electronics

MDPI

Review

Emotion Recognition in Conversations: A Survey Focusing on Context, Speaker Dependencies, and Fusion Methods

Yao Fu, Shaoyang Yuan, Chi Zhang * and Juan Cao

State Key Laboratory of Media Convergence and Communication, Communication University of China, Dingfuzhuang, Beijing 100024, China; fuyao@cuc.edu.cn (Y.F.); 202120081200019@cuc.edu.cn (S.Y.); caojuan@cuc.edu.cn (J.C.)

* Correspondence: zhangchi@cuc.edu.cn; Tel.: +86-186-1198-3683

Abstract: As a branch of sentiment analysis tasks, emotion recognition in conversation (ERC) aims to explore the hidden emotions of a speaker by analyzing the sentiments in utterance. In addition, emotion recognition in multimodal data from conversation includes the text of the utterance and its corresponding acoustic and visual data. By integrating features from various modalities, the emotion of utterance can be more accurately predicted. ERC research faces challenges in context construction, speaker dependency design, and multimodal heterogeneous feature fusion. Therefore, this review starts by defining the ERC task, developing the research work, and introducing the utilized datasets in detail. Simultaneously, we analyzed context modeling in conversations, speaker dependency, and methods for fusing multimodal information concerning existing research work for evaluation purposes. Finally, this review also explores the research, application challenges, and opportunities of ERC.

Keywords: emotion recognition in conversation; speaker dependency; context construct; fusion method; feature extraction; multimodal data

Citation: Fu, Y.; Yuan, S.; Zhang, C.; Cao, J. Emotion Recognition in Conversations: A Survey Focusing on Context, Speaker Dependencies, and Fusion Methods. *Electronics* **2023**, *12*, 4714. https://doi.org/10.3390/electronics12224714

Academic Editor: Alberto Fernandez Hilario

Received: 25 October 2023
Revised: 14 November 2023
Accepted: 16 November 2023
Published: 20 November 2023

1. Introduction

Sentiment analysis is the study of human attitudes and feelings in specific situations, focusing on understanding the emotions expressed by humans through the analysis of new aspects of human speech, voice, facial expressions, and behavior. Sentiment analysis typically identifies positive, negative, and neutral emotions. In contrast, emotion recognition aims to discern individuals' more nuanced emotions, such as hatred, joy, and disgust. The transmission of human emotions usually involves multiple sensory channels, such as hearing, vision, touch, and taste. Researchers utilize diverse modalities to convey emotional information, including text, images, audio, EEG(Electroencephalogram signals), etc. They further improve the effectiveness of multimodal emotional representation through methods like maximizing mutual information, difference learning, and evaluating consistency. They are committed to the multimodal recognition of multichannel human emotional signals to judge a person's emotional state more comprehensively and accurately. Emotion recognition covers text, audio, and video modalities. Compared to traditional sentiment analysis, ERC does not analyze utterances in isolation. Instead, it combines the context and dependency between speakers to track their emotional state within the conversation.

ERC research focuses on effective context modeling, speaker dependency, and feature fusion methods in multimodal data settings. Context modeling approaches typically use Graph Neural Networks(GNNs) [1], long short-term memory (LSTM) [2], and other structures to understand the relationships between speakers better. Multimodal feature fusion uses criteria such as mutual information maximization, differences, and consistency to strengthen and optimize its ability to fuse modal features effectively. In multimodal ERC algorithms, researchers primarily employ three core algorithmic steps: feature extraction, fusion, and classification. Feature extraction is responsible for extracting features

from multimodal conversations, while feature fusion employs various methods to facilitate and guide the integration of these features. The final step involves classifying the emotions of utterance in conversation. Let us take the multi-speaker conversation in the MELD(Multimodal EmotionLines Dataset) [3] as an example in Figure 1. There are complex interactions between speakers in a conversation, and each conversation contains text, audio, and video.

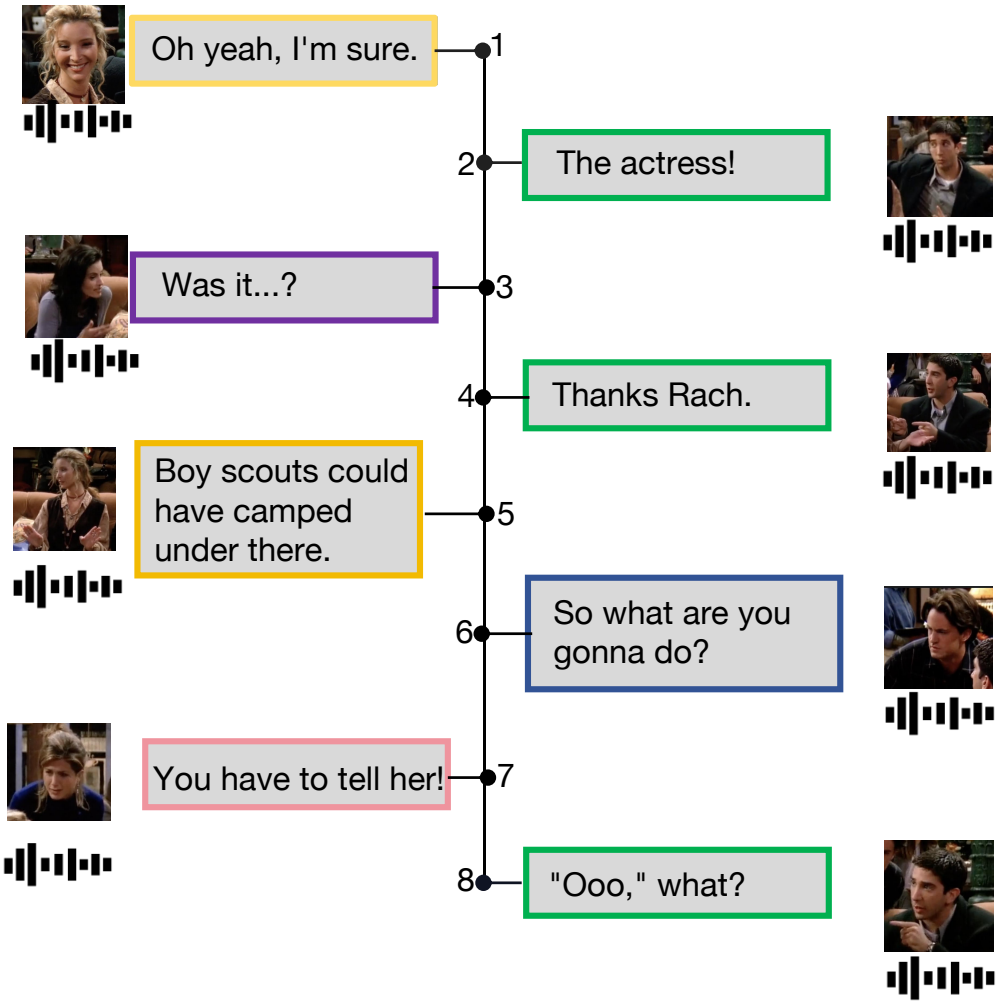

Figure 1. The Framework Used for Multimodal Emotion Recognition in Conversation.

As ERC is an emerging research field, an overview of its research progress and challenges, the available datasets, and the utilized benchmarks are helpful for future ERC research. Therefore, in this paper, we employ the current research resources to analyze and model the various factors that influence emotional dynamics in conversations. Furthermore, our paper not only provides insights into the current challenges and latest research findings in the field of ERC but also highlights future challenges and potential approaches to address these challenges. Our work aims to:

(1) We Summarize and analyze the relevant work about emotion recognition in conversation. Our work helps researchers fully understand the mainstream methods, motivations, and methods in this research field and provides detailed information on the available resources for beginners to learn.

(2) We classify the existing algorithms from three essential perspectives that affect multimodal emotion recognition tasks: speaker dependence, contextual context, and multimodal data fusion. We provide a detailed description and analysis of these methods, thoroughly evaluating each.

(3) We introduce real-life emotion recognition application scenarios and explain some problems encountered in these real scenarios. Furthermore, we elaborate on the challenges faced and suggest future research directions.

2. Emotion Categorization

Sentiment analysis is defined as the ability to perceive, integrate, and understand human emotions, and its corresponding classification system is an organizational structure for systematically classifying human emotions. An effective emotion classification system should be able to scientifically contain and reasonably classify and describe the various emotions that human beings possess. In the research field of sentiment analysis, some familiar and mainstream sentiment classification models are available, including the Ekman model [4], the pleasure, arousal, and domination (PAD) model [5], and the emotional wheel model of Plutchik [6]. Utilizing these taxonomic systems enables the intricate and elusive range of human emotions to be categorized into distinct categories, dimensions, and facets, facilitating the capture and comprehension of the diverse attributes inherent in each emotion. During the dialogue process, a machine can accurately perceive the user's emotions with the help of its estimated emotional classification mechanism and perform corresponding processing steps, which significantly enriches the content of the dialogue and makes the user respond with empathy, thereby improving the user's emotions.

Regarding discrete emotion classification systems, researchers believe that emotion is a psychological and physiological process caused by the cognition of developmental events and factors that trigger changes in internal and external psychological signals, thus dividing human emotions into limited categories, including emotions such as happiness, sadness, anger, disgust, and surprise, among others. According to different theories, two to eight basic emotions can be used to divide the standard of emotion classification. Ekman proposed seven characteristics for distinguishing between basic emotions and emotional phenomena: autonomous evaluations, specific antecedent events that are present in other primates, phenomena with rapid onsets, emotions with short durations, emotions with unconscious or involuntary appearances, emotions reflected in unique physiological systems such as the nervous system, and facial expressions. R. Plutchik proposed a Plutchik's wheel model composed of eight emotions based on the observation and research of human emotion expressions and the generalizability of emotions. This model is shown in Figure 2 below.

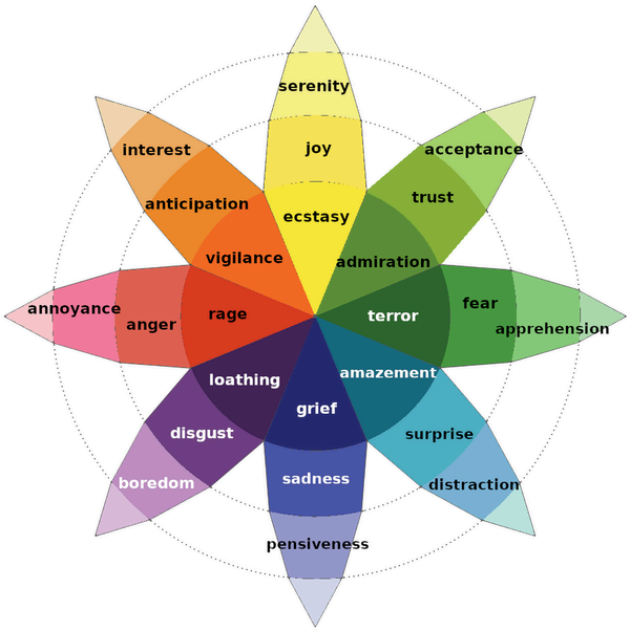

Figure 2. Plutchik's Wheel Model.

Discrete models classify sentiment into a limited number of discrete categories, which are limited in capturing similar sentiments and subtle variations. The dimensional emotion model provides a more detailed emotion description and measurement method. By treating emotion as a point in a multidimensional space and mapping the emotion to a continuous frequency spectrum, we can compare emotional states more intuitively and accurately through vectors and describe the complexity and variety of emotions. The dimensional emotion model in two-dimensional space typically utilizes the arousal-valence mode [7]. Valence reflects positive or negative evaluations of the body's intensity or activation of emotions. Arousal reflects an individual's will, low arousal indicates less energy or a lower emotion degree. This model is shown in Figure 3 below.

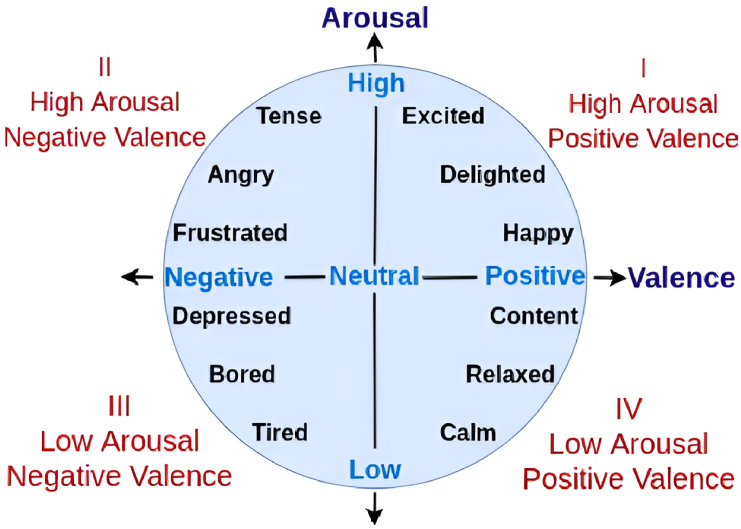

Figure 3. Arousal-Valence Mode.

Afterward, Mehrabian and Russell [8] proposed the most famous three-dimensional emotional model, the PAD model, through research on environmental psychology methods and the feeling-thinking-action model. In the PAD model, pleasure refers to the positive or negative feelings of emotions, arousal refers to the intensity or degree of emotional activation, and dominance refers to the degree to which emotions control the behavior of individuals. Considering emotion changes across these three dimensions, the PAD model provides a more comprehensive framework for emotion classification and understanding. This model is shown in Figure 4 below.

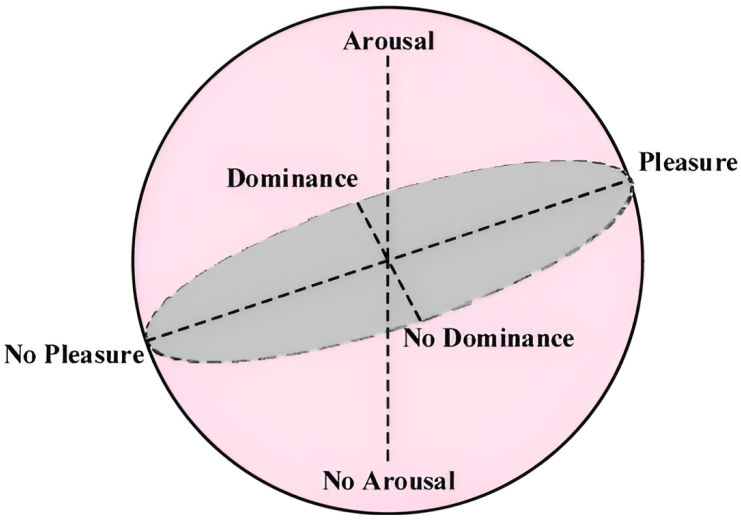

Figure 4. The PAD 3D Emotion Model.

The PAD model is widely used in emotion-related research, and it helps people describe and explain emotional experiences more accurately. The model can better capture emotions' subtle changes and complexities and provide valuable tools and theoretical foundations for sentiment analysis, user experience design, and psychological research.

3. Datasets

In this section, we introduce the datasets used for emotion recognition in conversation in detail, provide comprehensive information about these datasets, and describe the creation processes of the datasets, including the data collection and annotation steps. In addition, we provide the source of each dataset, ensuring the reliability and verifiability of the data.

3.1. MELD

The MELD dataset is based on video in the TV series 'Friends,', including multi-person conversations among nine main characters. To ensure that only single-speaker speech and images are included in the conversation, scenes containing only one speaker are extracted from the original videos. The entire dataset contains 1433 dialogues and 13,708 utterances. Each utterance has two annotation levels: the first annotation is one of seven emotions (neutral, happy, surprised, sad, angry, disgusted, or fearful), and the second level is one of three sentiments (positive, negative, or neutral). Non-neutral emotions account for more than 53% of the overall dataset. The MELD dataset is widely used in multimodal dialogue emotion recognition because it contains unambiguous facial expressions and speech emotions, providing more detailed information for emotion classification. However, one disadvantage of this dataset is that the conversation content is usually script-based, increasing the difficulty of emotion recognition. Nevertheless, the MELD dataset still provides researchers with a valuable resource for exploring and researching multimodal emotion recognition in conversation.

3.2. IEMOCAP

IEMOCAP [9] is an emotion recognition dataset recorded by the SAIL Laboratory of the University of Southern California. This dataset covers the conversation scripts of 10 professional actors, including text and emotional expression information. The dataset is divided into five sessions, each containing a male actor and a female actor. Each conversation consists of a scripted and freely interactive part where the actors interact in specific situations. In total, the dataset contains 151 conversations with 7433 utterances. Each utterance is annotated with one of six emotions: neutral, happy, sad, angry, frustrated, or excited, with non-neutral emotions accounting for 77% of the dataset. One drawback of this dataset is that, despite encompassing annotations for six distinct emotions, only four are typically utilized in training and recognition.This preference arises due to the relative ease in learning and distinguishing these four emotions compared to the others. The IEMOCAP dataset provides a valuable resource for researchers to explore and study ERC. It features authentic conversation and emotional expressions with professional actors, making the dataset more realistic and diverse.

3.3. DailyDialog

DailyDialog [10] is a high-quality multi-turn conversation dataset containing only plain text with less noise. The conversation in this dataset reflects dialogue scenes concerning different topics in daily life without a fixed speaker identity. In addition to 7 types of emotional annotations, the dataset also provides ten types of topic annotations and four types of dialogue behavior annotations. The entire dataset includes 12,218 conversations containing 103,607 sentences. The emotion annotations cover the following seven emotions: neutral, happiness, surprise, sadness, anger, disgust, and fear. Among them, non-neutral emotions account for 16.8% of the dataset. While DailyDialog may have limited applications in emotion recognition, it offers significant advantages, notably its expansive data scale. Nevertheless, a notable limitation of this dataset is the excessive proportion of

neutral emotions, which poses specific challenges to the task of emotion classification and may necessitate reduction for improved performance. Overall, the DailyDialog dataset provides researchers a valuable resource for exploring multi-turn emotion recognition and related fields.

3.4. EmoryNLP

The EmoryNLP [11] dataset was created by extracting text from the first four seasons of the TV series 'Friends' and adding emotion annotations. This dataset constructs a corpus containing 897 conversation utterances and 12,606 utterances from 97 TV episodes, providing variety. Each utterance is labeled with one of seven emotions: sad, mad, scared, powerful, peaceful, joyful, and neutral. The EmoryNLP dataset provides researchers with an exciting resource for emotion recognition and related research. This dataset uses TV drama conversations to provide rich emotional expressions and diverse emotional scenes, demonstrating a new perspective for exploring sentiment analysis.

3.5. EmotionLines

The EmotionLines [12] dataset is derived from two different data sources. The first data source includes multi-person conversations extracted from the TV series 'Friends', and the second data source consists of two-person conversations from private Facebook chat logs. The SocialNLP 2018 EmotionX Challenge used this dataset. The dataset can divided into two independent parts, each containing 1000 conversations and 29,245 sentences. The dataset has seven emotions: neutral, happiness, surprise, sadness, anger, disgust, and fear. Among them, non-neutral emotions account for 44.5% of the dataset. This dataset combines two conversational scenarios, TV dramas, and private chat records containing multi-person and two-person conversations.

3.6. EmoContext

This dataset is constructed from two-person conversations in plain text, with three utterances in each conversation, and only the last utterance has an emotional label.SemEval-2019 Task 3 [13] used this dataset. The dataset contains thirty-eight thousand four hundred twenty-one dialogues with 115,263 sentences. Four types of emotions are marked: happiness, sadness, anger, and other; non-neutral emotions account for 42.8% of the dataset. Its advantage is that the data scale is significant, and its disadvantages are that the dialogue length is too short, and only the last sentence is marked.

3.7. M3ED

Multimodal Multiscene, Multilabel Emotional Dialogue (M3ED) [14] is a large-scale, high-quality, multimodal, multiscene, multilabel emotional conversation dataset. The dataset includes three modalities: voice, text, and video. More than 900 conversation clips selected from 56 TV dramas, each utterance is marked with multi-emotional labels (for a total of 24,449 sentences) using the six mainstream essential emotional labels (happy, surprised, sad, angry, disgust, and fear) and a class of neutral emotions (for a total of seven discrete emotions). The inter-annotator agreement score reaches 0.59, significantly higher than the 0.43 of the MELD dataset and the 0.48 of the IEMOCAP dataset. In addition, this dataset is also the first multimodal interaction dataset in China, which has an essential supplementary role in the field of affective computing and is vital for promoting cross-cultural emotion analysis and recognition research.

We provide detailed statistics in Table 1, including the number of data samples, the conversation length, and the number of participants in each dataset. This information can assist researchers in assessing the comprehensiveness and inclusivity of various datasets, thereby enhancing their suitability for emotion recognition in conversation research. We also conduct a detailed analysis of the scale and classification of each emotion category in the datasets in Table 2, including the number of categories and the proportion of each emotion category, and explore the balance and distribution among the categories. Such

an analysis can help researchers understand the importance and attention of the different emotions in each dataset and provide a valuable reference for further emotion recognition research. Through these detailed introductions and analyses, readers can fully understand the characteristics and usability of the currently available emotion recognition datasets. These provide an essential foundation and guidance for subsequent research work.

Table 1. Comparison Among different emotion recognition datasets.

Dataset	Data Type	Train	Dev	Test
MELD	Utterance	9989	1109	2610
	Dialogue	1039	114	280
IEMOCAP	Utterance	5810	1623	
	Dialogue	120	31	
DailyDialog	Utterance	87,832	7912	7823
	Dialogue	11,118	7912	7863
EmoryNLP	Utterance	9934	1344	1328
Emotion Lines	Utterance	10,561/10,733	1178/1202	2764/2807
	Dialogue	720/720	80/80	200/200
EmoContext	Utterance	30,160	2755/5509	
M3ED	Utterance	17,427	2871	4201
	Dialogue	685	126	179

Table 2. Label distribution statistics of different emotion recognition datasets.

Label	MELD	IEMOCAP	Daily Dialog	Emory NLP	Emotion Lines	Emo Context	M3ED
Neutral	6436	1708	85,572	3776	6530	-	10,028
Happiness/Joy	2308	648	12,885	2755	1710	4669	2287
Surprise	1636	-	1823	-	1658	-	1051
Sadness	1002	1084	1150	844	498	5838	3957
Anger	1607	1103	1022	1332	772	5954	5234
Disgust	361	-	353	-	338	-	1497
Fear	358	-	74	1646	-	-	395
Frustration	-	1849	-	-	-	-	-
Excitement	-	1041	-	-	-	-	-
Peace	-	-	-	1190	-	-	-
Powerful	-	-	-	1063	-	-	-
Other	-	-	-	-	-	21,960	-

4. Feature Extraction

A rich repository of emotional content can be discerned from the various modalities of data generated within a conversation, including text, images, and video. Textual elements encapsulate emotional expressions within speech, images document facial expressions and bodily gestures, video records the dynamic evolution processes of emotional displays, and speech captures a spectrum of emotions through intonation and verbal cues. These different modalities are interrelated and complement each other's information. Feature extraction is a critical step of ERC. Effectively extracting emotional features from data can help computer systems and algorithms better understand emotions and represent and utilize multimodal data, thus achieving better results in sentiment analysis and emotion recognition tasks. We summarized the feature extraction techniques in multimodal models and listed them in Table 3, and listed extraction tools for unimodal ERC in Table 4.

Table 3. The Feature extraction techniques employed within multimodal models.

Model	Textual	Visual	Acoustic
MMGCN	TextCNN	DenseNet	OpenSmile
DialogueTRM	BERT	3D-CNN	OpenSmile
Emocaps	BERT	3D-CNN	OpenSmile
MMT	RoBERTa	DenseNet	OpenSmile
CMN	TextCNN	3D-CNN	OpenSmile
ICON	TextCNN	3D-CNN	OpenSmile
COGMEN	sBERT	OpenFace	OpenSmile/LibROSA
MMDFN	TextCNN	DenseNet	OpenSmile
GraphMFT	TextCNN	DenseNet	OpenSmile
MMDAG	RoBERTa	DenseNet	OpenSmile
Multilogue-Net	GloVe	Facet	OpenSmile
C-LSTM	TextCNN	3D-CNN	OpenSmile

Table 4. The feature extraction techniques employed within unimodal models.

Model	Textual
MuCDN	RoBERTa
DialogueRNN	Word2vec
EmoBERTa	RoBERTa
DialogueGCN	GloVe
DAG-ERC	RoBERTa/BERT
SGED	RoBERTa
S+Page [15]	GloVe
HiTrans	BERT

4.1. Textual Feature Extraction

When incorporating textual data into machine learning or deep learning frameworks, it is imperative to acknowledge that such data frequently exhibit an unstructured nature. It is necessary to conduct text feature extraction to adapt these data for use in these domains, which will convert the text data into a vector representation. The earliest text feature extraction method that emerged was the bag-of-words model [16]. This model is based on segmenting the utterances in the given corpus and creating a vocabulary where each word gives a unique index. The utterances are converted into a one-hot encoding vector, represented by placing the words in the bag of words at the corresponding index positions as 1 s. However, this approach must consider the order and contextual relationships between the words in a sentence. Moreover, as the corpus expands, the feature vectors corresponding to different sentences become high-dimensional and sparse, leading to computational and processing difficulties.

With the continuous progress exhibited by neural network technology, the word embedding technique has been widely used in text feature extraction tasks. This technique effectively reduces the computational burden imposed on the utilized model by calculating the similarities between words and mapping similar words into low-dimensional and dense vectors. Among the many available word embedding methods, Word2vec [17] is the most widely used approach. Word2vec is a word vector generation model that generates high-quality word vectors by training on a large-scale corpus. Depending on the utilized training method, Word2vec techniques can be classified into two categories: continuous bag-of-words (CBOW) and skip-gram models [18]. CBOW models require a word to be predicted based on the context of the current word, whereas the opposite is true for skip-gram models. These methods can capture the correlations between words and generate word vectors with low dimensionality and high density, thus exhibiting more generality.

In recent years, large-scale text pre-trained models (e.g., bidirectional encoder representations from transformers (BERT) [19] and RoBERTa [20]), which have achieved excellent results in numerous natural language processing (NLP) tasks, have gradually emerged

and become mainstream applications. BERT employs two critical tasks in its training process: masked language modeling (MLM) and next sentence prediction (NSP). These two tasks improve language models' comprehension of intersentence relations and contextual associations. RoBERTa, built upon BERT's foundation, was fine-tuned for the Masked Language Modeling (MLM) task by removing the Next Sentence Prediction (NSP) task and undergoing training on a more extensive and more diverse textual dataset. As opposed to BERT, which fixes masking markers in different training rounds, RoBERTa reselects the positions of the masking markers in every round; this practice endows the model with a more robust understanding of contextual relations. Due to the solid representational capabilities of pre-trained language models, they have become the most commonly chosen tools for representation learning by researchers. Compared to the traditional method of extracting word and sentence vectors, pre-trained models based on large-scale precondition training make the sentence features of the given text more semantic.

4.2. Visual Feature Extraction

Extraction of visual features involves gathering information from videos, which includes aspects like facial expressions, head movements, and body postures. This process is broadly categorized into two primary approaches. One approach involves the utilization of neural networks, specifically convolutional neural networks (CNNs), as introduced by Krizhevsky et al. in their work on ImageNet [21]. These networks are employed to conduct convolutional computations on video data, in which consists of multiple sequential image frames, deriving continuous visual features. For example, Tran et al. [22] proposed an efficient and straightforward deep three-dimensional CNN (3D-CNN) for capturing spatiotemporal features from an input video. Compared with the traditional 2D-CNN, the 3D-CNN adds a depth channel, which can be either a video frame or a different part of a stereo image, which makes the 3D-CNN more suitable for some scenes.

The second method employs specific software libraries, namely, OpenFace [23] and Facet [24]. OpenFace, the prevailing tool in current practice, initially conducts frame-level processing on the input video to extract a comprehensive set of features. These features encompass 68 key facial landmarks, 17 facial action units, and head posture, orientation, and eye gaze measurements. The user selects features based on a predetermined frame rate corresponding to the visual attributes in the current video. In addition, Facet extensively extracts visual features from the video, resulting in a more precise retrieval of facial action unit characteristics than OpenFace, which excels primarily in facial detection tasks.

These methodologies exhibit significant advantages and potential for practical applications in visual feature extraction, allowing researchers diverse options to cater to their distinct research requirements. Facet and OpenFace are frequently selected as the primary visual feature extraction tools across a broad spectrum of emotion recognition and classification tasks. This preference arises from their capacity to delve deeply into the information encapsulated within the visual modality by adopting a multifaceted approach. Consequently, researchers can effectively pinpoint and preserve vital task-related information.

4.3. Audio Feature Extraction

In emotion recognition during conversational interactions, acoustic signals assume a pivotal role. Audio feature extraction focuses on capturing key signal features from the sound of the current sentence that can represent emotions, including time-domain features such as the root-mean-square energy and zero-crossing rate and frequency-domain features such as mel-scale cepstral coefficients (MFCCs). These sound features help reveal the emotional fluctuations behind sounds and enrich the diversity of emotion recognition. Current dominant techniques for audio feature extraction primarily rely on open-source libraries, including widely used ones like LibROSA [25], openSMILE [26], and COVAREP [27].

LibROSA is a commonly used audio processing library that extracts acoustic features from audio data by default at a sampling rate of 22,050 Hz. These features include 1-dimensional logarithmic fundamental frequencies, 20-dimensional MFCCs, and

12-dimensional constant Q spectral coefficients. In addition, LibROSA also supports manual settings for extracting other features, such as zero-crossing rates. COVAREP is another audio feature extraction library that can extract various audio features. Its internal resampling speed is fast, and its feature extraction process often takes a shorter period. The openSMILE toolkit is a modular and flexible feature extraction tool capable of extracting a wide range of audio features. In addition, openSMILE also provides several statistical functions, such as the second-order mean, the discrete cosine transform, and linear predictive coding, which can further process the features extracted from audio data.

These open-source libraries offer a wide range of functionalities and flexibility. They empower us to choose appropriate feature extraction techniques. We can also use these libraries to process the extracted features further, enhancing audio features' accuracy and richness of information for multimodal dialogue emotion recognition tasks.

5. Methods

Two main research directions involve traditional ERC algorithms. First, emotions are affected by context in conversation, so the utilized model must construct this context to provide richer information. Second, emotions are also related to the speaker's state, but it is challenging to model the emotional dependencies between speakers effectively. Moreover, as researchers apply multimodal data in the context of emotion recognition during conversations, they are dedicated to leveraging multimodal information to advance the progress of this field. This multimodal information includes the linguistic content of conversations, a speech's vocal characteristics, and a speaker's facial expressions. An effective multimodal fusion mechanism can alleviate the deficiency of unimodal and obtain richer emotional information from different perspectives. This chapter focuses on three core research directions of ERC: methods for constructing dialogue context, modeling the dependencies between speakers, and fusing multimodal representations. By delving into these aspects, we can broaden our research horizon and bring new ideas to ERC.

5.1. Context Construction

Researchers utilize diverse methodologies and techniques, including sliding windows and hierarchical models, to thoroughly investigate context modeling approaches. This enables them to harness the surrounding utterances effectively. In this section, we explore the context modeling strategies and skills of different models to help researchers better understand the applications and effects of technical methods in emotion recognition tasks, and we demonstrate the comparative impact of various models in Table 5.

Table 5. Comparison of the performance of aforementioned methods on two primary datasets, IEMOCAP and MELD, with Wa-F1 indicating the weighted average F1 score. For more detailed insights, please refer to the paper.

	Happy	**Sad**	**Neutral**	**Angry**	**Excited**	**Frustrated**	**Wa-F1**	**Wa-F1**
DialogueRNN	33.83	69.83	57.76	62.5	64.45	59.46	59.89	57.03
C-LSTM	47	79.9	56.40	62.3	71.4	59.25	59.19	-
DialogueGCN	42.75	84.54	43.54	64.19	63.08	66.99	64.18	58.1
EmoCaps	71.91	85.06	64.48	68.99	78.41	66.76	71.77	64

5.1.1. Sequential Models

The preliminary endeavors in modeling the inter-dependencies among contextual utterances encompassed sequential models, such as LSTM, recurrent neural networks (RNNs) [28], and gated recurrent units (GRUs) [29]. These models iteratively extract historical utterances and retain the sequential organization of the conversational components, ensuring the suitability of the models for preserving the chronological continuity of the conversation. Wollmer et al. [30] adopted an LSTM-based RNN that could explicitly learn to perform clustering in the emotional space and simulate contextual knowledge to achieve

improved performance. Nevertheless, in models such as LSTM, constructing a context capable of distinguishing between individual speakers still needs to be addressed.

Majumder et al. [31] posited that the emotional attributes in a conversation are contingent upon three primary factors: the speaker, the contextual information provided by the preceding utterance, and the contextual information embedded within the prior utterance. Hence, DialogueRNN incorporates these three factors, employing three GRUs to encode the overarching context (global state), participant-specific context (party state), and emotional nuances (emotion representations). The DialogueRNN model can comprehensively consider this information and perform joint coding for emotion recognition tasks through such a division of labor.

An alternative Transformer-Based Context and Speaker-Sensitive Model context modeling approach involve utilizing expansive, pre-trained language models such as BERT and RoBERTa to comprehend the contextual intricacies in conversation. Researchers leverage these models to augment the process of integrating contextual information. Transformer-Based Context- and Speaker-Sensitive Model(HiTrans) [32] combines multiple utterances marked with [CLS] and packs them into an input sequence. An utterance sequence with a length exceeding 512 first divides itself into blocks, passes through BERT, and then passes through another transformer. The low-level transformer generates the current discourse representation, and the high-level transformer further generates global contextual information embedded in the discourse representation. Kim et al. [33] integrates the speaker's identity into the context, enriching the context extracted by the transformer model. Although pre-trained on massive data texts provides a large-scale pre-trained model with more powerful semantic capabilities, difficulties concerning computing resources and context sequence length limitations still need to be solved.

Furthermore, the conventional sequential modeling approach needs to be improved in its ability to capture the impact of prior states on the current utterance, as it needs to pay more attention to the influence of future states on the present situation. This approach must fully exploit the bidirectional dependencies inherent in contexts, prompting researchers to shift their focus toward bidirectional historical modeling. A bidirectional model can represent the history from the past to the present and from the present to the past. It simultaneously considers the forward and backward information in the dialogue history, facilitating a more comprehensive understanding of the context.

Researchers have employed bidirectional RNNs, LSTM networks, and transformer models to establish chronological records for bidirectional dialogue modeling. C-LSTM [34] takes the features of each utterance as its inputs and processes them through a bidirectional LSTM unit. One approach involves sequentially processing utterances in the forward direction, while another entails reverse sequential processing. This methodology enables the model to concurrently incorporate the contextual information preceding and following each utterance, subsequently merging the resulting output features from both directions to construct a holistic, contextual feature. Bidirectional LSTM models capture the context of utterances better than unidirectional LSTMs. Models such as a bidirectional GRU (BiGRU) [35] and BiTransformer [36] are committed to improving the expression ability of dialogue history through bidirectional modeling to understand the emotions and context more comprehensively, and these approaches have achieved specific results in ERC tasks. The success of this method offers valuable insights for advancing research in emotion recognition and contextual understanding.

5.1.2. Graph-Based Methods

The primary method for modeling context is stacking utterances, thus incorporating facial structural constraints to effectively capture extensive, multimodal, and diverse contextual information across longer distances. Simultaneously, researchers have observed that conversation can be construed as innate graph structures. The evident correlations and inter-dependencies prevail among these sentences. Furthermore, conversation conven-

tionally involves multi-turn exchanges marked by intricate dependencies and interaction patterns. The evident correlations and inter-dependencies prevail among these sentences.

These interaction patterns can ideally exploit GNNs' edges to facilitate the holistic modeling of the emotional dynamics within the conversation. They were additionally driven by advancements in human-computer interaction technology and contextual scenarios; diverse node types can be employed within a graph structure for modeling multimodal conversation data, encompassing the text, audio, and video modalities. Dynamically adapting to the dimensions and intricacies of conversation, GNNs can process dialogic graphs with varying dimensions and configurations. This adaptability enables them to capture finer and more intricate dependencies in interactions.

In graph-based ERC tasks, conventionally, utterances in conversation are represented as nodes, and predefined composition rules establish the connections. Graph-based research typically centers on utilizing GNNs and Graph Convolutional Networks(GCNs) [37] within the graph construction paradigm. This approach encompasses the following principal components.

Node definitions: Nodes represent nodes in a session, and standard graph-based conversations are represented by G = (V, E). When only considering the text modality, the set of nodes V corresponds to the number of sentences in the conversation, represented as V = N. However, in cases where the conversation involves multimodal settings, including videos, audio, and text, the size of the node set V is expanded to 3 N.

Edges: Edge construction relationships mainly depend on the conversation context, such as the temporal, speaker's relationship. In contrast, the weight of an edge represents the strength of the association between two nodes. Ghosal et al. [38] distinguish the links of utterance nodes according to their temporal relationships for the first time, and this approach clarifies the characteristics by which other speakers influence the target speaker. Experiments have proven that it is essential to distinguish between different contexts and speaker dependencies in relational modeling cases. On this basis, the Multimodal Fusion via Deep Graph Convolution Network(MMGCN) [39] models relationships under multimodal dialogue settings.

As depicted in Figure 5, MMGCN establishes two distinct categories of edges that connect internal interactions within the same modality and interactions across different modalities. DAG-ERC [40] treat each conversation as a Directed Acyclic Graph(DAG); each utterance only accepts information from some previous utterances and cannot propagate information back to itself and the words of its predecessor. However, DAG-ERC focuses on computing the information flow between utterances and does not consider the interactions among conversations under the multimodal data setting.

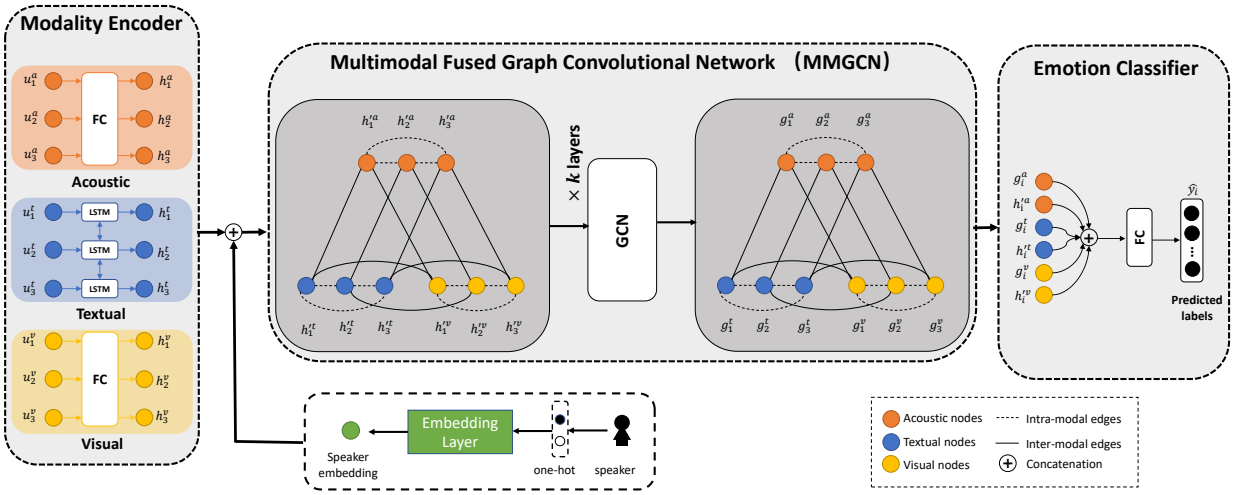

Figure 5. Multimodal Fusion via the Deep GCN (MMGCN) Proposed by Hu et al. [39].

On this basis, Xu et al. [41] proposed a multimodal DAG, which transmits information flows between nodes with the same modality and nodes across modalities. Information is only allowed to flow from previous utterances to the current utterance, and the current utterance is prohibited from passing information to itself or the previous utterances. Furthermore, due to the importance of the textual modality, this model only allows the visual and acoustic modalities to convey information to the textual modality, limiting the interaction and information flow between nonverbal modal features. This approach further improves the ability of a directed multimodal dialogue graph to capture multimodal conversation characteristics.

Weights: In the Edge weight setting, the MMGCN assumes that the higher the similarity between two nodes is, the more critical the information interaction between them is, so the edge weight between them is also greater. The MMGCN uses angular similarity to represent the weights between nodes. The graph comprises two distinct categories of edges: the connection between the same modality, as exemplified by Equation (1), and the connection between different modalities, as represented by Equation (2). In these equations, n_i and n_j refer to the feature representations of the i-th and j-th nodes within the graph, respectively. The symbol γ denotes a hyperparameter. The edge weight computation is performed as follows:

$$A_{ij} = 1 - \pi \arccos(\text{sim}(n_i, n_j)) \tag{1}$$

$$A_{ij} = \gamma(1 - \pi \arccos(\text{sim}(n_i, n_j))) \tag{2}$$

Since the neighboring nodes influence the current utterance node differently, GAT [42] is used to compute the edge weights. The traditional GAT, as defined in Equations (3) and (4), computes the scoring functions:

$$e_{ij} = \alpha([Wh_i, ||, Wh_j]), j \in N_i \tag{3}$$

$$A_{ij} = \frac{\exp(LeakyReLU(e_{ij}))}{\sum_{k \in N_i}} \exp(LeakyReLU(e_{ik})) \tag{4}$$

e_{ij} indicates the importance of the node's features to node i when incorporating the graph structure into the mechanism by considering Ni, which represents the neighborhood of node i in the graph. These values are subsequently normalized across all choices of j using the softmax function. Furthermore, the attention coefficients remain static in traditional GAT. Therefore, GATv2 [43] strategically relocates the LeakyReLU activation function between the weight matrix (W) and the subsequent non-linear layer, followed by the concatenation before applying a linear transformation with W. Empirical findings validate that GATv2 yields a more expressive attention mechanism, resulting in enhanced experimental performance:

$$A_{ij} = \frac{\exp(\sigma(a_\omega T[W_\omega x_i || W_\omega x_j]))}{\sum_{u_k \in N(u_i)} \exp(\sigma(a_\omega^T [W_\omega x_i || W_\omega x_k]))} \tag{5}$$

In this context, x_i corresponds to the feature representation of node u_i. Both u_j and u_k are neighboring nodes of u_i. Here, $u_k \in N(u_i)$ represents the neighborhood of u_i. ω_{ij} denotes the edge weight between u_i and u_j, and σ indicates the leaky rectified linear unit (LeakyReLU) non-linear activation function. W_ω and a_ω are adjustable parameters.

Each modality (e.g., text, sound, and images) uniquely expresses information and emotions in a multimodal ERC task. In addition to classic modeling methods based on intra-modal contextual relations and inter-modal interactions, Zhang et al. [44] transforms the task into a node classification problem; each sub-graph in the graph represents a conversation, and each node is an utterance in the conversation. In addition, nodes represent speakers in the whole graph. When constructing an edge, each utterance in

each conversation is first connected, and the weight of the connecting edge using angular similarity. Then, each utterance is connected with the corresponding speaker with an edge, and the weight of the edge is from the inverse speaking frequency of the speaker. Then, the graph is sent to a two-layer GCN to train and classify its nodes.

Graph-related research has yet to fully explore how to integrate the differences between different modalities effectively. Therefore, future research should further study a strategy for fusing multimodal features better to utilize the relationships and complementary information between different modalities. For example, we can explore how to design a more accurate weight calculation method to reflect the contributions of different modalities in ERC tasks.

5.1.3. Transformer-Based Methods

Transformers [45] are widely used in various downstream tasks in NLP(Natural language processing technology) due to their powerful sequence modeling and relationship capture capabilities. The multimodal emotion recognition task focuses mainly on the ability to represent different modal features and their fusion. Early ERC tasks in multimodal settings focused on improving the representation capabilities achieved for each modality. They improved the performance of traditional text-based emotion recognition models by enhancing the interaction and fusion of features. However, due to the representation capability differences among different modalities and data heterogeneity, interaction modeling introduces noise, which blurs the ability to represent multimodal information. Transformer's self-attention mechanism can capture the dependencies between utterances, making it an ideal choice for multimodal interaction modeling and learning that applies to various tasks and datasets.

DialogueTRM (A Novel Multi-Grained Interactive Fusion) [46], shown in Figure 6 below, explores different emotional behaviors from intra-modal and inter-modal perspectives. It builds a new layered transformer that can easily switch between sequential and feed-forward structures according to the contextual preferences within each modality. To achieve multimodal interaction fusion, it applies neuron- and vector-level feature interactions to learn the different contributions of individual modalities.

Li et al. [47] propose a new structure named Emoformer to extract multimodal emotion vectors from different modalities and fuse them with sentence vectors to be an emotion capsule and obtain emotional classification results through a context analysis model. A sequence-based approach employs a transformer-based context- and speaker-sensitive EDC model (Trans). It consists of two transformers. First, a pre-trained bidirectional transformer encoder generates a global utterance representation. Then, another high-level transformer captures the global information in the given dialogue, generates a global context, and combines speaker-sensitive tasks to judge whether sentences belong to the same speaker.

The Main Modal Transformer (MMT) [48] model utilizes a transformer architecture and consists of two attention mechanisms: cross-modal attention (Cm) and cross-task attention (Ct). Cross-modal attention learns the fusion relationships between different modalities. In contrast, cross-task attention learns the relationships between different tasks (e.g., sentiment analysis and emotion recognition). The main task of the MMT is to improve its multimodal feature fusion effect. It uses a two-level emotional cue extractor to extract emotional evidence. In addition, a cross-modal transformer (CMT) preserves the integrity of the dominant modal features and enhances the representations of weak modal features. Liu et al. [49] proposed a hierarchical dialog understanding model named HiDialog, shown in Figure 7, which performs sequence modeling by inserting unique tokens in conversations and introduces multi-turn and turn-level attention to learn embedding representations. In addition, the model utilizes a heterogeneous graph network [50] to optimize the learned embeddings.

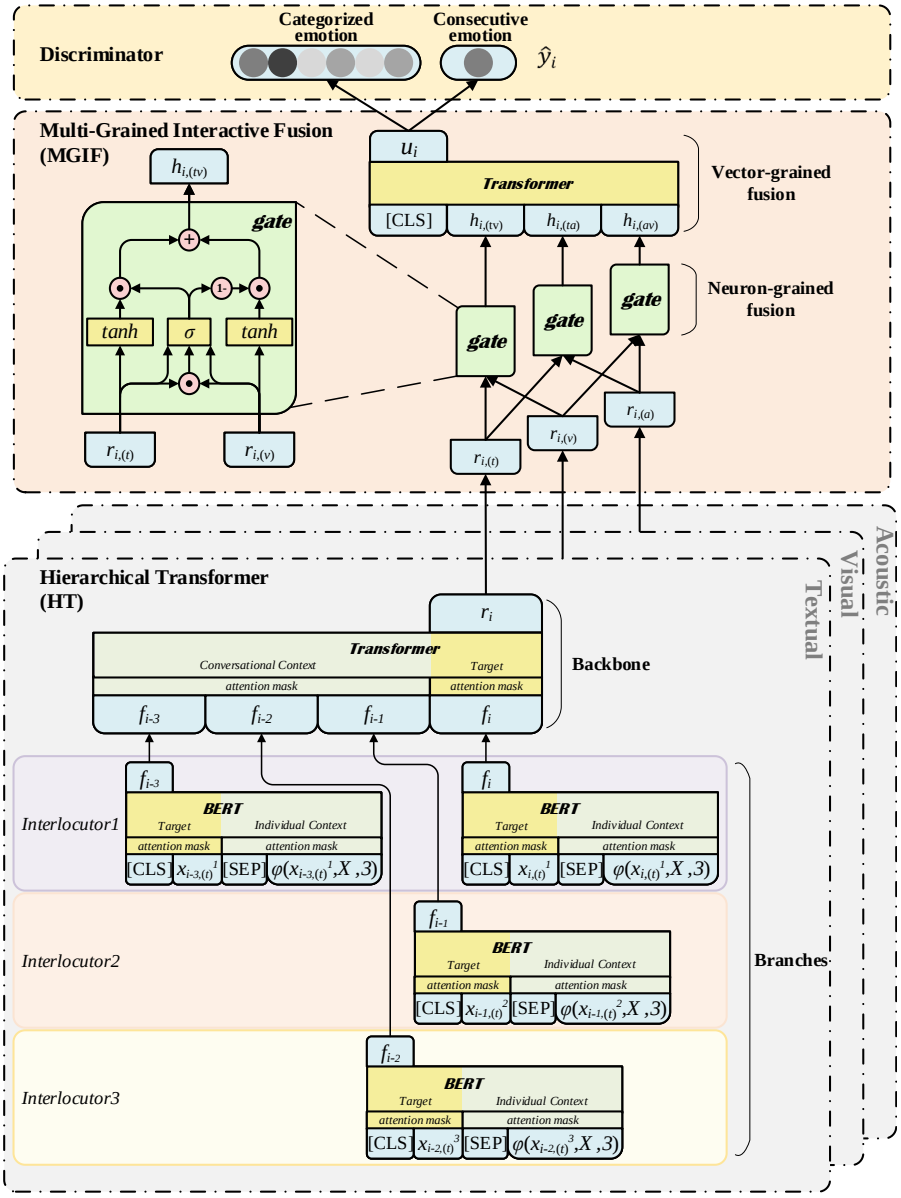

Figure 6. The DialogueTRM Method Proposed by Mao et al. [46].

This subsection summarizes the experimental results of different models in context modeling and comparisons among their performances. The performance of the models compared in terms of accuracy and generalizability, and the advantages and disadvantages of different models. The future direction is to examine the complementarity and redundancy in multimodal context features to improve the robustness and performance of prediction methods. By intensely studying the correlations and interactions between different modalities, more powerful multimodal feature fusion methods can be designed better to capture the rich information in these different modalities. In summary, there is some progress in multimodal ERC research. However, further in-depth research and explorations are still needed to give full play to the advantages of multimodal differences and rich features. We expect multimodal ERC's performance and application to be improved by adopting more effective fusion strategies and deeper feature analyses.

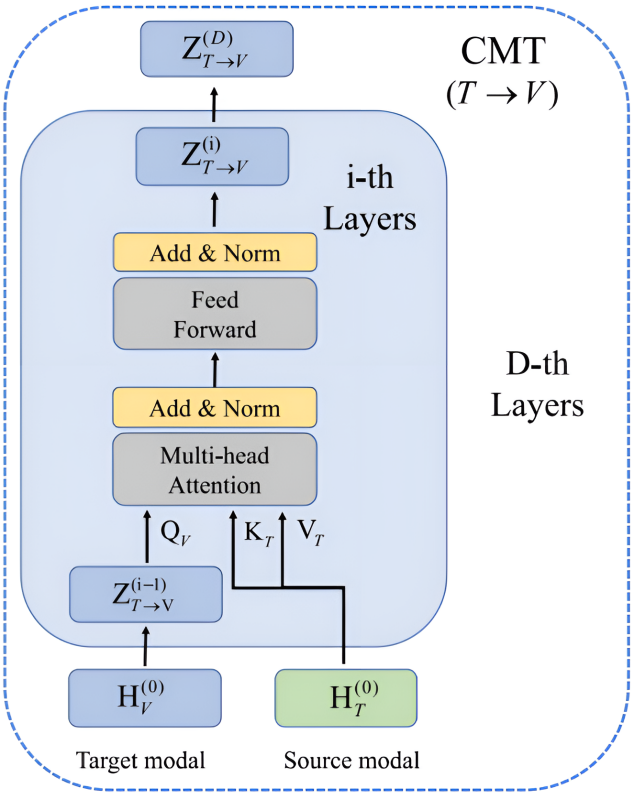

Figure 7. The architecture of CMT proposed by Zou et al. [48].

5.2. Speaker Dependency

Speaker dependency plays a pivotal role in conversational dynamics. It encompasses two distinct facets: intra-speaker dependency and inter-speaker dependency. Figure 8, sourced from the publicly accessible IEMOCAP dataset, illustrates a dialogue where Participant A (P_a) begins distressed during the initial exchanges (U_1 and U_3) and seeks consolation. Participant B (P_b), however, responds with sarcasm. Consequently, this interaction influences the speaker's emotional state to remain ostensibly neutral throughout the dialogue. In contrast, P_a persists in a state of distress at utterances U_1 and U_3, and due to the intra-speaker dependency, this state carries through to U_5. Here, P_a is further influenced by P_b's state—illustrative of inter-speaker dependency—and reacts with anger. This exemplification underscores the importance of examining speaker dependency within ERC tasks, offering insights into speakers' nuanced and implicit emotions in complex conversation environments.This section sorts and summarizes the speaker dependency modeling approaches proposed in existing work from several perspectives, we demonstrate the comparative impact of various models in Table 6.

Table 6. Comparison performance of the speaker-dependency modeling approach on two main datasets.

Methods	Happy	Sad	Neutral	Angry	Excited	Frustrated	Wa-F1	Wa-F1
CTNet	51.3	79.9	65.8	67.2	78.7	58.8	67.5	60.5
MM-DFN	42.22	78.98	66.42	69.77	75.56	66.33	68.18	59.46
ICON	32.8	74.4	60.6	68.2	68.4	66.2	63.5	56.3
SGED	–	–	–	–	–	–	68.53	65.46
COGMEN	51.9	81.7	68.6	66.0	75.3	58.2	67.6	–
HiTrans	–	–	–	–	–	–	64.50	61.94
Emoberta	–	–	–	–	–	–	68.57	65.61
DAG-ERC	–	–	–	–	–	–	68.03	63.65

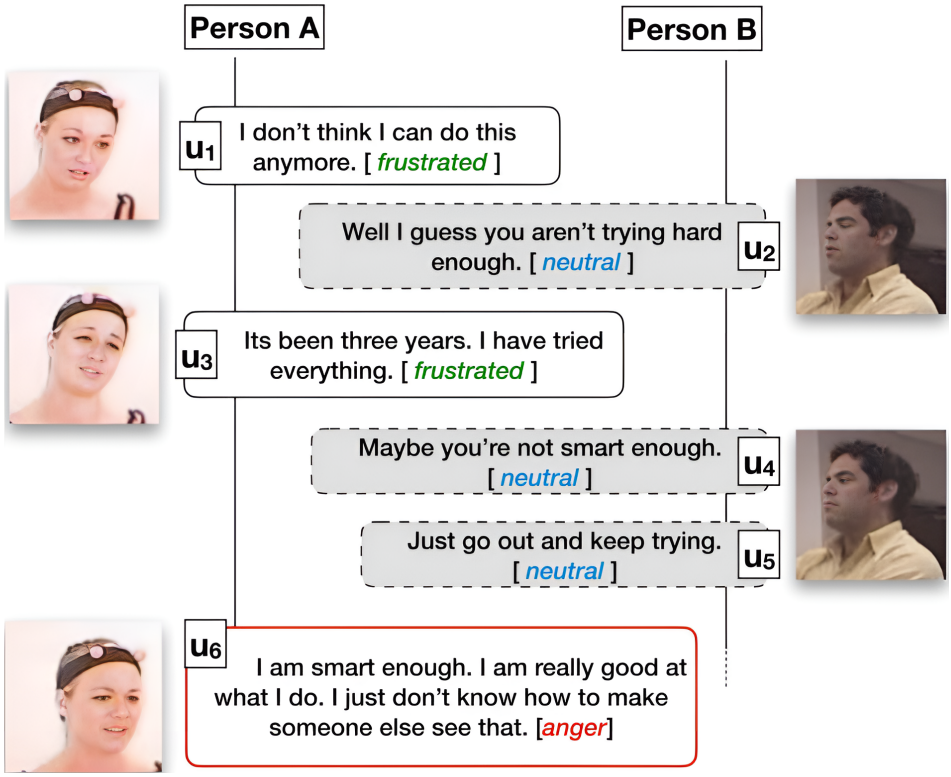

Figure 8. A Representative Dialogue Sample from the IEMOCAP Dataset.

5.2.1. Embedded Speaker Dependency

Embedded speaker dependency modeling refers to the implicit exploration of speaker dependencies by representing speaker information as sentence features, and the research concerning such methods has focused on how to obtain speaker features efficiently. Conversational Transformer Network (CTNet) [51] was primarily designed to address the intricate task of bimodal emotion recognition within conversational contexts, encompassing both textual and audio modalities. In pursuit of explicitly incorporating speaker-related information into multimodal sentence representations, CTNet undertakes a multistep process. CTNet extracts a 512-dimensional utterance-level speaker embedding from audio MFCCs through the x-vector system and combines the speaker embedding with the unimodal speaker embedding. The speaker embedding concatenated with the subsequent modeling steps' unimodal and cross-modal input features.

Unlike CTNet, which uses a sequential structure to simulate conversation, the Multimodal Dynamic Fusion Network (MM-DFN) [52] and MMGCN simulates the structure of an undirected graph. Before building the graph, to incorporate the speaker's information into the graph, the MMDFN and MMGCN first convert the speakers involved in the current dialogue into one-hot codes and then obtain the corresponding speaker features from the one-hot codes of the current sentence through a linear layer. Then, they splice the speaker features with the three modalities of the corresponding sentence to obtain the three modal features containing the speaker information. ConGCN primarily focuses on the textual unimodal aspect of conversations, representing dialogues as undirected graphs. Unlike conventional approaches, ConGCN initializes nodes in a graph, excluding those of the dialogue sentences, as random vectors. These initialized nodes serve a specific purpose within the graph (known as 'speaker nodes') and are dynamically updated during graph evolution. This update mechanism aids the model in exploring speaker-related characteristics, thereby facilitating the modeling of speaker dependencies. The strength of this embedded speaker dependency modeling approach lies in its simplicity. However, it needs to be improved in effectively capturing evolving speaker states throughout a dialogue, as this issue poses challenges in comprehensively modeling intricate speaker dependencies.

5.2.2. Dynamic Speaker Dependency Based on Sequential Structures

Due to the limitations of embedded speaker dependency modeling, researchers proposed dynamic speaker dependency modeling to improve the ability of models to capture the speaker dependencies that interact in conversations. Dynamic speaker dependency modeling means the utilized model generates or updates speaker-specific information based on historical dialogue or historical speaker information during the dialogue process. The existing dynamic speaker dependency modeling approaches are mainly divided into two categories, sequential structure-based and graph-based methods, according to their dialogue modeling techniques; this section provides an overview of the existing sequential structure-based dynamic speaker dependency modeling approaches.

Earlier work on ERC focused on conversational context, often using structures such as LSTM to integrate contextual information. However, this method ignores the essential impacts of the influence relationships between speakers and themselves on speaker sentiment in a dialogue. The Conversational Memory Network (CMN) [53] is the first method to consider the ERC task's inter-speaker dependency and intra-speaker dependence features. The CMN first obtains the three modal features corresponding to the current sentence, splices them to obtain multimodal features, and feeds the sentence within the context window to a GRU to obtain the features corresponding to the current moment containing contextual information. Afterward, the CMN sifts through the information between the current sentence features and the previously obtained sentence features containing contextual information through an attention mechanism, obtains the degree of influence between the two types of features in the form of weights, and performs a weighted summation operation to obtain the speaker features corresponding to the current sentence.

Interactive Conversational Memory Network(ICON) [54], on the other hand, introduces an interactive memory unit for multiparty conversation based on the CMN, and its self-influence module (SIM) designed to model intra-speaker dependency by integrating all historical sentences of the speaker corresponding to the current moment with the help of a GRU. Then, the GRU in the dynamic global influence module (DGIM) is used to model the influence between the SIM memory obtained from the SIM module at the current moment and the global memory obtained from the DGIM at the previous moment for the speakers; this step obtains the global state at the current moment and stores it in the memory unit. Subsequently, the memory unit is fused with the current discourse representation using the attention mechanism to determine the final emotion prediction. DialogueRNN is used to model the conversational utterance, the context, and the current speaker's emotional state with the help of three GRU units, representing the global state, the party-state, and the emotion representation, respectively.

Similar to DialogueRNN, Shenoy proposed a Context-Aware RNN (Multilogue-Net) [55], which uses multiple GRUs to account for the interlocutor state, the interlocutor's intention, the previous and future emotions, and the context of the dialogue. Unlike DialogueRNN, Multilogue-Net interacts with the information based on pairwise attention to obtain information for judging the speaker's emotion from different modalities. However, DialogueRNN and Multilogue-Net only consider intra-speaker dependency and ignore inter-speaker dependency. To this end, Zhao et al. [56] designed a Mutual Conversational Detachment Network (MuCDN) that splits the whole conversation into multiple sub-conversations, regarded as potential influence relations between speakers and calculates the relative lengths between sentences based on the discourse tree. Inter-speaker and intra-speaker GRUs can capture the dependencies between speakers within each sub-conversation. The former processes all sentences of the current speaker, while the latter handles all sentences of the non-current speaker. The drawbacks of this approach include disrupting the conversational flow, impeding the comprehension of contextual and posing challenges in revealing the intricacies of inter-speaker dependencies.

To address these issues, Bao et al. [57] introduces a novel structure for modeling speaker dependencies (SGED). SGED comprises two core components: a Conversational Context Encoder (CCE) and a Speaker State Encoder (SSE). The CCE generates current

sentence features enriched with contextual information, while the SSE explores intra- and inter-speaker dependencies. Intra-speaker dependencies are established by attending to the speaker dependency from the SSE in the previous moment for the current speaker, encompassing all sentences preceding the current moment. Eventually, the intra- and inter-speaker dependencies derived from the SSE undergo activation through a function, resulting in the speaker dependency at the present moment. The key advantages of this method lie in its effectiveness in probing the intricate interactions within a dialogue, and the SGED module seamlessly integrates with existing methods, enhancing overall performance.

5.2.3. Dynamic Speaker Dependency Based on Graph Structures

The previous section introduced a methodology for representing dialogues sequentially and modeling dynamic speaker dependencies. However, the disadvantage of a sequential structure is that distance limitations are observed between sentences, which makes it challenging to perform interactions between sentences at longer distances. Similarly, it is difficult for speaker features far from each other to interact effectively, limiting the sequential structure in terms of modeling long-distance contexts and speaker dependencies. With the continuous development of GNNs, researchers in the ERC field have found that simulating dialogues with graph structures can effectively address the shortcomings of sequential structures in cases with speaker interactions. During the process of simulating dialogue with a graph structure, the nodes in the graph often represent sentences at different moments. These sentences can be directly connected through edges, which means that no matter how far apart the sentences are, as long as they connect with edges, they can directly interact to explore the influence relationships between the speakers more effectively. In addition, the research development of GNNs and the advantages of graph structures in speaker interaction tasks have made exploring speaker dependencies with graph structures a mainstream strategy in recent years.

DialogueGCN defines various types of edges based on speaker identities. Sequential relationships help the model understand the flow of discourse and the corresponding speaker relationships, thus enabling it to simulate dialogues more accurately. Joshi et al. [58] proposed that the information reflecting speakers' emotions in dialogues comes from two primary sources: global information and local information. The global information is the context, and the local information includes the inter-/intra dependence between speakers. COGMEN uses the transformer encoder part of its position embedding module to obtain the global information and then constructs a directed graph by taking the sentence features containing global information as nodes, with four types of edges: past sentences of the current speaker, past sentences of the other speaker, future sentences of the current speaker, and future sentences of the other speaker. Different edges represent the relationships between the speaker's identities and the temporal information. The constructed directed graph is then fed into a Relational Graph Convolutional Network(R-GCNs) [59]and a graph transformer [60] to obtain sentence features with contextual dependencies and inter-/intra speaker dependencies.DAG-ERC combines the advantages of graphs and the constraints of conversation to determine the construction rules of the DAG: a direction constraint and a tele-information constraint. This compositional approach provides the relative position of the conversation and the speaker's identity, which helps the model better capture real-life contextual relationships and speaker dependencies.

In addition to the above methods, some speaker dependency modeling methods have also been developed from other perspectives, such as HiTrans, which designs an auxiliary task, determining whether two sentences are from the same speaker, and places the result into a final loss function to improve the model's sensitivity to speaker information. In addition, with the emergence and development of large-scale text pretraining models, some approaches centered on such models have emerged because training on a large corpus of conversation gives them strong speaker dependency comprehension capabilities. For example, Emoberta uses pre-trained RoBERTa as its core; for it to be able to grasp the influence relationships between speakers, Emoberta splices the sentences and their

corresponding speaker names before inputting them in RoBERTa, which enables the model to obtain both contextual and speaker information and thus integrate the influences of both aspects to give better judgments. However, the ERC method based on large pre-trained models is limited by the arithmetic power of the computer and the amount of training data, which makes it difficult to migrate between different scenarios effectively.

5.3. Multimodal Fusion

In prior research on ERC, a common oversight has been the neglect of differential emotional behaviors exhibited within and across various modalities when modeling conversational context. Developing effective modeling strategies for handling multimodal contextual information is instrumental in yielding more precise emotion prediction results. Furthermore, it is crucial to acknowledge the distinct contributions of emotional expression in multimodal settings. Words and sounds have proven more beneficial in predicting neutral emotions than the visual modality. Hence, when integrating multimodal information for emotion prediction, it becomes imperative to comprehend the unique contributions of each modality and transform them into fusion weights. However, a fundamental challenge arises because different modalities are often represented in distinct feature spaces, making assessing and quantifying their contributions directly intricate. Consequently, a significant avenue of research pertains to the effective fusion of multimodal contextual information in the context of multimodal ERC.

Compared with unimodal emotion recognition, multimodal emotion recognition has many advantages, including expressing richer information with the help of multiple representations. We demonstrate the comparative impact of various models in Table 7. Integrating visual information can reveal verbal cues such as facial expressions and body movements, while audio information helps convey characteristics such as the pitch and volume of a sound. This representation method for fusing multimodal data helps to consider emotional information from different perspectives comprehensively, thus making predictions more accurate, especially for classification situations that are easily confused in unimodal settings. In unimodal emotion recognition experiments, specific dialogue sentences can be incorrectly categorized as 'angry' or 'neutral,' particularly in the case of the 'frustrated' emotion category. However, when integrating concurrent analyses of the audio and video modalities, including visual cues such as frowning expressions in the video features and auditory characteristics such as increased volume in the audio features, the model enhances its ability to comprehend emotions, resulting in more precise classifications.

Table 7. Comparative performance analysis of the aforementioned multimodal fusion methods.

	IEMOCAP							MELD
Model	**Happy**	**Sad**	**Neutral**	**Angry**	**Excited**	**Frustrated**	**Wa-F1**	**Wa-F1**
GraphMFT	45.99	83.12	63.08	70.3	76.92	63.84	68.07	58.37
MMGCN	42.34	78.67	61.73	69	74.33	62.32	66.22	58.65
MMDAG	-	-	-	-	-	-	70.57	64.10
DialogueTRM	48.7	77.52	74.12	66.27	70.24	67.23	69.23	63.55

In early studies, to introduce multimodal features to ERC tasks with multimodal settings, researchers typically used cascading methods to integrate features to guide emotion recognition. The researchers proposed the CMN model by concatenating features from all three modalities but ignoring the interactions between the modalities. The CMN first uses the three modal features acquired from the input video, improving the accuracy and enhancing the robustness of emotion recognition from video. However, this method completely ignores the interactions between the processes of extracting helpful information in interactive scenarios. On this basis, researchers further proposed GME-LSTM [61] to perform multimodal information fusion in emotion recognition tasks for each utterance. The experiments showed that multimodal features to the LSTM model without an attention

mechanism would lead to declines in the F1 score, proving that although audio and video features provide rich content, they may also have noise.

Researchers in recent years have begun to explore more effective multimodal feature fusion methods, retain high-quality and adequate information during the fusion process, and reduce the influences of redundancy and noise, thereby improving the performance and accuracy of multimodal emotion recognition. Therefore, GME-LSTM uses a gating mechanism at each time step and passes the features through a temporal attention layer. At the same time, works such as DialogueTRM adopt attention mechanisms to guide effective multimodal feature fusion procedures.

However, as researchers have gradually discovered that GNNs can model long-distance contexts and different types of relationships, graph-based methods have become the mainstream approach. These methods ensure that their GNNs select the critical internal context and multimodal interaction information during learning by defining different types of multimodal sentence nodes and designing multimodal interaction information. The MMGCN was the first model to consider the combination of multimodal and contextual information. It uses undirected graphs to explore a more effective method for fusing multimodal interaction and contextual information. However, directly concatenating utterances from various modalities may introduce additional noise.

Furthermore, the MMGCN embeds utterances in a single GNN simultaneously with utterances within other modalities, which poses challenges for multimodal fusion. To solve the above problems, GraphMFT [62] adopts multiple improved GATs to extract the contextual dependencies within modalities and the complementary dependencies between modalities, thus effectively promoting the current discourse and intra-modal and inter-modal discourse interactions. For example, GraphMFT introduces a loss function with four subspaces to constrain the extracted multimodal features for alleviating the heterogeneity problem encountered in multimodal ERC, and it incorporates the PairCC strategy to solve the information propagation direction limitation. We employ a multisubspace mapping function and a PairCC strategy to address the heterogeneity gap. This approach models the given dialogue as three graphs (V-A, V-T, A-T) to capture the contextual information and complementary information between modalities.

In the current research, the problem with the utilized multimodal feature fusion algorithms is that the interactions among heterogeneous information from different modes must be more fully considered, and it is not sufficient to reflect real emotions under conflicts and differences between modal features. In response to this problem, researchers have explored different fusion strategies. For example, they proposed weighing different modalities while considering the importance differences among multimodal features and assigning weights via importance attention networks. Despite these efforts, multimodal ERC research still needs to be improved, and further consideration needs to be given to the existence of certain complementary and differential information in different modalities and the combination of contextual interactions. Aspects such as speaker dependence and dialogue context are also understudied in multimodal settings. In addition, in multimodal ERC tasks, multimodal connection fusion methods exist in various forms but cannot solve the modal conflict problem effectively. Therefore, when different modalities conflict, the fused modalities of the existing models interfere with each other, producing inaccurate results. Therefore, the current research focusing on multimodal fusion methods should consider how to use the relationships between modalities to eliminate modal conflicts, enabling the constructed models to describe better multimodal fusion features for recognizing emotions in dialogues.

To solve the above-mentioned problems, additional research can delve into the following avenues. First, more intricate and potent modality fusion strategies can be contemplated, ensuring the comprehensive incorporation of the variances between different modalities throughout the feature fusion phase. Second, it is imperative to underscore the pivotal role of contextual cues in enhancing the accuracy of multimodal emotion recognition methods during conversational interactions. This necessitates a comprehensive exploration

of strategies for optimizing the utilization of context and speaker-specific dependencies. In addition, introducing more advanced model structures handles multimodal conflicts. Improved multimodal fusion can provide more accurate and effective solutions for tasks such as emotion recognition. In future research, these directions should be further explored, and the existing methods should be continuously improved to advance the field of ERC.

6. Applications of ERC

As a pivotal undertaking within NLP, emotion recognition in conversations has found extensive applications across various domains, notably exemplified by the deployment of intelligent service robots. These sophisticated automatons discern users' emotional states by scrutinizing their ongoing and historical conversational interactions, subsequently adapting their response strategies to enhance user experiences. Recent years have witnessed the pervasive integration of multimodal information transmission. Consequently, multimodal data have emerged as the predominant mode of communication. For instance, within a multimodal context, an intelligent customer service robot orchestrates a comprehensive analysis of user emotions by assimilating facial features, textual inputs, and vocal cues, subsequently implementing context-sensitive policy adjustments. Below, we delineate a few illustrative instances of emotion recognition in conversational application domains.

6.1. Human-Computer Interaction

ERC plays a crucial role in improving human-computer interaction experiences. A machine can perceive and understand a user's emotional state and changes at different points during their dialogue to better respond to user needs and provide personalized services or answers. In virtual assistants, emotion-driven games, and emotion-oriented chatbots, sentiment analysis can help machines respond and make decisions based on the user's emotional state. In recent related research, Rashkin et al. [63]) introduced a novel dialogue dataset grounded in emotional contexts, comprising 25,000 scenes. Empirical experiments have demonstrated that a model trained on this dataset receives higher empathy ratings from human evaluators, thus enhancing user experiences in human-computer interactions. These studies aim to further augment the empathetic expressions generated by the model by leveraging user context and emotional information derived from the conversation.

ERC, combined with natural language generation technology, enables machines to communicate and express in an emotionally rich manner, enhancing the emotional interaction and communication capabilities between machines and humans. ERC has also applied to customer service and support. Businesses can analyze customer conversations with service representatives to identify a customer's emotional state, increase customer satisfaction and service quality, and resolve potential issues on time.

6.2. Mental Health Assessment

In recent years, dialogues that provide emotional support have gradually emerged with the development and refinement of theories related to emotional support.ERC can be used to assist with mental health assessments. Analyzing an individual's language and conversations can assess their emotional and mental health, which helps identify potential depression, anxiety, or other mental health issues. However, the need for well-designed tasks and suitable corpora hinders the related research progress. To this end, Liu et al. [64] defined the ESC (Emotional Support Conversation) task and proposed an ESC framework based on the theory of helping skills. Comforting and giving advice do not need to be performed in sequence. In addition, the author also constructed a high-quality ESC dataset ESConv with rich annotations and demonstrated the role of the ESConv dataset in training more emotional support systems through related experiments. Tu et al. [65] proposed a new emotional support method, MISC, which integrates COMET into emotional support conversations and uses an attention mechanism to learn from the obtained knowledge selectively, grasps the user's emotions and changes in emotional support dialogue.

6.3. Social Science Research

ERC is also of great significance in social science research. Analyzing social media data, online comments, and questionnaires reveals the expressions of and changes in human emotions in different social situations and events. This result helps researchers understand the evolution processes and factors influencing human emotional and psychological states. For example, mining user information on social media platforms such as Twitter and Facebook to obtain health- and drug-related information has sparked great interest in pharmacovigilance research. Such social media speeches can be analyzed to detect drug abuse, use responses, and drug-related emotional expressions. Most of these studies are based on aggregate results obtained for large populations rather than specific individuals. To research individuals or specific groups of people, Mahata et al. [66] designed a CNN-based model to identify personal drug intake comments mentioned in user conversations in tweets or blogs, thereby tracking the expression of information about the user's mood and other aspects of drug intake on social media.

7. Challenges

ERC research has made relatively strong progress. Deep learning models have demonstrated significant promise in extracting emotional features, capturing conversation dynamics, and modeling contextual factors, speaker characteristics, and emotional transitions. Nevertheless, numerous challenges persist despite these accomplishments.

7.1. ERC in Real Scenarios

ERC research has made significant progress, and some challenges have been successfully addressed. However, emotion recognition in realistic real-time human-computer interaction scenarios still faces additional challenges. As shown in Figure 9, at first, multimodal ERC data are obtained from capturing and collecting data in the multi-sensory world of human beings. The video angle, volume, and light changes observed in natural scenes all impact the recognition ability of an algorithm. The training data may also need to cover different regional dialects used in conversations. Problems such as losses, damage, and delays occur during data transmission. For multimodal scenarios, ensuring the complete transmission of the three data modes and delivering them to the model without delay is an enormous challenge. In addition, it is usually necessary to consider the ERC task in a real-time environment. It is impossible to accurately understand the complete speaker relationship and the impacts of future utterances, so the context relationship and speaker modeling process can only rely on previous iterations, which limits the constructed algorithm's capability to a certain extent.

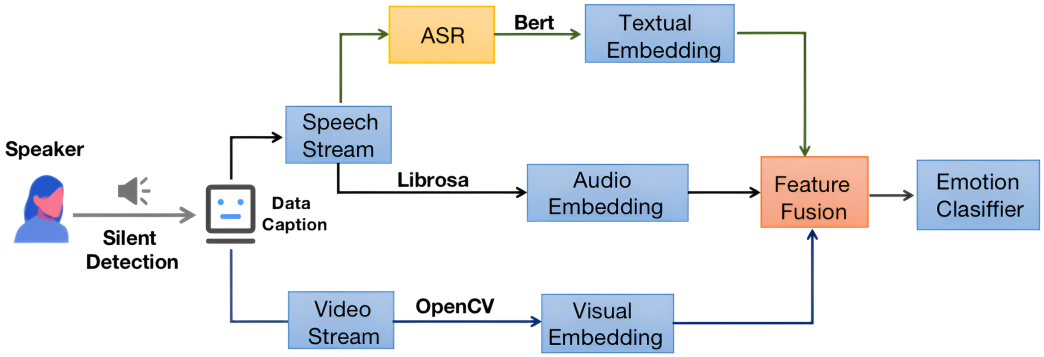

Figure 9. A Diagram Describing the ERC Flow in a Real-World Scenario.

To address the above challenge, it is crucial to leverage past conversation models better to model the correlation of emotions in future conversations. Furthermore, clearly defining roles and purposes in conversations, especially when dealing with unclear dynamics and difficult-to-determine speaker relationships, can better alleviate the limitations of natural ERC-constrained environments on the above issues.

7.2. Latency Due to Data Processing

ERC tasks can be divided into multimodal and plain text ERC. In terms of transmission speed, the speed of plain text can satisfy industrial users' needs. However, audio and video transmission induce considerable delays on the network channel, so they cannot be used in actual industrial cases. In real usage scenarios, the industrial transmission process typically uses video and audio streams to transmit and assemble data frames and perform feature extraction. Traditional encoder-based methods are no longer applicable in this case. In addition, the feature extraction of multimodal conversations is also time-consuming, especially with the traditional video feature extraction method. We used the feature extraction tools commonly employed by researchers in ERC tasks to evaluate the time required for feature extraction. We used pre-trained BERT for text, LibROSA for audio, and OpenFace for videos. We divided the videos in the MELD dataset into different experimental groups according to their lengths and performed feature extraction. During this process, we recorded each group's time spent on feature extraction and analyzed the results in detail, as shown in Table 8. According to the statistical results shown in the table, it can be observed that video feature extraction takes the longest time, and other modalities also have delays of more than one second. Such delays may have severe adverse effects on scenarios that require a high-quality human-computer interaction experience.

Therefore, when dealing with the transmission of longer video and audio data, we need to employ more flexible mechanisms to accommodate the transfer of smaller data units and utilize the features from these smaller units to predict emotions in the conversation. This approach helps address the delays caused by data transmission and processing and overcomes some limitations arising from incomplete contexts.

Table 8. Time-of-day statistics for the feature extraction process of multimodal dialogue emotion recognition in natural scenes.

Video Duration (Time)	Text Feature	Vision Feature	Audio Feature	Total Time
1 s	1.8 s	2.2 s	1.3 s	5.3 s
4 s	1.8 s	6.05 s	1.4 s	9.25 s
8 s	1.8 s	10.4 s	1.6 s	13.8 s

7.3. Classification of an Imbalanced Dataset

The current category labels for emotion recognition exhibit an unbalanced classification distribution, with significant variations among the number of samples in different emotion categories. However, this unbalanced classification task displays characteristics that require enhancement, particularly in its inherent bias towards the dominant category. In an unbalanced training set, the class with a large sample size significantly impacts the model training process, making it more inclined to predict the dominant category. Fewer emotional categories give the model a more vital ability to identify rare emotional categories with less training data. It is easy to confuse fused emotions.

Therefore, in the work related to dataset construction, it is necessary to focus more on enriching the data for emotional categories with fewer samples to narrow this gap. Although recent research has proposed some methods to deal with partially missing modalities in sentiment analysis tasks [67,68], a key challenge remains in the field of general ERC, namely, the lack of a flexible and broadly applicable framework and paradigm for effectively dealing with missing and emerging modalities. In cases with missing modalities, the existing methods mainly focus on using information from other available modalities to compensate for the impact of the missing modalities, thus maintaining ERC accuracy. In addition, some methods enable information transfer between different modalities by learning shared representations across the modalities to improve the robustness of the model to the missing modalities. However, more flexible and broadly applicable frameworks and paradigms are still needed to deal with missing modalities. Such a framework should be

able to adapt to different situations and various types of missing modalities and automatically adapt to emerging concerns. Further research can explore domain knowledge and transfer learning techniques to deal with missing modalities and develop general methods and algorithms that apply to different tasks and application scenarios.

In summary, while various methods have been proposed to address ERC challenges in the context of partially missing modalities, there is still a need for a flexible and broadly applicable framework and paradigm to handle both missing and emerging modalities. Moreover, it is essential to develop general methods and algorithms tailored to different emotion recognition tasks and application scenarios.

8. Conclusions and Future Research Ideas

Emotions are essential in understanding human research by machines, an essential component of cognitive behavior. For a machine to truly understand humans, it must be able to recognize, understand, and respond to human expressions of emotion. With the continuous development and in-depth study of artificial intelligence technology, the influence of emotion in AI research will continue to expand, especially in fields such as AIGC, which require machines to possess enhanced cognitive abilities to achieve higher cognitive functionality. Therefore, this paper provides a comprehensive review of the recent research progress, focusing on the research motivations and experimental effects of context modeling, speaker dependency considerations, and related knowledge integration to understand the significance of these studies better.

Furthermore, we have synthesized methods for extracting features from multimodal data and conducted an in-depth exploration of relevant datasets in this research domain. We have also identified the challenges currently faced by researchers, such as issues in real-time contexts and imbalances in data, and have proposed potential solutions to these problems. By providing researchers with a comprehensive perspective and valuable references, we aim to facilitate the ongoing advancement of the ERC task, ultimately enriching human-computer interaction.

Author Contributions: Conceptualization, C.Z.; Writing—original draft, Y.F.; validation, S.Y.; writing—review and editing, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the National Key R&D Program of China under grant 2020AAA0108701 and in part by "the Fundamental Research Funds for the Central Universities" (CUC21GZ014).

Data Availability Statement: Data are contained within the article.

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

Abbreviations

The following abbreviations are used in this manuscript:

ERC	Emotion recognition in conversation
GNN	Graph Neural Network
GRU	Gate Recurrent Unit
ESC	Emotional Support Conversation

References

1. Scarselli, F.; Gori, M.; Tsoi, A.C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Trans. Neural Netw.* **2009**, *20*, 61–80. [CrossRef] [PubMed]
2. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
3. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.

4. Ekman, P.; Sorenson, E.R.; Friesen, W.V. Pan-Cultural Elements in Facial Displays of Emotion. *Science* **1969**, *164*, 86–88. [CrossRef] [PubMed]
5. Bakker, I.; van der Voordt, T.; Vink, P.; de Boon, J. Pleasure, Arousal, Dominance: Mehrabian and Russell revisited. *Curr. Psychol.* **2014**, *33*, 405–421. [CrossRef]
6. Plutchik, R. *Emotions and Life: Perspectives from Psychology, Biology, and Evolution*; American Psychological Association: Washington, DC, USA, 2003.
7. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [CrossRef]
8. Mehrabian, A.; Russell, J.A. *An Approach to Environmental Psychology*; The MIT Press: Cambridge, MA, USA, 1974.
9. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
10. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. DailyDialog: A manually labeled multi-turn dialogue dataset. *arXiv* **2017**, arXiv:1710.03957.
11. Zahiri, S.M.; Choi, J.D. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. *arXiv* **2017**, arXiv:1708.04299.
12. Chen, S.Y.; Hsu, C.C.; Kuo, C.C.; Ku, L.W. EmotionLines: An emotion corpus of multi-party conversations. *arXiv* **2018**, arXiv:1802.08379.
13. Chatterjee, A.; Narahari, K.N.; Joshi, M.; Agrawal, P. SemEval-2019 Task 3: EmoContext contextual emotion detection in text. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 39–48.
14. Zhao, J.; Zhang, T.; Hu, J.; Liu, Y.; Jin, Q.; Wang, X.; Li, H. M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 5699–5710.
15. Liang, C.; Yang, C.; Xu, J.; Huang, J.; Wang, Y.; Dong, Y. S+ page: A Speaker and Position-Aware Graph Neural Network Model for Emotion Recognition in Conversation. *arXiv* **2021**, arXiv:2112.12389.
16. Zhang, Y.; Jin, R.; Zhou, Z.H. Understanding bag-of-words model: A statistical framework. *Int. J. Mach. Learn. Cybern.* **2010**, *1*, 43–52. [CrossRef]
17. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
18. Mikolov, T.; Sutskever, I.; Chen, K.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 27th Annual Conference on Neural Information Processing Systems 2013, Lake Tahoe, NV, USA, 5–10 December 2013; Volume 26.
19. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
20. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
21. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
22. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
23. Amos, B.; Ludwiczuk, B.; Satyanarayanan, M. OpenFace: A general-purpose face recognition library with mobile applications. *CMU Sch. Comput. Sci.* **2016**, *6*, 20.
24. Zhu, Q.; Yeh, M.-C.; Cheng, K.-T.; Avidan, S. Fast human detection using a cascade of histograms of oriented gradients. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1491–1498.
25. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. Librosa: Audio and music signal analysis in Python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; Volume 8, pp. 18–25.
26. Eyben, F.; Wöllmer, M.; Schuller, B. OpenSMILE: The Munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 29 October 2010; pp. 1459–1462.
27. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; Scherer, S. COVAREP—A collaborative voice analysis repository for speech technologies. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 960–964.
28. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent neural network regularization. *arXiv* **2014**, arXiv:1409.2329.
29. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734.
30. Wöllmer, M.; Metallinou, A.; Eyben, F.; Schuller, B.; Narayanan, S.S. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. *Interspeech* **2010**. Available online: https://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/deliver/index/docId/76287/file/wollmer10c_interspeech.pdf (accessed on 15 November 2023).

31. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. Dialoguernn: An attentive RNN for emotion detection in conversations. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6818–6825.
32. Li, J.; Ji, D.; Li, F.; Zhang, M.; Liu, Y. HiTrans: A Transformer-Based Context- and Speaker-Sensitive Model for Emotion Detection in Conversations. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 4190–4200.
33. Kim, T.; Vossen, P. EmoBERTa: Speaker-aware emotion recognition in conversation with RoBERTa. *arXiv* **2021**, arXiv:2108.12009.
34. Sedoc, J.; Gallier, J.; Foster, D.; Ungar, L. Semantic Word Clusters Using Signed Spectral Clustering. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, 30 July–4 August 2017.
35. Ghosal, D.; Akhtar, M.S.; Chauhan, D.; Poria, S.; Ekbal, A.; Bhattacharyya, P. Contextual Inter-Modal Attention for Multi-Modal Sentiment Analysis. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3454–3466.
36. Kiela, D.; Bhooshan, S.; Firooz, H.; Perez, E.; Testuggine, D. Supervised Multimodal Bitransformers for Classifying Images and Text. *arXiv* **2019**, arXiv:1909.02950.
37. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
38. 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.
39. Hu, J.; Liu, Y.; Zhao, J.; Jin, Q. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. *arXiv* **2021**, arXiv:2107.06779.
40. Shen, W.; Wu, S.; Yang, Y.; Quan, X. Directed acyclic graph network for conversational emotion recognition. *arXiv* **2021**, arXiv:2105.12907.
41. Xu, S.; Jia, Y.; Niu, C.; Zan, H. MMDAG: Multimodal Directed Acyclic Graph Network for Emotion Recognition in Conversation. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 13 June 2022; pp. 6802–6807.
42. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
43. Brody, S.; Alon, U.; Yahav, E. How Attentive Are Graph Attention Networks? *arXiv* **2021**, arXiv:2105.14491.
44. Zhang, D.; Wu, L.; Sun, C.; Li, S.; Zhu, Q.; Zhou, G. Modeling Both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-Speaker Conversations. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 10–16.
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J. Attention is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
46. Mao, Y.; Sun, Q.; Liu, G.; Wang, X.; Gao, W.; Li, X.; Shen, J. DialogueTRM: Exploring the Intra- and Inter-Modal Emotional Behaviors in the Conversation. *arXiv* **2020**, arXiv:2010.07637.
47. Li, Z.; Tang, F.; Zhao, M.; Zhu, Y. Emocaps: Emotion capsule based model for conversational emotion recognition. *arXiv* **2022**, arXiv:2203.13504.
48. Zou, S.; Huang, X.; Shen, X.; Liu, H. Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation. *Knowl.-Based Syst.* **2022**, *258*, 109978. [CrossRef]
49. Zou, S.; Huang, X.; Shen, X.; Liu, H. Hierarchical Dialogue Understanding with Special Tokens and Turn-level Attention. *arXiv* **2023**, arXiv:2305.00262.
50. Zhang, C.; Song, D.; Huang, C.; Swami, A.; Chawla, N.V. Heterogeneous Graph Neural Network. *arXiv* **2020**, arXiv:2003.02102.
51. Lian, Z.; Liu, B.; Tao, J. CTNet: Conversational Transformer Network for Emotion Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 985–1000. [CrossRef]
52. Hu, D.; Hou, X.; Wei, L.; Jiang, L.; Mo, Y. MM-DFN: Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022.
53. Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; Zimmermann, R. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LO, USA, 1–6 June 2018; pp. 2122–2132.
54. Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; Zimmermann, R. Icon: Interactive Conversational Memory Network for Multimodal Emotion Detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2594–2604.
55. Shenoy, A.; Sardana, A. Multilogue-net: A Context-Aware RNN for Multi-Modal Emotion Detection and Sentiment Analysis in Conversation. *arXiv* **2020**, arXiv:2002.08267.
56. Zhao, W.; Zhao, Y.; Qin, B. MuCDN: Mutual Conversational Detachment Network for Emotion Recognition in Multi-Party Conversations. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 7020–7030.
57. Bao, Y.; Ma, Q.; Wei, L.; Zhou, W.; Hu, S. Speaker-guided Encoder-Decoder Framework for Emotion Recognition in Conversation. *arXiv* **2022**, arXiv:2206.03173.

58. Joshi, A.; Bhat, A.; Jain, A.; Singh, A.V.; Modi, A. COGMEN: COntextualized GNN based Multimodal Emotion recognition. *arXiv* **2022**, arXiv:2205.02455.
59. Schlichtkrull, M.; Kipf, T.N.; Bloem, P.; Van Den Berg, R.; Titov, I.; Welling, M. Modeling Relational Data with Graph Convolutional Networks. In Proceedings of the Semantic Web: 15th International Conference, Crete, Greece, 3–7 June 2018.
60. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph Transformer Networks. In Proceedings of the NeurIPS, Montreal, QC, Canada, 2–8 December 2018.
61. Chen, M.; Wang, S.; Liang, P.P.; Baltrušaitis, T.; Zadeh, A.; Morency, L.-P. Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning. In Proceedings of the ICMI '17: Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017.
62. Li, J.; Wang, X.; Lv, G.; Wang, X.; Lv, G.; Zeng, Z. GraphMFT: A Graph Network Based Multimodal Fusion Technique for Emotion Recognition in Conversation. *Neurocomputing* **2023**, *550*, 126427. [CrossRef]
63. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv* **2018**, arXiv:1811.00207.
64. Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; Yu, Z.; Huang, M. Towards emotional support dialog systems. *arXiv* **2021**, arXiv:2106.01144.
65. Tu, Q.; Li, Y.; Cui, J.; Wang, B.; Wen, J.R.; Yan, R. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. *arXiv* **2022**, arXiv:2203.13560.
66. Mahata, D.; Friedrichs, J.; Shah, R.R.; Jiang, J. Detecting personal intake of medicine from Twitter. *IEEE Intell. Syst.* **2018**, *33*, 87–95. [CrossRef]
67. Zhao, J.; Li, R.; Jin, Q. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. 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), Virtual Event, 1–6 August 2021; pp. 2608–2618.
68. Wang, N.; Cao, H.; Zhao, J.; Chen, R.; Yan, D.; Zhang, J. M2R2: Missing-Modality Robust emotion Recognition framework with iterative data augmentation. *IEEE Trans. Artif. Intell.* **2022**, *4*, 1305–1316. [CrossRef]