GEORGIA STATE UNIVERSITY
Department of Mathematics and Statistics

Deciphering Emotions: A Survey of Conversation-based Recognition Methods

By

Jesse Annan

December 2023

ABSTRACT

Emotion Recognition in Conversation (ERC) is an increasingly popular, yet unresolved, and challenging task in Natural Language Processing. Its aim is to identify the emotional states of speakers engaged in dialogue. ERC has various applications in human-computer interaction, social media analysis, mental health assessment, and affective computing. Solving the ERC task is a necessary step towards creating an empathetic Autonomous Digital Human (ADH). An ADH is an entity finely attuned to emotions, enhancing its ability to engage users in a profoundly natural and empathetic manner. In this survey, we review recent advances in ERC methods, particularly focusing on multimodal approaches that utilize various information sources like text, audio, and visual cues. We categorize existing methods into two types: RNN-based and transformer-based, then discuss their respective advantages and disadvantages. Additionally, we compare their performance using two widely-used benchmark datasets: IEMOCAP (dyadic) and MELD (dialogue). Furthermore, we present the primary challenges and open issues in ERC, highlighting potential directions for future research.

Keywords: *Natural Language Processing, Emotion Recognition, Autonomous Digital Human, Conversation Analysis, Dialogue Processing, Transformer, RNN*

Contents

List of Abbreviations

ERC Emotion Recognition in Conversation

ADH Autonomous Digital Human

IEMOCAP Interactive Emotional Dyadic Motion Capture

MELD Multimodal EmotionLines Dataset

CNN Convolutional Neural Network

CMT Cross-Model Transformer

MTCNN Multi-task Cascade Convolutional Network

MGIF Multi-Grained Interactive Fusion

SWFC Sample-Weighted Focal Constructive

Chapter 1

INTRODUCTION

Introduction

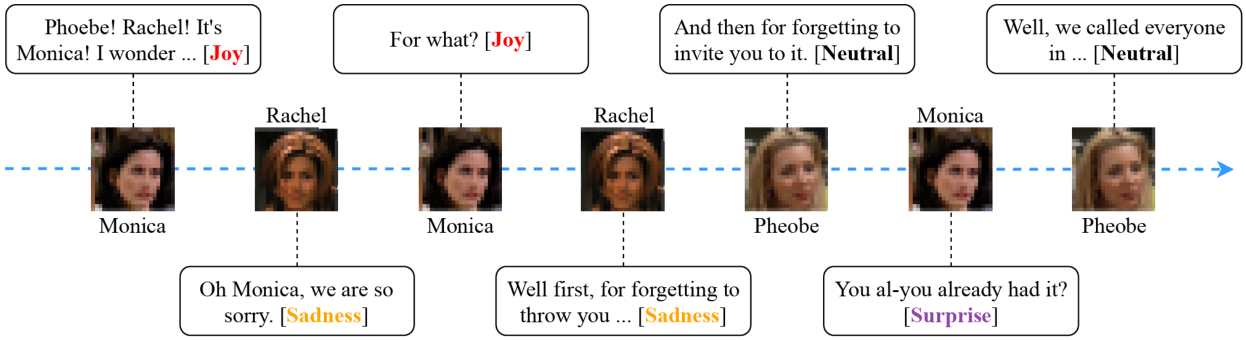
Phoebe! Rachel! It's Monica! I wonder ... [Joy]
For what? [Joy]
And then for forgetting to invite you to it. [Neutral]
Well, we called everyone in ... [Neutral]
Rachel
Rachel
Monica
Monica
Monica
Pheobe
Pheobe
Oh Monica, we are so sorry. [Sadness]
Well first, for forgetting to throw you ... [Sadness]
You al-you already had it? [Surprise]

Figure 1.1: An example of Emotion Recognition in Conversations. It also depicts emotional shift of speakers in a dialogue in comparison with their previous emotions. *(Source: [20])*

Humans perceive and understand each other through many channels of communication, whether conveyed through text, audio, or visual cues. Each individual channel has a unique advantage in expressing or communicating the emotional state of a speaker. To understand emotions on a theoretical level, research on emotions has been studied as far back as 1872 [2]. Charles Robert Darwin shared his research on the similarity of emotional cues shared by humans and animals in his book "The Expression of the Emotions in Man and Animals." This book concerns the biological aspects of emotional behavior [8]. Paul Ekman [13], a psychologist and pioneer in the study of emotions, laid the groundwork by identifying and categorizing universal emotional expressions, illuminating the connection with facial expressions. Ekman identified the six basic emotions as *anger, surprise, disgust, enjoyment, fear,* and *sadness*. Beyond facial expressions, the intonation of sound within speech has emerged as a pivotal aspect of emotion recognition. It refers to the variation of pitch, intensity, and duration of speech sounds. The intonation of sound can convey different emotions, such as happiness, sadness, anger, fear, surprise, and disgust [21]. More recent research has found that small changes in sound can affect human emotions as much as shifts in the tone of a person's voice [4]. This era of increased interactions with devices to perform any task at all has led to the development of an efficient ERC. Its aim is to identify the emotional states of speakers engaged in dialogue. ERC has various applications in human-computer interaction, social media analysis, mental health assessment, and affective computing. One of the core aspirations in artificial intelligence is to develop intelligent systems or an empathetic ADH that can effectively follow multi-modal instructions aligned with human intent to complete various real-world tasks in the wild [3, 22]. ERC is an increasingly popular and challenging task in NLP that aims to identify the emotional state of speakers engaged in dialogue and hence pave a way to create an ADH. Recent works on ERC present several solutions with multi-modal datasets and challenges such as conversational context modeling, emotional shifts of the interlocutors, and others. In this survey, we review the recent advances in ERC methods, focusing on the multi-modal approaches that leverage different sources of information, such as text, audio, and visual cues. We categorize the existing methods into two types: RNN-based, and transformer-based methods. We discuss the advantages and disadvantages of each type and compare their performance on several benchmark datasets. We also present the main challenges and open issues in ERC and suggest some possible directions for future research. The methods we cover in this survey are: CMN, DialogueRNN, DialogueTRM, DialogXL, EmoBERTa, Hi-Trans, M2FNet, FacialMMT, and MultiEMO. These methods are selected based on their novelty, popularity, and effectiveness in ERC. We provide a brief overview of each method and highlight their key contributions and limitations. We also provide a table that summarizes the main characteristics and results of each method.

Chapter 2

CORPORA EXPLORATION

2.1 Understanding Diverse Corpora for Emotion Recognition

The pursuit of creating an empathetic system capable of recognizing emotional cues to enhance its performance in specified tasks has motivated the assembly of realistic emotionally annotated datasets. Two prominent corpora for the ERC task are the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [6] and the Multimodal EmotionLines Dataset (MELD) [29], an extension of the earlier dyadic datasets, EmotionLines [9]. IEMOCAP emphasizes the significance of multi-modal data in understanding emotions, given that emotions are conveyed through both verbal and non-verbal channels. This benchmark comprises approximately 12 hours of recorded video involving ten actors (five females, five males) engaged in dyadic sessions. These sessions encompass facial expressions and motion capture during spontaneous and scripted communication. The emotional categories in IEMOCAP include anger, happiness, sadness, neutral, excitement, and frustration. Additionally, it provides continuous emotions such as activation, valence, and dominance. MELD introduces a multi-party conversation setting that is more challenging to classify than the dyadic variants available in previous datasets [6, 9, 26]. Each utterance in MELD is annotated with emotion and sentiment labels, covering audio, visual, and textual modalities. MELD extends EmotionLines by annotating emotions across multiple modalities (text, audio, and visual) and accurately tracking the timestamps of utterances within the same episodes (see Table 2.2). EmotionLines, which solely focused on textual data, in some instances sampled utterances from different episodes, potentially introducing noise due to the lack of contextual relevance between previous and current utterances or emotional states of the interlocutors(see Table 2.3). SEMAINE [26] and AVEC [33] represent two multi-modal datasets widely utilized in ERC. The SEMAINE dataset involves interactions between a human and an operator (either a machine or a person simulating a machine). AVEC, a subset of the SEMAINE, is an ongoing series of challenges centered around recognizing continuous emotional states like arousal, valence, and dominance. There are additional multi-modal emotion and sentiment analysis datasets, as highlighted in [29], such as MOUD [27] and MOSI [41]. However, these datasets contain individual narratives rather than dialogues. Uni-modal (textual) dataset such as ALM [1] and ISEAR [31] are also two favoured benchmarks for ERC in text [2]. ALM entails sentence-level annotation labels for around 185 children's stories, including those from Grimm and Potter. In contrast, ISEAR is a dataset formulated by a group of psychologists, encompassing reported situations where students experienced emotions like joy, fear, anger, sadness, disgust, shame, and guilt.

2.2 Multi-modal Corpora Comparison

Dataset	Type	# dialogues			# utterances		
		train	dev	test	train	dev	test
IEMOCAP [6]	acted	120		31	5810		1623
SEMAINE [26]	acted	58		22	4386		1430
MELD [29]	acted	1039	114	280	9989	1109	2610

Table 2.1: Comparison among IEMOCAP, SEMAINE, and MELD datasents. The utterances in MELD is nearly double the size of the other datasets. *(Source: [29])*

Utterance	S	E	Incorrect Splits		Corrected Splits	
			Start Time	End Time	Start Time	End Time
Chris says they're closing down the bar	3	6	00:05:57,023	00:05:59,691	00:05:57,023	00:05:58,734
No way!	3	6	00:05:57,023	00:05:59,691	00:05:58,734	00:05:59,691

Table 2.2: Example of timestamp alignment using Gentle alignment tool. *(Source: [29])*

S	E	Utterance	Speaker	Emotion	Sentiment
6	4	What are you talking about? I never left you! You've always been my agent!	Joey	surprise	negative
		Really?!	Estelle	surprise	positive
		Yeah!	Joey	joy	positive
		Oh well, no harm, no foul.	Estelle	neutral	neutral
5	20	Okay, you guys free tonight?	Gray	neutral	neutral
		Yeah!!	Ross	joy	positive
		Tonight? You-you didn't say it was going to be at nighttime.	Chandler	surprise	negative

Table 2.3: A dialogue in EmotionLines where utterances from two different episodes are present. *(Source: [29])*

Chapter 3

METHODS OF EMOTION RECOGNITION IN CONVERSATION

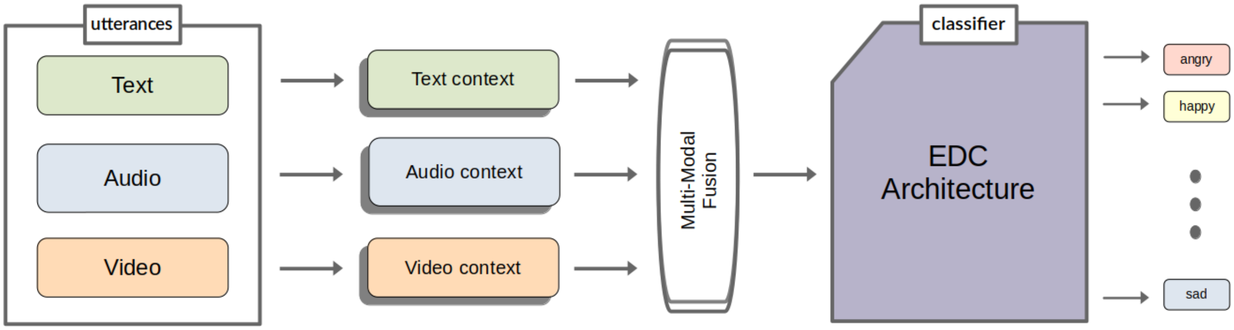

Figure 3.1: Condensed Structure of an Emotion Recognition in Conversation Framework

3.1 Techniques for Uni-modal Feature Extraction in Emotion Recognition

3.1.1 Textual Feature Extraction

As presented in [19], Kim offers a technique for preparing textual data for various classification tasks like sentiment analysis and determining text subjectivity. This method involves text classification by utilizing the softmax probability derived from key text features. These key features are acquired by initially concatenating vector representations of the words constituting the given text. Subsequently, a convolution operator, functioning as a filter, is applied to generate feature maps. Finally, a max pooling layer is employed to acquire the most important feature map. This innovative approach has been adopted by [15, 24, 28] for textual feature representation. Specifically, in the extraction of textual modality from dyadic conversations [6], Convolutional Neural Network (CNN) with sizes 3, 4, and 5, each having 50 feature maps, are utilized. These are then subjected to max-pooling with a window of size 2 and passed through a fully connected layer with 100 neurons to represent the textual modality. Both Mao et al. [25] and Li et al. [20] leverage BERT's [12] exceptional representation learning abilities along with the vanilla transformer [39] for extracting contextual representations. These transformer-based models are structured hierarchically in a two-stage manner. Initially, BERT serves as the low-level transformer for textual representation. Subsequently, its final representation is parsed to the vanilla transformer as the second stage, transforming the textual representation into a more contextual representation of the utterances. However, Mao et al. [25] notes that their hierarchical structure for textual representation takes into account the inter-speaker information within the textual representation. DialogXL, a novel architecture introduced by Shen et al. [34], Chudasama et al. [10], Shi et al. [35], and Kim et al. [18], follow the trend of text representation through transformer-based models to capture contextual utterance representation. In particular, [10], [35], and [18] generated deeper inter-utterance context by parsing the textual modality through the RoBERTa model proposed by Kim et al. [23]. Similarly, Zheng et al. [43], employing BERT, a transformer-based model, effectively utilizes dialogue context and speaker emotional dynamics. It does so by initially concatenating the input utterance and all its contextual utterances as input, selecting the hidden representation generated by the first token as the textual representation.

3.1.2 Audio Feature Extraction

In [15, 24, 25, 28, 35], audio features get extracted through OpenSMILE [14], an open-source software automatically deriving audio features like pitch and voice intensity. The audio vector standardizes and reduces dimensions to 100 via a fully connected neural layer from the initial 6373 produced by OpenSMILE software. Similarly, Shi et al. [35] extracts contextualized audio features of 256 dimensions using [24]. However, Zheng et al. [43] acquires word-level audio representation based on the Wav2Vec2.0 model [5], with dimensions of 768. On the other hand, Chudasama et al. [10] introduces a novel feature extractor model. The proposed model, employing the standard ResNet18 [16] and inspired by the triplet network to emphasize the importance of the triplet loss function [32], processes audio signals utilizing warping and additive white Gaussian noise, resulting in Mel Spectrograms, and uses the the Mel Spectrograms to generate contextualized features for the audio signals.

3.1.3 Visual Feature Extraction

3D-CNN has been utilized by researchers [15, 24, 25, 28] to capture facial expressions and the visual environment in visual data. As highlighted by Tran et al. [37], the 3D-CNN not only extracts relevant features from individual image frames but also captures spatiotemporal features across frames, facilitating the identification of emotional expressions such as smiles or frowns. However, while the 3D-CNN excels in achieving state-of-the-art results in object classification in three-dimensional data [17], it presents challenges in visual feature extraction. Additionally, research by Shi et al. [35] emphasizes that gathering redundant surrounding information for each utterance might lead to misinterpretation of the speaker's genuine emotional tendencies due to the influence of irrelevant scene information. To address this, they propose the VisExtNet framework. This framework comprises a Multi-task Cascade Convolutional Network (MTCNN) [42] for precise facial feature extraction and utilizes ResNet-101 [16] pre-trained on VGGFace2 [7] to extract 100-dimensional emotion-rich cues of interlocutors. Finally, contextual information is extracted using DialogueRNN with 256 dimensions. Furthermore, Zheng et al. [43] also acknowledges the challenge of environmental noise and the existing frameworks' inability to distinguish the main speaker from others. To resolve these issues, Zheng et al. introduces a three-stage face sequence extraction method. Firstly, it distinguishes faces from the background by employing multi-modal rules and an active speaker detection model [36]. Then, it applies InfoMap [30], an unsupervised face clustering technique, to identify the number of face clusters in the sequence. Finally, face matching is conducted to determine the real speaker's face.

3.2 Fusion Strategies for Multi-modal Emotion Recognition

As described in [25], multi-modal fusion aims to generate a unified representation to enhance specific tasks involving multiple modalities, such as building classifiers or other predictors. This process generally falls into two categories: linear weighting fusion, which utilizes techniques like concatenation, bi-linear operations, or addition to combine different modalities, and interactive weighting fusion, which involves methods like differential operations, gating mechanisms, or attention mechanisms to fuse modalities at various sub-view granularity. In numerous studies, including [10, 24, 35], the approach commonly adopted for multi-modal features involves direct concatenation. As clearly explained in [43], to achieve interactions between different modalities, they apply the Cross-Model Transformer (CMT) layer [38]. Firstly, they fuse the text and audio modalities, alternating the two modalities as the query vector, then concatenating them to obtain the text-audio fused representation and in a similarly manner, the text-audio fused with visual modality to obtain the utterance-level text-audio-visual fused representation.

3.3 Challenges and Advances in Model Architectures

To classify utterance emotions, Hazarika et al. [15] employed temporal contextual information from past utterances by the speaker and involved parties. They filtered this information using an attention mechanism over a GRU generated memory representation of speaker and party history to extract meaningful summary of utterances. Building on this work, Majumder et al. [24] expanded the multi-modal extraction process. They acknowledged limitations highlighted in CMN [15] and hypothesized that an utterance's emotion depends on the main speaker, preceding context, and emotions exhibited by conversation participants. They proposed a model with an updated speaker state and employed attention scores to focus on relevant utterances. Utilizing a GRU, they jointly encoded utterances and speaker states. In a different approach, Li et al. [20] stacked two transformers for contextual representation: BERT for textual representation and a vanilla transformer for broader context. Their model's output is pre fed to their emotion detection and classification task and an auxiliary task determining if two utterances originated from the same speaker. This two-stage transformer encoding surpassed the results of DialogueRNN [24]. Another model, Shen et al. [34], drew ideas from XLNet [40] and Transformer-XL [11], utilizing memory of previous utterances for inter- and intra-speaker context. DialogXL employed four attention mechanisms: global, local, speaker, and listener self-attention, capturing utterance-level understanding by considering intra-speaker dependency and inter-speaker emotion dependencies. The speaker self-attention relies on (the memory of) previous utterances made by the speaker to capture intra-speaker dependency, where as, the listener self-attention relies on previous utterances by other speakers to capture inter-speaker emotion dependencies. Kim et al. [18] used RoBERTa to extract contextual information in emotional conversations. By inserting speaker names before each utterance and segmenting utterances, they enabled RoBERTa to classify based on inter- and intra-speaker states. This simple architecture proved effective due to awareness of speaker effects on listeners. Mao et al. [25] adopted a hierarchical structure similar to Hi-Trans [20], employing BERT for multi-modal representation and a vanilla transformer to encode inter-speaker information from textual, visual, and acoustic modalities asynchronously. Their proposed fusion method, Multi-Grained Interactive Fusion (MGIF), contributed to improved performance. Focusing on proper visual representations without external noise, Zheng et al. [43] introduced a framework distinguishing faces from backgrounds, conducting unsupervised face clustering, and applying face matching to identify the main speaker. They captured emotions in facial clusters with a computer vision algorithm and applied a softmax operation to concatenated textual, acoustic, and visual representations for emotion classification. Finally, Shi et al. [35] introduced MultiAttn, a novel fusion method based on bidirectional multi-head cross-attention, effectively integrating multi-modal information. They also proposed the Sample-Weighted Focal Constructive (SWFC) loss to alleviate the difficulty of classifying minority and semantically similar emotions. The influence of their loss function is depicted in Table 3.1.

	Model	Year	Fusion	Benchmarks	
				IEMOCAP [6]	MELD [29]
text-only	CNN [19]	2014		48.18	55.02
	CMN [15]	2018		56.13	-
	DialogueRNN [24]	2018		62.75	57.03
	Hi-Trans [20]	2020		64.50	61.94
	DialogXL [34]	2020		65.94	62.41
	EmoBERTa [18]	2020		**67.42**	65.61
	M2FNet [10]	2022		66.20	**66.23**
	MultiEMO [35]	2023		64.48	61.23
multi-modal	BC-LSTM [28]	2017	Concat	56.19	56.32
	DialogueTRM [25]	2020	MGIF [25]	69.23	63.55
	M2FNet [10]	2022	Concat	69.89	66.71
	FacialMMT [43]	2023	CMT [38]	-	66.58
	MultiEMO [35]	2023	Concat	**72.84**	**66.74**

Table 3.1: Quantitative (F1 weighted average score) comparison with text-only and multi-modal based on IEMOCAP and MELD benchmarks.

Chapter 4

SUMMARY

4.1 Strengths and Weaknesses of ERC Models

The CNN model [19] introduced by Yoon Kim possesses two notable strengths and has influenced papers such as DialogueRNN [24]. This model has the ability to automatically learn hierarchical representations of text, capturing both word-level and sentence-level features. Additionally, it utilizes filters to identify crucial phrases or word combinations contributing to classification tasks. However, it treats words as discrete entities and may face challenges in comprehending long-range dependencies crucial for certain text understanding tasks. Poria et al. [28] made one of the initial attempts to model ERC using multi-modal data. They consider the broader context of the video, such as tone of voice. However, incorporating surrounding scenes in emotion detection might potentially introduce uncertainty or ambiguity into the task. Hazarika et al. proposed the CMN model [15], effectively capturing temporal dependencies and contextual information through memory mechanisms. While it considers conversational history to understand emotions within current dialogues, its memory mechanism is limited and struggles with longer sequences. Additionally, there's a lack of information on how the memory network represents and utilizes context in dialogues. This led to the development of DialogueRNN (Majumder et al.), one of the most cited papers in ERC research. DialogueRNN incorporates an attention mechanism to focus on relevant parts of the conversation and considers the temporal context of the conversation. Despite performing well on multi-modal ERC tasks, DialogueRNN is ineffective in capturing long-range dependencies in conversations. The emergence of transformer-based models like Hi-Trans shares a common weakness in interpreting how the model makes predictions given an utterance. However, Hi-Trans shows sensitivity to different speakers within a conversation, enhancing the accuracy of detecting emotions. Shen et al. introduced DialogXL, stemming from ideas presented in XLNet and Transformer-XL. DialogXL's attention mechanisms cater to both intra-speaker (between speakers) and inter-speaker (within a speaker's history) dependencies in multi-party conversations. Its memory-based approach allows for a context-aware understanding of emotions within ongoing conversations. While novel, the incorporation of four attention mechanisms and a memory-based approach make the model complex. EmoBERTa, a simple RoBERTa-based model introduced by Kim et al., accounts for inter-speaker and intra-speaker dynamics and benefits from RoBERTa's pre-trained language representations.

It performs well even with limited labeled data but is constrained to textual context. The hierarchical structure of DialogueTRM [25] encodes both intra- and inter-speaker information and introduces a novel fusion mechanism, MGIF, for a more contextual representation of multi-modal features. Zheng et al.'s FacialMMT [43] model focuses on distinguishing surrounding noise from speakers and properly capturing facial expressions of the main speaker. However, for short sequences, facial features like mouth movement are hard-coded, making the model complex. Chudasama's M2FNet [10] introduces a new feature extractor using an adaptive margin-based triplet loss function to learn emotion-relevant features from audio and visual data. However, it doesn't consider the temporal dynamics of the conversation. The more recent MultiEMO [35] by Shi et al. captures relationships among different modalities using a multi-head attention mechanism and introduces the SWFC loss function, addressing challenges in classifying minority or semantically similar emotions. Yet, the complex fusion mechanisms might hinder the interpretability of the model's decision-making process.

Chapter 5

REMARKS

In this paper, we survey the best existing approaches for ERC. The trends revolve around two major architectures: RNN variants (such as GRU), providing the model with a memory of preceding utterances, and transformer-based models, enabling the modeling of inter- and intra-speaker dependencies in utterances. Models like MultiEMO [35] and DialogueTRM [25] that introduced novel fusion mechanisms showed performance improvements. Therefore, future research might benefit from exploring novel methods to maintain contextualized information in uni-modal features and examining the apparent correlation of context within different modalities [35]. The results presented in Table 3.1 highlight the importance of an attention mechanism in a model's architecture to learn crucial segments of an utterance for the ERC task.

Bibliography

[1] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586, 2005.

[2] N. Alswaidan and M. E. B. Menai. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62:2937–2987, 2020.

[3] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[4] E. Asutay and D. Västfjäll. 368Sound and Emotion. In *The Oxford Handbook of Sound and Imagination, Volume 2*. Oxford University Press, 09 2019.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[7] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74, 2018.

[8] Charles Robert Dawin. The expression of the emotions in man and animals, 1872.

[9] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, L.-W. Ku, et al. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*, 2018.

[10] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe. M2fnet: Multi-modal fusion network for emotion recognition in conversation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4652–4661, 2022.

[11] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] P. Ekman et al. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

[14] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[15] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, 2018:2122, 2018.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[18] T. Kim and P. Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*, 2021.

[19] Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[20] J. Li, D. Ji, F. Li, M. Zhang, and Y. Liu. Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200, 2020.

[21] Z. Li, F. Tang, M. Zhao, and Y. Zhu. Emocaps: Emotion capsule based model for conversational emotion recognition. *arXiv preprint arXiv:2203.13504*, 2022.

[22] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[24] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI conference on artificial intelligence*, 33(01):6818–6825, 2019.

[25] Y. Mao, Q. Sun, G. Liu, X. Wang, W. Gao, X. Li, and J. Shen. Dialoguetrm: Exploring the intra-and inter-modal emotional behaviors in the conversation. *arXiv preprint arXiv:2010.07637*, 2020.

[26] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17, 2011.

[27] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency. Utterance-level multimodal sentiment analysis. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982, 2013.

[28] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency. Context-dependent sentiment analysis in user-generated videos. *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883, 2017.

[29] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[30] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123, 2008.

[31] K. R. Scherer and H. G. Wallbott. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310, 1994.

[32] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[33] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456, 2012.

[34] W. Shen, J. Chen, X. Quan, and Z. Xie. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13789–13797, 2021.

[35] T. Shi and S.-L. Huang. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766, 2023.

[36] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021.

[37] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[38] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2019:6558, 2019.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[41] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[42] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[43] W. Zheng, J. Yu, R. Xia, and S. Wang. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15445–15459, 2023.