] designed a psychological assessment model that transforms sentence-level textual inputs into emotion-specific vectors, which are then processed using a CNN equipped with an attention module to emphasize emotionally relevant terms. Similarly, Liu et al. [29] proposed a multistage attention framework based on temporal convolutional networks (TCNs), wherein attention mechanisms are applied at both the word and concept levels to weight features dynamically based on contextual relevance. Their model showed notable gains on benchmark datasets, such as TREC and MR, particularly in handling subtle and short-text inputs.

Jia [30] presented a sentiment classification framework for Chinese microblogs that integrates BERT embeddings with a CNN architecture enhanced by an attention-pooling mechanism. The model's input comprises microblog texts that are first encoded using BERT to capture global semantics. The CNN component then extracts local n-gram features; the attention-pooling layer replaces the conventional max or average pooling by assigning learned and emotion-relevant weights to each convolutional feature. This design enables the

model to emphasize emotionally salient regions of text, resulting in marked improvements in the classification accuracy on three-class Chinese sentiment datasets.

To enhance the performance in dialogue emotion classification tasks, this study incorporated attention mechanisms by replacing the conventional pooling layers with an attention-pooling layer. Unlike traditional pooling operations that treat all features equally or select only the maximum value, attention-pooling computes adaptive weights based on the relevance of each feature. The attention-pooling mechanism can dynamically assign varying weights to these features on the basis of their relevance, thereby enhancing the CNN’s capability to selectively focus on emotionally salient information and improving overall emotion classification performance.

2.4. Summary

This subsection systematically reviews traditional methods, deep learning models, and PLMs, highlighting their strengths and drawbacks. Traditional supervised methods require extensive manual feature engineering. Deep learning methods struggle with long dialogue contexts. Single-model PLMs overlook localized emotional cues. To overcome these challenges, this study proposes and investigates EmoBERTa–CNN—a hybrid architecture that combines the global semantic modeling strengths of EmoBERTa with the local feature extraction capabilities of CNN—designed to integrate the respective advantages of prior approaches while overcoming their limitations in handling complex conversational emotion recognition tasks.

A review of attention mechanisms highlights their proven efficacy in enabling models to dynamically focus on salient information, thereby enhancing feature representation and model performance in various NLP tasks.

Table 1 presents a comparative summary of representative studies in the field of emotion recognition. It contrasts key aspects, such as dataset type, whether multi-turn dialogue is supported, the use of global and local features, and the incorporation of attention mechanisms. This tabular overview facilitates a clearer understanding of methodological trends and reveals how our proposed model aligns with, and advances beyond, existing approaches in handling complex conversational emotion recognition tasks.

Table 1. Comparison of prior studies on emotion recognition.

Paper	Dataset	Method	Input Tokens	Emotions	Multi-Turn	Global	Local	Attention
[10]	Six emotional categories	Naive Bayes, SVM	\	6	X	X	O	X
[11]	Six emotional categories	Naive Bayes, SMO	\	6	X	X	O	X
[12]	Six emotional categories	Bi-GRU+CNN	128	6	X	X	O	X
[13]	Three categories (emotional polarity)	LSTM+CNN	128	3	X	O	O	O
[14]	IEMOCAP	DialogueRNN	128	6	O	O	X	X
[15]	IEMOCAP, MELD	DialogueGCN	128	6, 7	O	O	O	X
[16]	IEMOCAP, MELD	RoBERTa-based	512	6, 7	O	O	X	X
[17]	IEMOCAP, MELD	XLNet-large	512	6, 7	O	O	X	X
[18]	Three categories (emotional polarity)	BERT-based	256	3	O	O	X	X

Table 1. Cont.

Paper	Dataset	Method	Input Tokens	Emotions	Multi-Turn	Global	Local	Attention
[19]	ISEAR, SemEval	BERT+CNN	512	6, 3	X	O	O	X
[20]	ISEAR	BERT+LSTM	128	6	X	O	O	O
[21]	SemEval	BERT+GRU	128	3	O	O	O	X
[22]	IEMOCAP, MELD	BERT+GCN	512	6, 7	O	O	O	X
[23]	MELD	BERT+GRU+GNN	256	7	O	O	O	X
[24]	Six emotional categories	RoBERTa+TF-IDF	512	6	X	O	O	O
[25]	Sentiment140, Emotion Dataset	RoBERTa + R-GCN+CRF	256	3, 7	X	O	O	X
[26]	Turkish social media	PTM+CNN	256	3	X	O	O	X
[28]	Text Review	CNN	128	5	X	O	X	O
[29]	Movie Review	TCN	64	2	X	O	O	O
[30]	Three categories (emotional polarity)	BERT+CNN	128	3	X	O	O	O
our	SemEval,MELD	EmoBERTa+CNN	128	3,7	O	O	O	O

Note: “O” indicates presence; “X” indicates absence; “\” denotes was not used; “bold” indicates the proposed model.

3. Methods

3.1. Overview

To address the challenge of ERC, wherein existing models struggle to simultaneously capture global contextual semantics and local emotional features, this study devised a novel hybrid architecture referred to as EmoBERTa–CNN. By combining the strengths of Transformer-based PLMs (EmoBERTa [16]) and CNNs, the proposed model aims to improve emotion classification performance by effectively modeling global semantic information and localized emotional cues.

As illustrated in Figure 1, the EmoBERTa–CNN model operates in two stages. In Stage 1, input multi-turn conversational texts are encoded into semantic vectors by the EmoBERTa-based global contextual encoder module, which applies Transformer-based layers for deep contextual modeling to generate global semantic representations. In Stage 2, these representations are passed to the CNN-based local emotion classifier, wherein the convolution and attention-pooling operations extract fine-grained local emotional features. A fully connected layer aggregates these features for precise emotion category prediction, enabling the model to effectively leverage both global and local information.

3.2. Bidirectional Encoder Representations from Transformers for Emotion Recognition (EmoBERTa)

The EmoBERTa-based global contextual encoder module is a variant of the PLM RoBERTa, optimized specifically for ERC. EmoBERTa retains the powerful capabilities of Transformer architectures to model global contextual semantics using textual sequences. Through further pretraining on an emotion-aware corpus, EmoBERTa becomes more effective at capturing global semantic features and emotion-related characteristics within conversational texts.

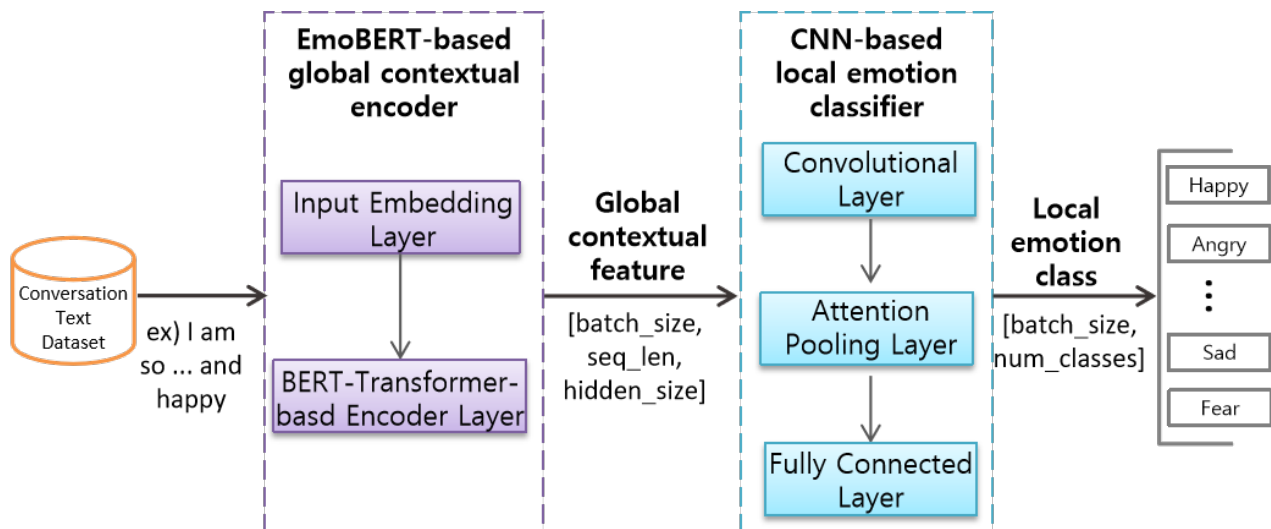


Figure 1. Overview of the EmoBERTa-CNN model.

As is evident in Figure 2, the EmoBERTa architecture follows a standard Transformer structure, in which the input is a raw text sequence and the output is represented as matrices. Assuming that an input textual sequence comprises n tokens, with each token represented by embeddings of dimension 1024 (used in this study), the EmoBERTa input and output are matrices with $n \times 1024$ dimensions. The EmoBERTa model comprises multiple stacked Transformer encoder layers. Specifically, the EmoBERTa-large variant employed in this study utilizes 24 Transformer encoder layers, enabling the model to deeply capture semantic and emotional contexts within conversational sequences.

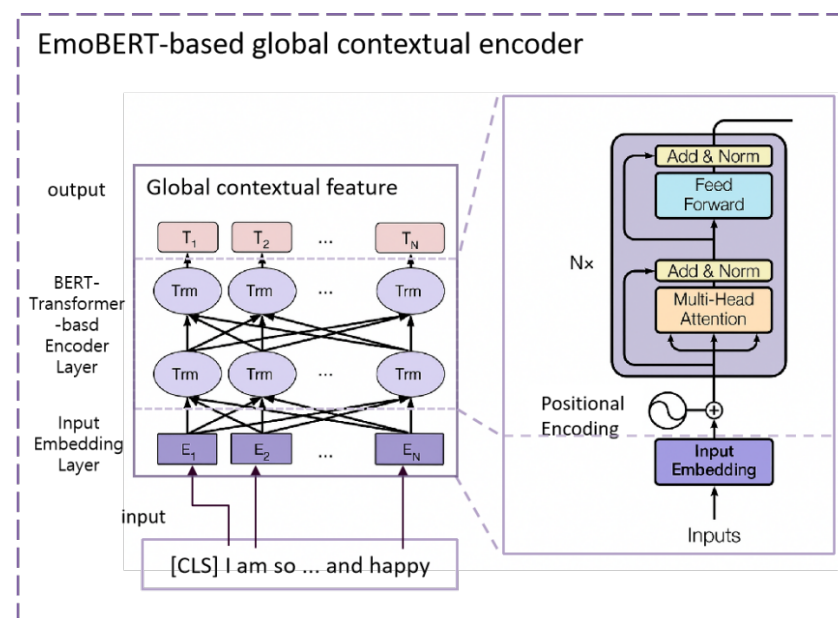


Figure 2. Architecture of the EmoBERTa-based global contextual encoder module.

3.2.1. Input Representation on EmoBERTa

In the EmoBERTa-based global contextual encoder module, the embedding layers comprise the following three components: token, segment, and position embeddings.

The token embedding layer transforms each word in the input text into a vector representation with fixed dimensions. Specifically, EmoBERTa represents each token as a 1024-dimensional vector. For the model to process text as vectors, two essential pre-

processing steps are required before the embedding operation. First, tokenization divides the input text into basic units (tokens)—such as words or subwords—that the model can recognize. Second, the token sequence is augmented with special structured markers. A [CLS] token is prepended to capture the global representation of the entire sequence and [SEP] tokens are inserted at the end of each text segment to denote segment boundaries or sequence termination. The resulting sequence, comprising the [CLS] marker, tokenized text, and one or more [SEP] markers, is then fed into the embedding layer for numerical encoding.

Segment embeddings are primarily designed to differentiate distinct segments within input sequences, particularly in text classification tasks involving the relationships between two sentences. In the context of ERC, EmoBERTa leverages segment embeddings to explicitly capture and represent emotional interactions among multiple conversational utterances, thereby facilitating clearer contextual distinctions.

Position embeddings encode positional information for each token within a sentence. Because the Transformer architecture itself does not inherently capture sequential-order information, these positional embeddings are essential for EmoBERTa to incorporate token ordering within sequences. To form the final input representation fed into EmoBERTa's encoder layers, these three types of embeddings—token, segment, and positional—are summed element-wise to yield comprehensive contextual embeddings suitable for deep semantic and emotional feature extraction.

3.2.2. Encoder Layer

The EmoBERTa encoder comprises 24 stacked Transformer blocks, each incorporating a multihead self-attention mechanism with 16 attention heads. Each encoder layer accepts input sequences with a maximum length of 512 tokens and produces comprehensive semantic representations, either as individual hidden state vectors or as sequences of hidden state vectors, at each time step.

The self-attention mechanism explicitly models dependencies and captures the structural relationships between the tokens within a sentence. Specifically, multihead self-attention is implemented via the scaled dot-product attention mechanism [31], computed as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where K , Q , and V denote the key, query, and value matrices, respectively, and d_k denotes the dimensionality of the query and key matrices. Multihead attention performs parallel computations by projecting queries, keys, and values through multiple linear transformations, resulting in multiple attention heads. The outputs from these attention heads are concatenated and further linearly projected to form the final attention output, as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{Head}_1, \dots, \text{Head}_h)W^O \quad (2)$$

$$\text{Head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (3)$$

where W_i^Q , W_i^K and W_i^V are learnable projection matrices for each Head_i , and W^O projects the concatenated outputs back to the original dimensionality. This design enables the model to jointly attend to information from different representation subspaces. Each encoder layer also incorporates a position-wise feedforward network (FFN) comprising two linear transformations activated by a rectified linear unit (ReLU) function, defined as follows:

$$\text{FFN}(X) = \max(0, XW_1 + b_1)W_2 + b_2 \quad (4)$$

The output from the FFN is input to the subsequent encoder layer. This process is repeated iteratively through all encoder layers, and the final output of the top encoder layer is transmitted to a classification layer, which ultimately performs the emotion classification. Because the self-attention mechanism inherently lacks positional awareness, EmoBERTa employs positional embeddings to encode the sequential order of the tokens. These embeddings are computed using sine and cosine functions, as follows:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (5)$$

$$PE(pos, 2i + 1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (6)$$

where pos is the position index of the token and d_{model} denotes the embedding dimension. This positional encoding helps the model to learn relative positional relationships efficiently, enhancing the effectiveness of the sequential semantic representation.

Consequently, the EmoBERTa encoder layers efficiently leverage the contextual information from both directions of the tokens in the input sequence, yielding richer and more robust feature representations.

3.3. Convolutional Neural Network (CNN) Module

The proposed CNN-based local emotion classifier takes the global semantic embeddings generated by the EmoBERTa-based contextual encoding module as input and extracts critical local emotional features through a sequence of convolutional, attention-pooling, dropout, and fully connected layers, as illustrated in Figure 3.

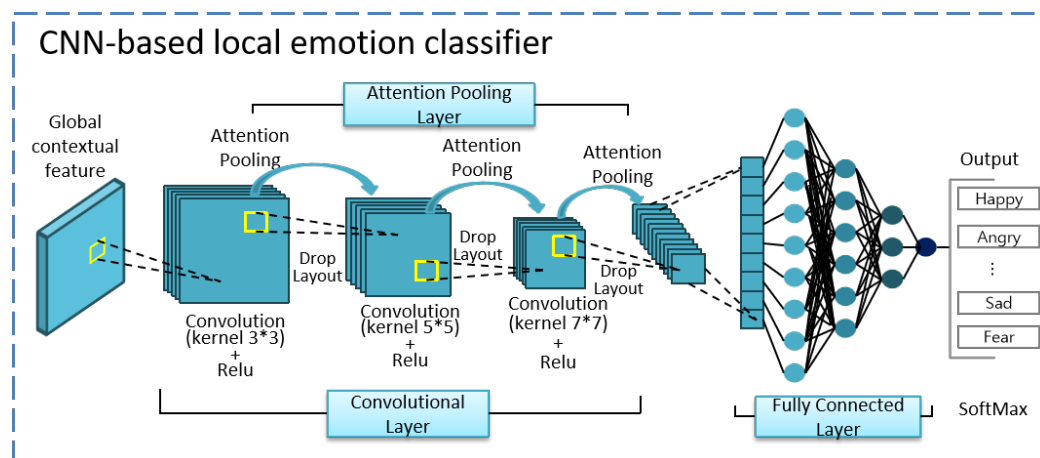


Figure 3. Architecture of the CNN-based local emotion classifier.

The convolutional layer applies filters of varying sizes to capture the local semantic features from the input matrix (EmoBERTa embeddings). Given an input embedding vector, the convolutional operation employs multiple filters, each with a fixed width (equal to the embedding dimension, 1024) and varying heights to extract diverse n-gram features [32]. Mathematically, the convolutional feature extraction can be represented as follows:

$$S_j = g(w \cdot [v_i; v_{i+1}; \dots; v_{i+h-1}] + b) \quad (7)$$

where w denotes a convolutional filter of size $h \times m$, $[v_i; v_{i+1}; \dots; v_{i+h-1}]$ represents a sliding window over consecutive tokens, b is the bias term. g represents the activation function used in this study, namely, the ReLU.

An attention-pooling mechanism is introduced after the convolutional layer. A formal description of the attention-pooling operation is given in Equations (8)–(10). First, each convolutional feature vector c_i is transformed via nonlinear activation, where W and b represent the learnable weight matrix and bias of the attention layer (Equation (8)). Next, the attention coefficient a_i is computed by applying the softmax function to the dot products between each e_i and a randomly initialized context vector E (Equation (9)). Finally, these coefficients weight the original convolutional features to produce the pooled output (Equation (10)):

$$e_i = \tanh(Wc_i + b) \quad (8)$$

$$a_i = \frac{\exp(e_i^\top E)}{\sum_{j=1}^n \exp(e_j^\top E)} \quad (9)$$

$$c' = \sum_{i=1}^n a_i c_i \quad (10)$$

Here, c' denotes the attention-pooled feature vector, which enables the model to focus selectively on emotion-related cues extracted by the CNN.

After the pooling step, dropout regularization mitigates potential overfitting. Specifically, this study adopted a dropout rate of 0.5, randomly disabling half of the neuron connections during training to enhance model generalization.

To ensure optimal model performance, the hyperparameters of the CNN module—such as the number of filters, kernel sizes, and layer depth—were empirically tuned through iterative experimentation. Initially, each parameter was constrained to a reasonable range informed by prior empirical studies. Within this range, various configurations were manually tested; those that consistently achieved stable convergence and higher F1-scores on the validation set were selected. The final architecture includes three convolutional layers, each with 256 filters and kernel sizes of [3, 5, 7]. This design enables the model to effectively extract local emotional patterns across multiple n-gram levels, thereby improving its ability to capture fine-grained emotional cues from conversational text.

Finally, the fully connected layer aggregates the local emotional features extracted from the preceding layers, transforming them into a fixed-length vector suitable for the final emotion classification. A softmax activation function is utilized at this stage to predict the emotion labels by calculating the probabilities for each class. The softmax function is formally expressed as follows:

$$p_i = \frac{\exp(e_i)}{\sum_{j=1}^C \exp(e_j)} \quad (11)$$

where p_i denotes the probability of the i -th class, e_i is the output score for class i , and C is the total number of classes. The softmax function transforms raw output scores into a normalized probability distribution over all classes, ensuring that the probabilities are non-negative and sum to one.

Although recurrent models, such as Bi-GRU and LSTM, are commonly used to capture sequential dependencies in dialogue, CNNs were deliberately selected in this study for local feature extraction due to their efficiency in capturing n-gram patterns and lower computational complexity. Unlike RNN-based architectures, CNNs support parallel computation and are less prone to vanishing gradient issues. In emotion recognition tasks, where short phrases and affective keywords often convey key emotional signals, CNNs effectively extract such fine-grained patterns through convolutional operations. Moreover, since the EmoBERTa encoder in this architecture already models long-range sequential context, CNNs function as a lightweight yet effective component for enhancing local emotional representations. These advantages collectively make CNNs a suitable and computationally efficient choice for modeling localized features in multi-turn dialogue scenarios.

4. Results

4.1. Hyperparameters

The performance of the proposed emotion detection system is sensitive to several key hyperparameters, including batch size, hidden layer size, number of layers, learning rate, and dropout rate. These parameters must be systematically tuned to optimize classification accuracy. The model was trained using the Adam optimizer combined with a categorical cross-entropy loss function. Using categorical cross-entropy enhances the model's ability to learn the probability distribution over distinct emotion classes.

To determine suitable hyperparameter values for the EmoBERTa–CNN model, we employed a two-stage approach. First, we defined a reasonable search space for each hyperparameter, guided by prior studies in the literature and empirical insights from preliminary experiments. Within these ranges, we manually tuned key parameters by iteratively adjusting the CNN learning rate, kernel size, number of filters, batch size, and dropout rate. Each configuration was assessed based on its performance on the validation set, with particular attention to F1-score stability across multiple runs. The final hyperparameter set was selected based on consistent convergence behavior and improved F1-scores observed across repeated trials. The selected values are summarized in Table 2.

Table 2. Hyper-parameters for EmoBERTa–CNN.

Hyper-Parameters	Values
EmoBERTa learning rate	1×10^{-6}
CNN learning rate	1×10^{-5}
Optimizer	Adam
Loss function	Categorical cross-entropy
Batch size	16
Dropout	0.5
Kernel sizes	[3, 5, 7]
Epochs	10

4.2. Dataset

The proposed model was evaluated using two datasets: SemEval 2019 Task 3 [33] and MELD [34]. The SemEval 2019 dataset includes three classes of emotions: “angry”, “happy”, and “sad”. The training and testing samples for each emotion category and the overall number of instances are listed in Table 3. The MELD dataset is a multiparty (more than two interlocutors in a dialogue) conversational dataset. The data were collected from the TV series *Friends*. The MELD dataset contains seven classes of emotions: neutral, joy, surprise, anger, sadness, disgust, and fear. The training and testing samples for each emotion category and the overall number of instances are listed in Table 4.

Table 3. Description of the SemEval dataset.

Dataset Split	Total	Happy	Sad	Angry
Training	15212	4243	5463	5506
Testing	832	284	250	298

Table 4. Description of the MELD dataset.

Dataset Split	Total	Neutral	Joy	Surprise	Anger	Sadness	Disgust	Fear
Training	9989	4710	1743	1205	1109	683	271	268
Testing	2610	1256	402	345	281	208	68	50

4.3. Evaluation Metrics

The standard metrics, namely, recall, precision, and F1-score, were used for the evaluation. Here, the positive sentences correctly classified into the emotion class are True Positive (TP). The negative sentences classified as positive in the emotion class are designated as False Positive (FP). Sentences genuinely belonging to a specific emotion class and erroneously classified as not belonging to that class are designated as False Negative (FN). The precision, recall, and F-measure were computed as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (12)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (13)$$

$$\text{F-measure} = \frac{2 \times P \times R}{P + R} \quad (14)$$

Additionally, considering the potential imbalance in the distribution of emotion categories within the dataset, this study employed a weighted average F1-score [35] to objectively evaluate the performance of the model across all classes, as follows:

$$\text{WeightedF1} = \frac{\sum_i (F1_i \times N_i)}{\sum_i N_i} \quad (15)$$

where N_i denotes the number of instances of the i th emotion class in the dataset. This technique effectively reduces the impact of class imbalance, leading to a more robust evaluation of the model's overall classification capabilities.

4.4. Analysis of Results

To assess the proposed EmoBERTa-CNN model's performance on ERC tasks, extensive experiments were conducted utilizing two publicly available datasets: SemEval 2019 Task 3 and MELD. We conducted comparative analyses between our proposed approach and SOTA models to confirm the effectiveness of our method.

For the SemEval 2019 Task 3 dataset, the experiments were conducted on the following three emotional classes: "angry", "happy", and "sad". We compared the proposed EmoBERTa-CNN model with recent models, namely, SymantoResearch [21], BERT-CNN [19], and EmoBERTa [16]. As shown in Table 5, our EmoBERTa-CNN-Att model achieved the highest overall F1-score of 96.0%, significantly outperforming all the other benchmark methods. The version without attention (i.e., EmoBERTa-CNN) achieved a competitive F1-score of 95.6%.

Table 5. Comparison of the performances of our and several recent models using the SemEval dataset.

Model	F1-Score
SymantoResearch [21]	76.8%
BERT-CNN [19]	94.3%
EmoBERTa [16]	93.3%
EmoBERTa-CNN (our)	95.6%
EmoBERTa-CNN-Att (our)	96.0%

Note: The F1-score in bold indicates the best performance among all compared models in this table.

To further analyze performance, we computed the detailed precision, recall, and F1-scores for the individual emotion classes, as presented in Table 6. The EmoBERTa-CNN-Att model demonstrated strong performance across the three emotion categories, achieving F1-scores of 0.96 for "angry", 0.95 for "sad", and 0.97 for "happy". The EmoBERTa-CNN

model (without the attention mechanism) also yielded competitive results, with F1-scores of 0.95 for “angry”, 0.95 for “sad”, and 0.97 for “happy”.

Table 6. Individual emotions scores using the SemEval test dataset.

Emotions Model	Angry			Sad			Happy		
	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑
SymantoResearch [21]	0.73	0.79	0.76	0.82	0.80	0.81	0.75	0.72	0.73
BERT-CNN [19]	0.91	0.97	0.94	0.95	0.90	0.92	0.98	0.96	0.97
EmoBERTa [16]	0.91	0.95	0.93	0.96	0.88	0.92	0.94	0.96	0.95
EmoBERTa-CNN (our)	0.95	0.95	0.95	0.94	0.96	0.95	0.98	0.96	0.97
EmoBERTa-CNN-Att (our)	0.95	0.96	0.96	0.95	0.95	0.95	0.98	0.96	0.97

Note: Each emotion is evaluated using Precision (P), Recall (R), and F1-score (F1), where ↑ indicates that higher values reflect better performance. Bolded F1-scores denote the best results for each emotion class across all models.

These results indicate that both proposed models generally outperformed other benchmark models, including SymantoResearch [21] and EmoBERTa [16], regarding the F1-scores for individual emotions. For instance, for the “angry” emotion, EmoBERTa-CNN-Att (0.96) and EmoBERTa-CNN (0.95) surpassed BERT-CNN [19] (0.94), EmoBERTa [16] (0.93), and Symanto Research [21] (0.76). A similar trend was observed for the “sad” emotion. With regards to the “happy” emotion, both our models achieved an F1-score of 0.97, which is on par with BERT-CNN [19] and higher than the other compared methods.

Figure 4 presents the confusion matrices for the EmoBERTa, EmoBERTa-CNN, and EmoBERTa-CNN-Att models on the SemEval dataset, offering a detailed visualization of their classification performances for each emotion category. As shown in Figure 4a, the baseline EmoBERTa model correctly classified 777 labels (284 for angry, 220 for sad, and 273 for happy). As depicted in Figure 4b, the proposed EmoBERTa-CNN model demonstrated an improved performance by correctly identifying 795 labels (282 “angry”, 239 “sad”, and 274 “happy”). Furthermore, the EmoBERTa-CNN-Att model achieved the highest number of correct classifications, accurately predicting 798 labels (287 “angry”, 237 “sad”, and 274 “happy”) (Figure 4c).

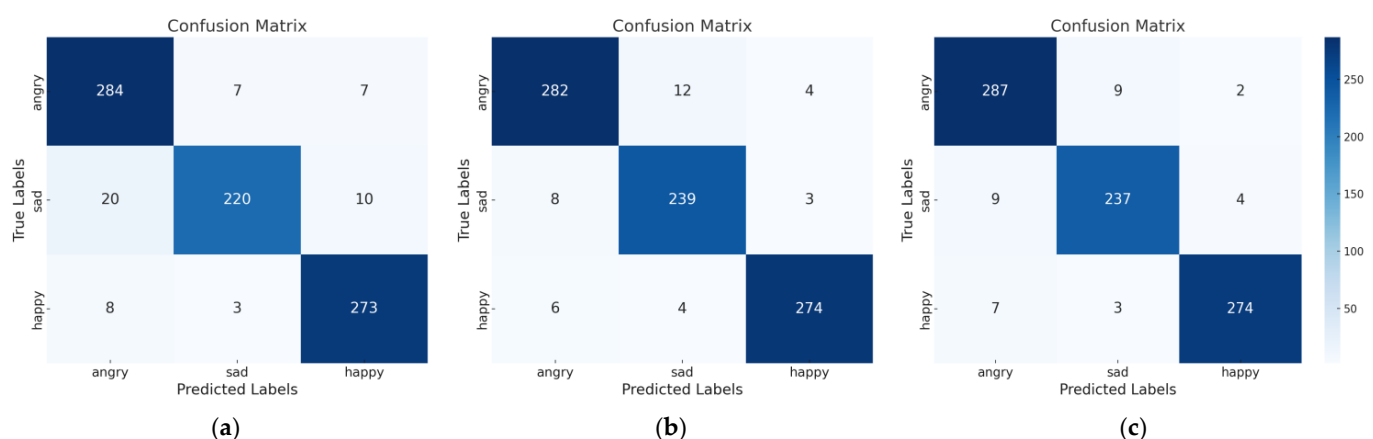


Figure 4. (a) Confusion matrix of the EmoBERTa model on the SemEval dataset; (b) confusion matrix of the EmoBERTa-CNN model on the SemEval dataset; and (c) confusion matrix of the EmoBERTa-CNN-Att model on the SemEval dataset.

This progression from EmoBERTa to EmoBERTa-CNN, and subsequently to EmoBERTa-CNN-Att, clearly underscores the enhanced predictive accuracy and stability of the proposed architectures. Furthermore, the attention mechanism in EmoBERTa-CNN-Att con-

tributes to more accurate label predictions and a reduction in misclassification across emotion categories.

Experiments involving seven emotional categories (“neutral”, “joy”, “surprise”, “anger”, “sadness”, “disgust”, and “fear”) were conducted using the MELD dataset. We comprehensively compared the proposed EmoBERTa–CNN and EmoBERTa–CNN–Att models with recent advanced methods such as DialogXL [17], EmotionFlow [23], and EmoBERTa [16].

As evident from the results summarized in Table 7, EmoBERTa–CNN (without attention) achieved a weighted average F1-score of 76.39%. Critically, integrating the attention mechanism into EmoBERTa–CNN–Att further boosted this performance to a weighted average F1-score of 79.45%. This represents a significant increase, of 3.06 percentage points, over the EmoBERTa–CNN baseline, emphasizing the effectiveness of attention mechanisms in enhancing the model’s emotional classification capabilities. Furthermore, this score of 79.45% for EmoBERTa–CNN–Att was also the highest achieved among all the compared models, demonstrating clear improvements over other existing advanced methods.

Table 7. Comparing performances of the proposed models and several recent methods using the MELD dataset.

Model	F1-Score
DialogXL [17]	62.41%
Emotionflow [23]	65.05%
EmoBERTa [16]	66.59%
EmoBERTa–CNN (our)	76.39%
EmoBERTa–CNN–Att (our)	79.45%

Note: The F1-score in bold indicates the best performance among all compared models in this table.

The model performance was further analyzed by computing the detailed precision, recall, and F1-scores for the individual emotion classes (Table 7). These results demonstrate the impact of the architectural enhancements. First, comparing EmoBERTa–CNN to the baseline EmoBERTa [16] showed consistent improvements in the F1-scores across all seven emotion categories (Neutral: 0.86 vs. 0.81; Joy: 0.74 vs. 0.63; Surprise: 0.72 vs. 0.58; Anger: 0.69 vs. 0.57; Sadness: 0.63 vs. 0.36; Disgust: 0.43 vs. 0.34; Fear: 0.31 vs. 0.20). This indicates that the integrated CNN and EmoBERTa hybrid architecture outperformed the single-model architectures in emotion classification tasks.

More importantly, introducing the attention mechanism into the EmoBERTa–CNN–Att model yielded further significant gains. As shown in Table 8, EmoBERTa–CNN–Att consistently outperformed EmoBERTa–CNN (without attention) across individual emotion F1-scores, with the exception of surprise: Neutral (0.88 vs. 0.86); Joy (0.78 vs. 0.74); Surprise (0.71 vs. 0.72); Anger (0.73 vs. 0.69); Sadness (0.67 vs. 0.63); Disgust (0.59 vs. 0.43); and Fear (0.48 vs. 0.31). These improvements underscore the effectiveness of the attention mechanism in enabling the model to better focus on salient emotional cues within the contextualized representations of each emotion. Consequently, the EmoBERTa–CNN–Att model surpassed its nonattention counterpart and significantly outperformed the original EmoBERTa baseline across all individual emotion classes.

Table 8. Individual emotions scores using the MELD test dataset.

Emotion Model	Neutral			Joy			Surprise			Anger			Sadness			Disgust			Fear		
	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑
EmoBERTa [16]	0.75	0.88	0.81	0.66	0.61	0.63	0.55	0.61	0.58	0.59	0.55	0.57	0.73	0.24	0.36	0.62	0.24	0.34	0.28	0.16	0.20
EmoBERTa–CNN (our)	0.88	0.85	0.86	0.71	0.78	0.74	0.66	0.78	0.72	0.65	0.73	0.69	0.66	0.59	0.63	0.70	0.31	0.43	0.55	0.22	0.31
EmoBERTa–CNN–Att (our)	0.88	0.88	0.88	0.78	0.78	0.78	0.64	0.80	0.71	0.75	0.72	0.73	0.78	0.59	0.67	0.63	0.56	0.59	0.54	0.44	0.48

Note: Each emotion is evaluated using Precision (P), Recall (R), and F1-score (F1), where ↑ indicates that higher values correspond to better performance. Bolded F1-scores indicate the best results for each emotion class across all models.

As Figure 5 shows, the confusion matrix analysis provides a visual representation of the classification performance and illustrates the progressive improvements achieved by our enhancements.

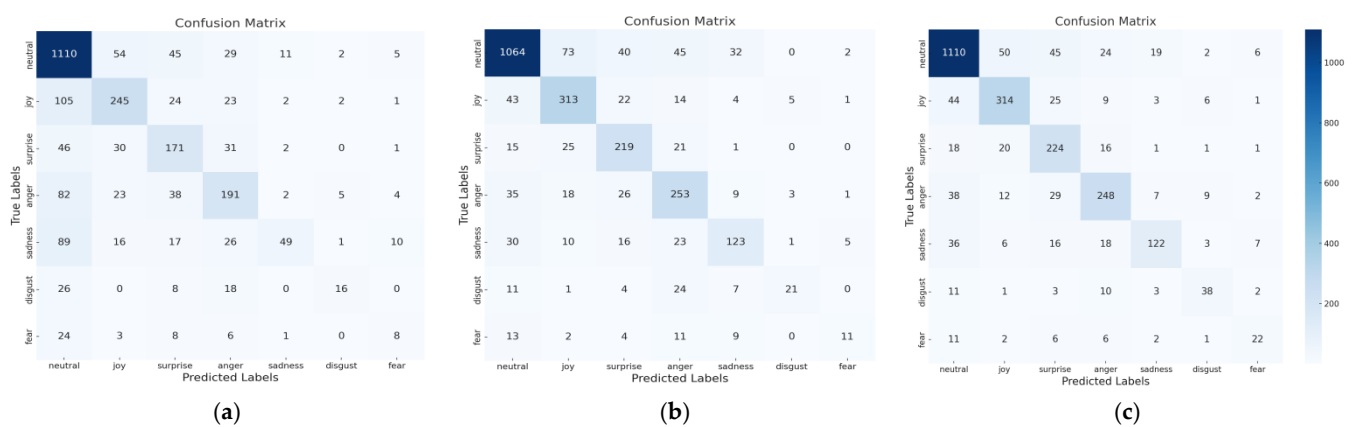


Figure 5. (a) Confusion matrix of the EmoBERTa model on the MELD dataset; (b) confusion matrix of the EmoBERTa–CNN model on the MELD dataset; and (c) confusion matrix of the EmoBERTa–CNN–Att model on the MELD dataset.

Figure 5a depicts the confusion matrix for the baseline EmoBERTa model. Although this establishes its foundational performance, visible off-diagonal values indicate areas of confusion between certain emotion classes. The results reveal that this baseline model correctly predicted 1790 labels.

Figure 5b shows the confusion matrix for the EmoBERTa–CNN model. Integrating the CNN architecture significantly reduced confusion among emotion classes compared to the EmoBERTa baseline. EmoBERTa–CNN increased the total number of correctly predicted labels from 1790 (EmoBERTa) to 2004. This improvement is evident by the values concentrated more along the diagonal and the reduced magnitudes in the off-diagonal cells, clearly demonstrating its strengthened capability to discriminate fine-grained emotions.

Finally, Figure 5c shows the confusion matrix for the proposed EmoBERTa–CNN–Att model. Introducing the attention mechanism yielded discernibly enhanced classification accuracy. The EmoBERTa–CNN–Att model increased the total number of correctly predicted labels by 2078. This is reflected in the even stronger diagonal dominance and further minimized misclassifications across the emotion categories, as compared to EmoBERTa–CNN.

Despite these improvements, a qualitative error analysis based on the confusion matrices (Figures 4 and 5) and detailed performance metrics (Tables 6 and 8) revealed several limitations. In the SemEval dataset, the most frequent misclassifications occurred between the sad and angry categories. Specifically, instances labelled as sad were occasionally misclassified as angry, and vice versa, likely due to overlapping lexical and syntactic features. Similarly, the MELD dataset showed consistent confusion between emotionally

similar categories, such as sadness and fear, as well as surprise and joy. These categories share subtle contextual cues that present considerable challenges for accurate classification. Additionally, utterances containing implicit emotional undertones—such as sarcasm or irony, often embedded in conversational humor—were sometimes misclassified as neutral or incorrectly interpreted as anger. Furthermore, rare classes, such as fear and disgust, exhibited relatively higher misclassification rates, highlighting the model’s sensitivity to class imbalance in the MELD dataset.

5. Conclusions

This study investigated ERC and addressed the limitations of existing approaches in simultaneously capturing global contextual semantics and fine-grained local emotion features. To tackle these challenges, we devised EmoBERTa–CNN, a hybrid model that integrates the strengths of Transformer-based PLMs and CNNs.

This architecture comprises a pretrained EmoBERTa model combined with a CNN. The EmoBERTa model is based on RoBERTa and is fine-tuned using emotion-aware corpora. This enables EmoBERTa to capture global semantic contexts and emotional cues embedded in conversational dialogues. A CNN module was subsequently incorporated to extract local emotional features from the global semantic representations produced by EmoBERTa, thereby enabling precise emotion classification. To further enhance the capability of the CNN to capture emotionally salient information, an attention-pooling mechanism was integrated following the CNN module. In contrast to traditional pooling methods that uniformly treat all features or select only the strongest activation, attention pooling dynamically assigns weights to extracted local features based on their relevance to the emotion classification task. This selective aggregation enables a more precise focus on critical emotional cues, resulting in the improved overall discriminative power of the EmoBERTa–CNN architecture.

Extensive experiments were conducted using two benchmark datasets, namely, SemEval 2019 Task 3 and MELD. The results demonstrate that the EmoBERTa–CNN model consistently outperforms several state-of-the-art baselines in terms of the F1-score across various emotional categories. Confusion matrix analyses confirmed that EmoBERTa–CNN significantly reduced the misclassification between different emotional categories, thus achieving superior predictive accuracy.

In summary, the proposed model effectively resolved the challenges of simultaneously modeling the global context and local emotion features, substantially improving emotion recognition performance in conversational settings. Furthermore, this study contributes to the development of human–computer interaction by enabling dialogue systems to more accurately perceive and effectively respond to users’ emotional states. By enhancing the emotional intelligence of conversational agents, the proposed model supports the development of more empathetic, context-aware, and human-like interaction systems, which are essential for applications such as virtual assistants, mental health support platforms, and social robots.

Despite the encouraging experimental results achieved in this study, several limitations persist. First, the model depends heavily on manually annotated datasets and extensive training corpora, both of which are resource-intensive to create and maintain. This reliance poses challenges for applications in low-resource settings, where such data may be unavailable. Second, although the model performs well on standard benchmarks, its robustness in real-world, noisy, and cross-lingual environments remains underexplored. In future work, we intend to investigate semi-supervised and self-supervised learning frameworks to reduce dependence on annotated data. We also plan to explore model compression techniques and lightweight architectural variants to support deployment on

resource-constrained platforms. Finally, we aim to extend the framework to multilingual emotion recognition and evaluate its generalizability across a wider range of conversational domains and user contexts.

Author Contributions: Conceptualization, M.Z. and K.C.; Methodology, M.Z. and A.Y.; Software, M.Z.; Validation, M.Z. and X.S.; Formal analysis, M.Z.; Writing—original draft, M.Z.; Writing—review & editing, A.Y., J.P. and J.R.; Visualization, M.Z.; Supervision, J.R. and K.C.; Project administration, K.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant for the R and D project, funded by the National Center for Mental Health (grant number: MHER23B01).

Data Availability Statement: The datasets used in this study are publicly available. The Multimodal EmotionLines dataset (MELD) can be accessed at <https://affective-meld.github.io> (accessed on 19 March 2025), and the SemEval-2019 Task 3 dataset is available at <https://github.com/chenyangh/SemEval2019Task3/tree/master> (accessed on 19 March 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MELD	Multimodal EmotionLines Dataset
NLP	Natural language processing
ERC	Emotion recognition in conversation
TER	Textual Emotion Recognition
CNNs	Convolutional Neural Networks
RNNs	Recurrent Neural Networks
PLMs	Pre-trained language models
SVM	Support Vector Machines
GCNs	Graph Convolutional Networks
GRUs	Gated recurrent units
LSTMs	Long Short-Term Memory networks
GNNs	Graph Neural Networks
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Approach
EmoBERTa	Emotion-aware Bidirectional Encoder Representations from Transformers
SOTA	State-of-the-art
PTM	Pretrained Transformer Model

References

1. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [\[CrossRef\]](#)
2. Alva, M.Y.; Nachamai, M.; Paulose, J. A Comprehensive Survey on Features and Methods for Speech Emotion Detection. In Proceedings of the IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Tamil Nadu, India, 5–7 March 2015; pp. 1–6. [\[CrossRef\]](#)
3. Saxena, A.; Khanna, A.; Gupta, D. Emotion Recognition and Detection Methods: A Comprehensive Survey. *J. Artif. Intell. Syst.* **2020**, *2*, 53–79. [\[CrossRef\]](#)
4. Pereira, P.; Moniz, H.; Carvalho, J.P. Deep Emotion Recognition in Textual Conversations: A Survey. *Artif. Intell. Rev.* **2024**, *58*, 10. [\[CrossRef\]](#)
5. Zhang, T.; Zheng, W.; Cui, Z.; Zong, Y.; Li, Y. Spatial–Temporal Recurrent Neural Network for Emotion Recognition. *IEEE Trans. Cybern.* **2018**, *49*, 839–847. [\[CrossRef\]](#)
6. Zhao, J.; Gui, X.; Zhang, X. Deep Convolution Neural Networks for Twitter Sentiment Analysis. *IEEE Access* **2018**, *6*, 23253–23260. [\[CrossRef\]](#)

7. Munikar, M.; Shakya, S.; Shrestha, A. Fine-Grained Sentiment Classification Using BERT. In Proceedings of the 2019 Artificial Intelligence for Transforming Business and Society (AITB), Kathmandu, Nepal, 5 November 2019; IEEE: New York, NY, USA; pp. 1–5. [\[CrossRef\]](#)
8. Liu, Y. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692. [\[CrossRef\]](#)
9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2 June 2019; pp. 4171–4186. [\[CrossRef\]](#)
10. Asghar, M.Z.; Subhan, F.; Imran, M.; Kundi, F.M.; Khan, A.; Shamshirband, S.; Mosavi, A.; Koczy, A.R.V.; Csiba, P. Performance Evaluation of Supervised Machine Learning Techniques for Efficient Detection of Emotions from Online Content. *arXiv* **2019**, arXiv:1908.01587. [\[CrossRef\]](#)
11. Gaiind, B.; Syal, V.; Padgalwar, S. Emotion Detection and Analysis on Social Media. *arXiv* **2019**, arXiv:1901.08458. [\[CrossRef\]](#)
12. Bharti, S.K.; Varadhaganapathy, S.; Gupta, R.K.; Shukla, P.K.; Bouye, M.; Hingaa, S.K.; Mahmoud, A.; Kumar, V. Text-Based Emotion Recognition Using Deep Learning Approach. *Comput. Intell. Neurosci.* **2022**, *2022*, 2645381. [\[CrossRef\]](#)
13. Jiang, M.; Zhang, W.; Zhang, M.; Wu, J.; Wen, T. An LSTM-CNN Attention Approach for Aspect-Level Sentiment Classification. *J. Comput. Methods Sci. Eng.* **2019**, *19*, 859–868. [\[CrossRef\]](#)
14. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 6818–6825. [\[CrossRef\]](#)
15. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. *arXiv* **2019**, arXiv:1908.11540. [\[CrossRef\]](#)
16. Kim, T.; Vossen, P. EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa. *arXiv* **2019**, arXiv:2108.12009. [\[CrossRef\]](#)
17. Shen, W.; Chen, J.; Quan, X.; Xie, Z. DialogXL: All-in-One XLNet for Multi-party Conversation Emotion Recognition. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 13789–13797. [\[CrossRef\]](#)
18. Song, L.; Xin, C.; Lai, S.; Wang, A.; Su, J.; Xu, K. CASA: Conversational Aspect Sentiment Analysis for Dialogue Understanding. *J. Artif. Intell. Res.* **2022**, *73*, 511–533. [\[CrossRef\]](#)
19. Abas, A.R.; Elhenawy, I.; Zidan, M.; Othman, M. BERT-CNN: A Deep Learning Model for Detecting Emotions from Text. *Comput. Mater. Contin.* **2022**, *71*, 2943–2961. [\[CrossRef\]](#)
20. Kumar, P.; Raman, B. A BERT-Based Dual-Channel Explainable Text Emotion Recognition System. *Neural Netw.* **2022**, *150*, 392–407. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Basile, A.; Franco-Salvador, M.; Pawar, N.; Štajner, S.; Rios, M.C.; Benajiba, Y. SymantoResearch at SemEval-2019 Task 3: Combined Neural Models for Emotion Classification in Human-Chatbot Conversations. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 330–334. [\[CrossRef\]](#)
22. Wu, X.; Feng, C.; Xu, M.; Zheng, T.F.; Hamdulla, A. DialoguePCN: Perception and Cognition Network for Emotion Recognition in Conversations. *IEEE Access* **2023**, *11*, 141251–141260. [\[CrossRef\]](#)
23. Song, X.; Zang, L.; Zhang, R.; Hu, S.; Huang, L. EmotionFlow: Capture the Dialogue Level Emotion Transitions. In Proceedings of the ICASSP 2022—IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 22–27 May 2022; pp. 8542–8546. [\[CrossRef\]](#)
24. Alqarni, F.; Sagheer, A.; Alabbad, A.; Hamdoun, H. Emotion-Aware RoBERTa Enhanced with Emotion-Specific Attention and TF-IDF Gating for Fine-Grained Emotion Recognition. *Sci. Rep.* **2025**, *15*, 17617. [\[CrossRef\]](#)
25. Yan, J.; Pu, P.; Jiang, L. Emotion-RGC Net: A Novel Approach for Emotion Recognition in Social Media Using RoBERTa and Graph Neural Networks. *PLoS ONE* **2025**, *20*, e0318524. [\[CrossRef\]](#)
26. Najafi, A.; Varol, O. TurkishBERTweet: Fast and Reliable Large Language Model for Social Media Analysis. *Expert Syst. Appl.* **2024**, *255*, 124737. [\[CrossRef\]](#)
27. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [\[CrossRef\]](#)
28. Du, T.B.; Yu, C.H.; Wen, Z.; Kong, X. Psychological Assessment Model Based on Text Emotional Characteristics. *J. Jilin Univ.* **2019**, *57*, 927–932.
29. Liu, Y.; Li, P.; Hu, X. Combining Context-Relevant Features with Multi-Stage Attention Network for Short Text Classification. *Comput. Speech Lang.* **2022**, *71*, 101268. [\[CrossRef\]](#)
30. Jia, K. Sentiment Classification of Microblog: A Framework Based on BERT and CNN with Attention Mechanism. *Comput. Electr. Eng.* **2022**, *101*, 108032. [\[CrossRef\]](#)
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 2. [\[CrossRef\]](#)
32. Wang, J.; Wang, Z.; Zhang, D.; Yan, J. Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification. In Proceedings of the IJCAI, Melbourne, Australia, 19–25 August 2017; Volume 350, pp. 3172077–3172295. [\[CrossRef\]](#)

33. Huang, C.; Trabelsi, A.; Zaïane, O. ANA at SemEval-2019 Task 3: Contextual Emotion Detection in Conversations through Hierarchical LSTMs and BERT. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 49–53. [\[CrossRef\]](#)
34. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A Multimodal Multi-party Dataset for Emotion Recognition in Conversations. *arXiv* **2018**, arXiv:1810.02508. [\[CrossRef\]](#)
35. Hinojosa Lee, M.C.; Braet, J.; Springael, J. Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-scores. *Appl. Sci.* **2024**, *14*, 9863. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.