

Received May 9, 2019, accepted June 12, 2019, date of publication July 16, 2019, date of current version August 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929050

Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances

SOUJANYA PORIA¹, NAVONIL MAJUMDER¹, RADA MIHALCEA^{1,2}, AND EDUARD HOVY⁴

¹ISTD, Singapore University of Technology and Design, Singapore 487372

²CIC, Instituto Politécnico Nacional, Mexico City 07738, Mexico

³Computer Science and Engineering, University of Michigan, Ann Arbor, MI 48109, USA

⁴Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Corresponding author: Soujanya Poria (sporia@sutd.edu.sg)

This work was supported in part by the Michigan Institute for Data Science, in part by the National Science Foundation under Grant 1815291, in part by the John Templeton Foundation under Grant 61156, and in part by the DARPA under Grant HR001117S0026-AIDA-FP-045.

ABSTRACT Emotion is intrinsic to humans and consequently, emotion understanding is a key part of human-like artificial intelligence (AI). Emotion recognition in conversation (ERC) is becoming increasingly popular as a new research frontier in natural language processing (NLP) due to its ability to mine opinions from the plethora of publicly available conversational data on platforms such as Facebook, YouTube, Reddit, Twitter, and others. Moreover, it has potential applications in health-care systems (as a tool for psychological analysis), education (understanding student frustration), and more. In Addition, ERC is also extremely important for generating emotion-aware dialogues that require an understanding of the user's emotions. Catering to these needs calls for effective and scalable conversational emotion-recognition algorithms. However, it is a difficult problem to solve because of several research challenges. In this paper, we discuss these challenges and shed light on recent research in this field. We also describe the drawbacks of these approaches and discuss the reasons why they fail to successfully overcome the research challenges in ERC.

INDEX TERMS Emotion recognition, sentiment analysis, dialogue systems, natural language processing.

I. INTRODUCTION

Emotion is often defined as an individual's mental state associated with thoughts, feelings, and behavior. Stoics like Cicero organized emotions into four categories - *metus* (fear), *aegritudo* (pain), *libido* (lust), and *laetitia* (pleasure). Later, evolutionary theories of emotion were initiated in the late 19th century by Darwin and Prodder [1]. He hypothesized that emotions evolved through natural selection and, hence, have cross-culturally universal counterparts. In recent times, Plutchik [2] categorized emotion into eight primary types, visualized by the wheel of emotions. Further, Ekman [3] argued for the correlation between emotion and facial expression.

Natural language is often indicative of one's emotion. Hence, emotion recognition has been enjoying popularity in the field of NLP [4], [5], due to its widespread applications in opinion mining, recommender systems, health-care, and

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhou.

so on. Strapparava and Mihalcea [6] addressed the task of emotion detection in news headlines. A number of emotion lexicons [7], [8] have been developed to tackle the textual emotion recognition problem.

Only in the past few years has emotion recognition in conversation (ERC) gained attention from the NLP community [9]–[12] due to the growing availability of public conversational data. ERC can be used to analyze conversations that take place on social media. It can also aid in analyzing conversations in real times, which can be instrumental in legal trials, interviews, e-health services, and more.

Unlike vanilla emotion recognition of sentences/utterances, ERC ideally requires context modeling of the individual utterances. This context can be attributed to the preceding utterances, and relies on the temporal sequence of utterances. Compared to the recently published works on ERC [10]–[12], both lexicon-based [8], [13], [14] and modern deep learning-based [4], [5] vanilla emotion recognition approaches fail to work well on ERC datasets as these works ignore the conversation specific factors such as the presence

of contextual cues, the temporality in speakers' turns, or speaker-specific information. Fig. 4a and Fig. 4b show an example where the same utterance changes its meaning depending on its preceding utterance.

A. TASK DEFINITION

Given the transcript of a conversation along with speaker information of each constituent utterance, the ERC task aims to identify the emotion of each utterance from a set of pre-defined emotions. Fig. 2 illustrates one such conversation between two people, where each utterance is labeled by the underlying emotion. Formally, given the input sequence of N number of utterances $[(u_1, p_1), (u_2, p_2), \dots, (u_N, p_N)]$, where each utterance $u_i = [u_{i,1}, u_{i,2}, \dots, u_{i,T}]$ consists of T words $u_{i,j}$ and spoken by party p_i , the task is to predict the emotion label e_i of each utterance u_i .

B. CONTROLLING VARIABLES IN CONVERSATIONS

Conversations are broadly categorized into two categories: task oriented and chit-chat (also referred to as non-task oriented). Both types of conversations are governed by different factors or pragmatics [15], such as topic, interlocutors' personality, argumentation logic, viewpoint, intent [16], and so on. Fig. 1 shows how these factors play out in a dyadic conversation. Firstly, topic (*Topic*) and interlocutor personality (P_*) always influence the conversation, irrespective of the time. A speaker makes up his/her mind (S_*^t) about the reply (U_*^t) based on the contextual preceding utterances ($U_*^{<t}$) from both speaker and listener, the previous utterance being the most important one since it usually makes the largest change in the joint task model (for task-oriented conversations) or the speaker's emotional state (for chit-chat). Delving deeper, the pragmatic features, as explained by Hovy [15], like argumentation logic, interlocutor viewpoint, inter-personal relationship and dependency, situational awareness are encoded in speaker state (S_*^t). Intent (I_*^t) of the speaker is decided based on previous intent I_*^{t-2} and speaker state S_*^t , as the interlocutor may change his/her intent based on the opponent's utterance and current situation. Then, the speaker formulates appropriate emotion E_*^t for the response based on the state S_*^t and intent I_*^t . Finally, the response U_*^t is produced based on the speaker state S_*^t , intent I_*^t , and emotion E_*^t . We surmise that considering these factors would help representing the argument and discourse structure of the conversation, which leads to improved conversation understanding, including emotion recognition.

Early computational work on dialogue mostly focused on task-oriented cases, in which the overall conversational intent and step-by-step sub-goals played a large part [17], [18]. Cohen and Levesque [19] developed a model and logic to represent intentions and their connections to utterances, whose operators explicate the treatment of beliefs about the interlocutor's beliefs and vice versa, recursively. Emotion however played no role in this line of research. In more recent work, chatbots and chit-chat dialogue have become more prominent, in part due to the use of

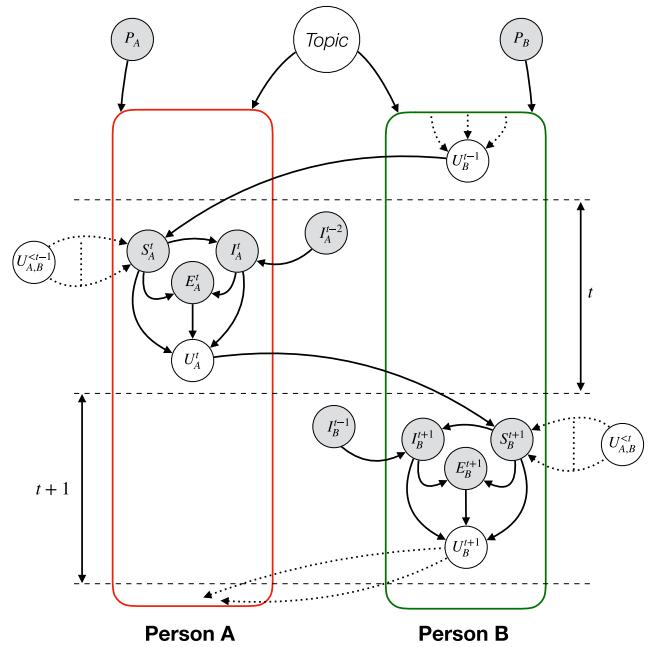


FIGURE 1. Interaction among different controlling variables during a dyadic conversation between persons A and B. Grey and white circles represent hidden and observed variables, respectively. P represents personality, U represents utterance, S represents interlocutor state, I represents interlocutor intent, E represents emotion and *Topic* represents topic of the conversation. This can easily be extended to multi-party conversations.

distributed (such as embedding) representations that do not readily support logical inference.

On conversational setting, D'Mello *et al.* [20] and [21] worked with small datasets with three and four emotion labels, respectively. This was followed by Phan *et al.* [22], where emotion detection on conversation transcript was attempted. Recently, several works [23], [24] have devised deep learning-based techniques for ERC. These works are crucial as we surmise an instrumental role of ERC in emotion-aware a.k.a. affective dialogue generation which has fallen within the topic of “text generation under pragmatics constraints” as proposed by Hovy [15]. Fig. 3 illustrates one such conversation between a human (*user*) and a medical chatbot (*health-assistant*). The assistant responds with emotion based on the *user*'s input. Depending on whether the *user* suffered an injury earlier or not, the *health-assistant* responds with excitement (evoking urgency) or happiness (evoking relief). As ERC is a new research field, outlining research challenges, available datasets, and benchmarks can potentially aid future research on ERC. In this paper, we aim to serve this purpose by discussing various factors that contribute to the emotion dynamics in a conversation. We surmise that this paper will not only help the researchers to better understand the challenges and recent works on ERC but also show possible future research directions. The rest of the paper is organized as follows: Section II presents the key research challenges; Section III and Section IV cover the datasets and recent progress in this field; finally Section V concludes the paper.

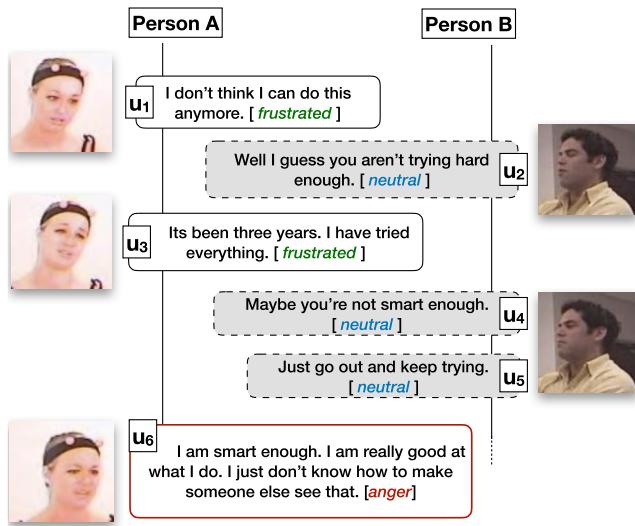


FIGURE 2. An abridged dialogue from the IEMOCAP dataset.

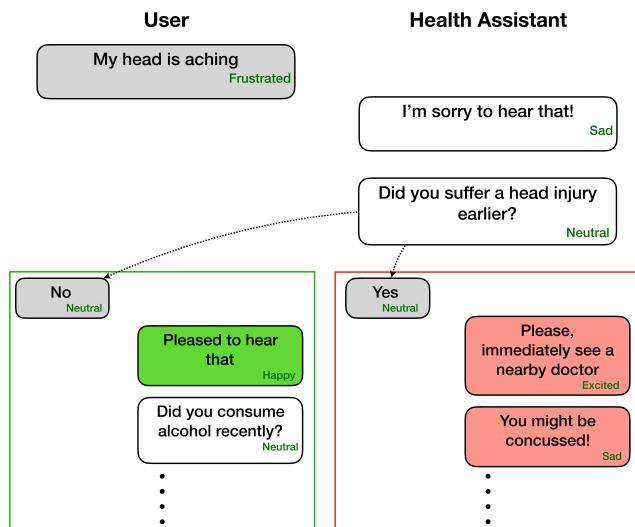


FIGURE 3. Illustration of an affective conversation where the emotion depends on the context.

II. RESEARCH CHALLENGES

Recent works on ERC, e.g., DialogueRNN [11] or ICON [23], strive to address several key research challenges that make the task of ERC difficult to solve:

A. CATEGORIZATION OF EMOTIONS

Emotion is defined using two type of models — categorical and dimensional. Categorical model classifies emotion into a fixed number of discrete categories. In contrast, dimensional model describes emotion as a point in a continuous multi-dimensional space.

In the categorical front, Plutchik [2]’s wheel of emotions defines eight discrete primary emotion types, each of which has finer related subtypes. On the other hand, Ekman [3] concludes six basic emotions — anger, disgust, fear, happiness, sadness, and surprise.

Most dimensional categorization models [25], [26] adopt two dimensions — valence and arousal. Valence represents the degree of emotional positivity and arousal represents the intensity of the emotion. In contrast with the categorical models, dimensional models map emotion into a continuous spectrum rather than hard categories. This enables easy and intuitive comparison of two emotional states using vector operations, whereas comparison is non-trivial for categorical models. As there are multiple categorization and dimensional taxonomies available, it is challenging to select one particular model for annotation. Choosing a simple categorization model e.g., Ekman’s model has a major drawback as these models are unable to ground complex emotions. On the other hand, complex emotion models such as Plutchik’s model make it very difficult for the annotators to discriminate between the related emotions, e.g., discerning anger from rage. Complex emotion models also increase the risk of obtaining a lower inter-annotator agreement.

The popular ERC dataset IEMOCAP [27] adopted both categorical and dimensional models. However, newer ERC datasets like DailyDialogue [28] have employed only categorical model due to its more intuitive nature. Most of the available datasets for emotion recognition in conversation adopted simple taxonomies, which are slight variants of Ekman’s model. Each emotional utterance in the EmoContext dataset is labeled with one of the following emotions: happiness, sadness, and anger. The majority of the utterances in EmoContext do not elicit any of these three emotions and are annotated with an extra label: others. Naturally, the inter-annotator agreement for the EmoContext dataset is higher due to its simplistic emotion taxonomy. However, the short context length and simple emotion taxonomy make ERC on this dataset less challenging.

B. BASIS OF EMOTION ANNOTATION

Annotation with emotion labels is challenging as the label depends on the annotators perspective. Self-assessment by the interlocutors in a conversation is arguably the best way to annotate utterances. However, in practice it is unfeasible as real-time tagging of unscripted conversations will impact the conversation flow. Post-conversation self-annotation could be an option, but it has not been done yet. As such, many ERC datasets [27] are scripted and annotated by a group of people uninvolved with the script and conversation. The annotators are given the context of the utterances as prior knowledge for accurate annotation. Often pre-existing transcripts are annotated for quick turn-around, as in EmotionLines [10].

The annotators also need to be aware of the interlocutors perspective for situation-aware annotation. For example, the emotion behind the utterance “Lehman Brothers’ stock is plummeting!!” depends on whether the speaker benefits from the crash. The annotators should be aware of the nature of association between the speaker and Lehman Brothers for accurate labeling.

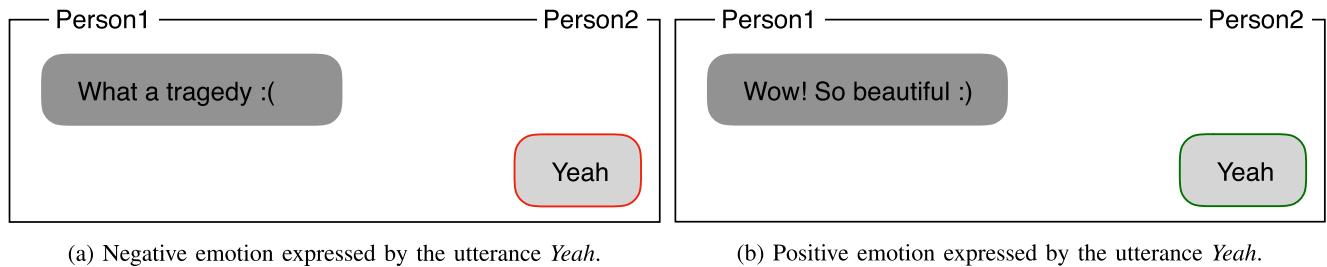


FIGURE 4. Role of context in emotion recognition in conversation.

C. CONVERSATIONAL CONTEXT MODELING

Context is at the core of the NLP research. According to several recent studies [29], [30], contextual sentence and word embeddings can improve the performance of the state-of-the-art NLP systems by a significant margin.

The notion of context can vary from problem to problem. For example, while calculating word representations, the surrounding words carry contextual information. Likewise, to classify a sentence in a document, other neighboring sentences are considered as its context. In Poria *et al.* [31], surrounding utterances are treated as context and they experimentally show that contextual evidence indeed aids in classification.

Similarly in conversational emotion-detection, to determine the emotion of an utterance at time t , the preceding utterances at time $< t$ can be considered as its context. However, computing this context representation often exhibits major difficulties due to emotional dynamics.

Emotional dynamics of conversations consists of two important aspects: *self* and *inter-personal dependencies* [32]. Self-dependency, also known as *emotional inertia*, deals with the aspect of emotional influence that speakers have on themselves during conversations [33]. On the other hand, inter-personal dependencies relate to the emotional influences that the counterparts induce into a speaker. Conversely, during the course of a dialogue, speakers also tend to mirror their counterparts to build rapport [34]. This phenomenon is illustrated in Fig. 2. Here, P_a is frustrated over her long term unemployment and seeks encouragement (u_1, u_3). P_b , however, is pre-occupied and replies sarcastically (u_4). This enrages P_a to appropriate an angry response (u_6). In this dialogue, *emotional inertia* is evident in P_b who does not deviate from his nonchalant behavior. P_a , however, gets emotionally influenced by P_b . Modeling self and inter-personal relationship and dependencies may also depend on the topic of the conversation as well as various other factors like argument structure, interlocutors' personality, intents, viewpoints on the conversation, attitude towards each other etc.. Hence, analyzing all these factors are key for a true self and inter-personal dependency modeling that can lead to enriched context understanding.

The contextual information can come from both local and distant conversational history. While the importance of local context is more obvious, as stated in recent works, distant

context often plays a less important role in ERC. Distant contextual information is useful mostly in the scenarios when a speaker refers to earlier utterances spoken by any of the speakers in the conversational history.

The usefulness of context is more prevalent in classifying short utterances, like "yeah", "okay", "no", that can express different emotions depending on the context and discourse of the dialogue. The examples in Fig. 4a and Fig. 4b explain this phenomenon. The emotions expressed by the same utterance "Yeah" in both these examples differ from each other and can only be inferred from the context.

Finding contextualized conversational utterance representations is an active area of research. Leveraging such contextual clues is a difficult task. Memory networks, RNNs, and attention mechanisms have been used in previous works, e.g., HRLCE or DialogueRNN, to grasp information from the context.

D. SPEAKER SPECIFIC MODELING

Individuals have their own subtle way of expressing emotions. For instance, some individuals are more sarcastic than others. For such cases, the usage of certain words would vary depending on if they are being sarcastic. Let's consider this example, P_a : "The order has been cancelled.", P_b : "This is great!". If P_b is a sarcastic person, then his response would express negative emotion to the order being canceled through the word *great*. On the other hand, P_b 's response, *great*, could be taken literally if the canceled order is beneficial to P_b (perhaps P_b cannot afford the product he ordered). Since, necessary background information is often missing from the conversations, speaker profiling based on preceding utterances often yields improved results.

E. LISTENER SPECIFIC MODELING

During a conversation, the listeners make up their mind about the speaker's utterance as it's spoken. However, there is no textual data on the listener's reaction to the speaker while the speaker speaks. A model must resort to visual modality to model the listener's facial expression to capture the listener's reaction. However, according to DialogueRNN, capturing listener reaction does not yield any improvement as the listener's subsequent utterance carries their reaction. Moreover, if the listener never speaks in a conversation, his/her reaction remains irrelevant. Nonetheless, listener modeling can be

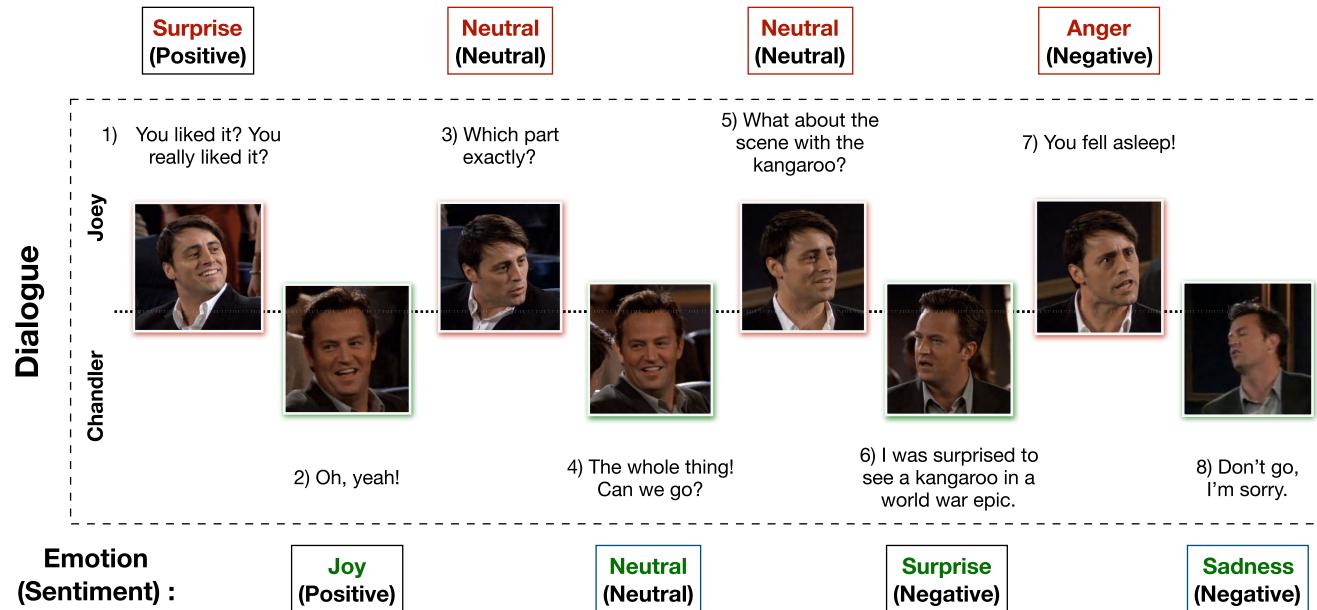


FIGURE 5. Emotion shift of speakers in a dialogue in comparison with speaker's previous emotion. Red and blue colors are used to show the emotion shift of Joey and Chandler respectively.

useful in the scenarios where continuous emotion recognition of every moment of the conversation is necessary, like audience reaction during a political speech, as opposed to emotion recognition of each utterance.

F. PRESENCE OF EMOTION SHIFT

Due to *emotional inertia*, participants in a conversation tend to stick a particular emotional state, unless some external stimuli, usually the other participants, invoke a change. This is illustrated in Fig. 5, where Joey changes his emotion from *neutral* to *anger* due to the last utterance of Chandler, which was unexpected and rather shocking to Joey. This is a hard problem to solve, as the state-of-the-art ERC model, DialogueRNN is more accurate in emotion detection for the utterances without emotional shift or when the shift is to a similar emotion (e.g., from fear to sad).

The state-of-the-art methods keeps mimicking the same emotion for a particular party, since an abrupt change of emotion is unlikely. Hence, these methods fail in most cases where a change occurs. To tackle this, a new problem of detecting emotion shift can be framed:

- 1) based on the historical utterances and the present utterance, is there an emotion shift (binary classification)?
- 2) if there is a shift then what is the target emotion (multi-label classification)?

As baseline, the performance of conditional random field (CRF) would be interesting as it models label dependencies.

G. FINE-GRAINED EMOTION RECOGNITION

Fine-grained emotion recognition aims at recognizing emotion expressed on explicit and implicit topics. It involves a deeper understanding of the topic of the conversation,

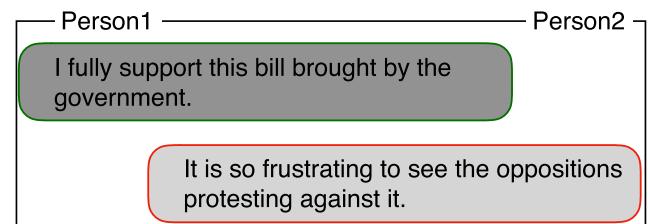


FIGURE 6. Fine-grained emotion understanding: An example.

interlocutor opinion, and stand. For example, in Fig. 6, while both persons take a supportive stand for the government's bill, they use completely opposite emotions to express it. It is not possible for a vanilla emotion recognizer to understand the positive emotion of both the interlocutors on the aspect of government's bill. Only by interpreting *Person 2*'s frustration about the opposition's protest against the bill can a classifier infer *Person 2*'s support for the bill. On the other hand, even though *Person 1* does not explicitly express his/her opinion on the opposition, from the discourse of the conversation, it can be inferred that *Person 1* holds a negative opinion on the opposition.

H. MULTIPARTY CONVERSATION

In a multiparty conversation, more than two participants are involved. Naturally, emotion recognition in such conversations is more challenging in comparison with dyadic conversations due to the difficulty in tracking individual speaker states and handling co-references.

I. PRESENCE OF SARCASM

Sarcasm is a linguistic tool that uses irony to express contempt. An ERC system incapable of detecting sarcasm mostly fails to predict emotion of the sarcastic utterances correctly.

TABLE 1. Label distribution statistics in different emotion recognition datasets.

Label	DailyDialog	MELD	EmotionLines	IEMOCAP	EmoContext
Neutral	85572	6436	6530	1708	-
Happiness/Joy	12885	2308	1710	648	4669
Surprise	1823	1636	1658	-	-
Sadness	1150	1002	498	1084	5838
Anger	1022	1607	772	1103	5954
Disgust	353	361	338	-	-
Fear	74	358	255	-	-
Frustrated	-	-	-	1849	-
Excited	-	-	-	1041	-
Other	-	-	-	-	21960

Sarcasm detection in a conversation largely depends on the context and discourse of the conversation. For example, the utterance “*The part where Obama signed it*” can only be detected as sarcastic if we look at the previous utterance “*What part of this would be unconstitutional?*”. Sarcasm nature is also person dependent, which again warrants speaker profiling in the conversation.

J. EMOTION REASONING

The ability to reason is necessary for any explainable AI system. In the context of ERC, it is often desired to understand the cause of an expressed emotion by a speaker. As an example, we can refer to 2. An ideal ERC system, with the ability of emotion reasoning, should perceive the reason for *Person_A*’s *anger*, expressed in *u₆* of Fig. 2. It is evident upon observation that this *anger* is caused by the persistent nonchalant behavior of *Person_B*. Readers should not conflate emotion reasoning with context modeling, which we discuss earlier in this section. Unlike context modeling, emotion reasoning does not only find the contextual utterances in conversational history that triggers the emotion of an utterance, but also determines the function of those contextual utterances on the target utterance. In Fig. 2, it is the indifference of *Person_B*, reflected by *u₄* and *u₅*, that makes *Person_A* angry. Similarly, in Fig. 5, Joey expresses anger once he ascertains Chandler’s deception in the previous utterance. It is hard to define a taxonomy or tagset for emotion reasoning. At present, there is no available dataset which contains such rich annotations. Building such dataset would enable future dialogue systems to frame meaningful argumentation logic and discourse structure, taking one step closer to human-like conversation.

III. DATASETS

In the last few years, emotion recognition in conversation has gained major research interest, mainly because of its potential application in dialogue systems to generate emotion-aware and empathetic dialogues [12]. The primary goal of ERC task is to label each utterance in the conversation with an emotion label. In this section, we discuss the publicly available ERC datasets as well as the shortcomings of these datasets.

There are a few publicly available datasets for ERC - IEMOCAP [27], SEMAINE [35], Emotionlines [10], MELD [36], DailyDialog [28], and EmoContext [37]. A detailed comparison of these datasets is drawn in Table 2. Out of these five datasets, IEMOCAP, SEMAINE, and MELD are multimodal (containing acoustic, visual and textual information) and the remaining two are textual. Apart from SEMAINE dataset, rest of the datasets contains categorical emotion labels. In contrast, each utterance of SEMAINE dataset is annotated with four real valued affective attributes: valence ($[-1, 1]$), arousal ($[-1, 1]$), expectancy ($[-1, 1]$), and power ($[0, \infty)$). We also show the emotion label distribution of these datasets in Table 1. In EmoContext dataset, an emotion label is assigned to only the last utterance of each dialogue. None of these datasets can be used for emotion reasoning as they lack necessary annotation details required for the reasoning task. Readers should also note that, all these datasets do not contain fine-grained and topic level emotion annotation.

IV. RECENT ADVANCES

In this section we give a brief introduction to the recent work on this topic. We also compare the approaches and report their drawbacks. As depicted in Fig. 1, recognizing emotion of an utterance in a conversation primarily depends on these following three factors:

- 1) the utterance itself and its context defined by the interlocutors’ preceding utterances in the conversation, as well as intent and the topic of the conversation,
- 2) the speaker’s state comprising variables like personality and argumentation logic and,
- 3) emotions expressed in the preceding utterances.

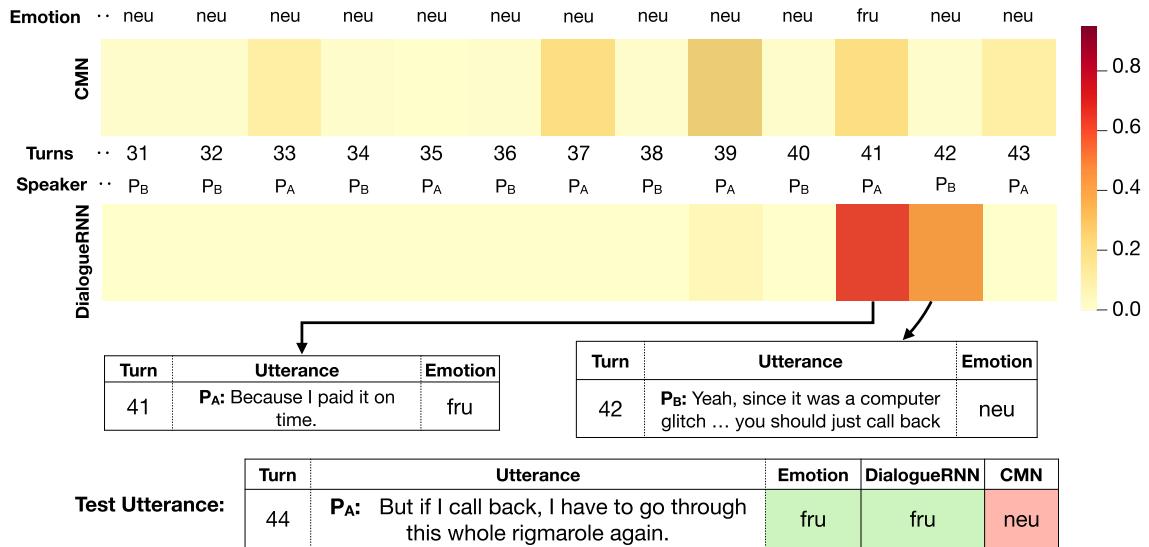
Although, IEMOCAP and SEMAINE have been developed almost a decade ago, most of the works that used these two datasets did not consider the aforementioned factors.

A. BENCHMARKS AND THEIR DRAWBACKS

Based on these factors, a number of approaches to address the ERC problem have been proposed recently. Conversational memory network (CMN), proposed by Hazarika *et al.* [38] for dyadic dialogues, is one of the first ERC approaches that

TABLE 2. Comparison among IEMOCAP, SEMAINE, emotionLines, MELD, and dailydialog datasets.

Dataset	# dialogues			# utterances		
	train	val	test	train	val	test
IEMOCAP	120		31	5810		1623
SEMAINE	63		32	4368		1430
EmotionLines	720	80	200	10561	1178	2764
MELD	1039	114	280	9989	1109	2610
DailyDialog	11,118	1000	100	87,832	7912	7863
EmoContext	30,159	2754	5508	90,477	8262	16,524

**FIGURE 7.** Comparison of attention scores over utterance history of CMN and DialogueRNN. Higher attention value signifies more important contextual information. Note: Figure taken from Majumder et al. [11].

utilizes distinct memories for each speaker for speaker-specific context modeling. Later, Hazarika *et al.* [23] improved upon this approach with interactive conversational memory network (ICON), which interconnects these memories to model self and inter-speaker emotional influence. None of these two methods actually exploit the speaker information of the target utterance for classification. This makes the model blind to speaker-specific nuances.

DialogueRNN [11] aims to solve this issue by considering the speaker information of the target utterance and, further, modeling self and inter-speaker emotional influence with a hierarchical multi-stage RNN with attention mechanism. On both IEMOCAP and SEMAINE datasets, DialogueRNN outperformed (Table 3 and Table 4) the other two approaches.

The need to grasp inter-speaker dependency for ERC is also acknowledged and modeled in [39] by leveraging quantum theory and LSTM. Their network, Quantum-Inspired Interactive Networks (QIN) has outperformed CMN and ICON on IEMOCAP and MELD datasets. Recently, Yeh *et al.* [9] proposed an ERC method called

Interaction-aware Attention Network (IANN) by leveraging inter-speaker relation modeling. Similar to ICON and CMN, IANN (Fig. 8) utilizes distinct memories for each speaker.

All of these models affirm that contextual history, modeling self and inter-speaker influence are beneficial to ERC (shown in Fig. 7 and Fig. 10). Further, DialogueRNN shows that the nearby utterances are generally more context rich and ERC performance improves when the future utterances, at time $> t$, are available. This is indicated by Fig. 10, where DialogueRNN uses both past and future utterances as context with roughly the same frequency. Also, the distant utterances are used less frequently than the nearby utterances. On the other hand, CMN and ICON do not use future utterances as context at all. However, for real-time applications, systems cannot rely on future utterances. In such cases, CMN, ICON, and DialogueRNN with fixed context window would be befitting.

All these networks, namely CMN, ICON, IANN, and DialogueRNN, perform poorly on the utterances with emotion shift. In particular, the cases where the emotion of the target utterance differs from the previous utterance, DialogueRNN

TABLE 3. Comparison between DialogueRNN and baseline methods on IEMOCAP dataset; bold font denotes the best performances. Average(w) = Weighted average. ICON results differ from the original paper [23] as in our experiment, we disregard their contextual feature extraction and pre-processing part. More details can be found in Majumder et al. [11].

Methods	IEMOCAP													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average(w)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN	27.77	29.86	57.14	53.83	34.33	40.14	61.17	52.44	46.15	50.09	62.99	55.75	48.92	48.18
memnet	25.72	33.53	55.53	61.77	58.12	52.84	59.32	55.39	51.50	58.30	67.20	59.00	55.72	55.10
bc-LSTM	29.17	34.43	57.14	60.87	54.17	51.81	57.06	56.73	51.17	57.95	67.19	58.92	55.21	54.95
bc-LSTM+Att	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19
CMN	25.00	30.38	55.92	62.41	52.86	52.39	61.76	59.83	55.52	60.25	71.13	60.69	56.56	56.13
ICON	22.22	29.91	58.78	64.57	62.76	57.38	64.71	63.04	58.86	63.42	67.19	60.81	59.09	58.54
DialogueRNN	25.69	33.18	75.10	78.80	58.59	59.21	64.71	65.28	80.27	71.86	61.15	58.91	63.40	62.75

TABLE 4. Comparison between DialogueRNN and baseline methods on SEMAINE dataset; Acc. = Accuracy, MAE = Mean Absolute Error, r = Pearson correlation coefficient; bold font denotes the best performances. More details can be found in Majumder et al. [11].

Methods	SEMAINE							
	Valence		Arousal		Expectancy		Power	
	MAE	r	MAE	r	MAE	r	MAE	r
CNN	0.545	-0.01	0.542	0.01	0.605	-0.01	8.71	0.19
memnet	0.202	0.16	0.211	0.24	0.216	0.23	8.97	0.05
bc-LSTM	0.194	0.14	0.212	0.23	0.201	0.25	8.90	-0.04
bc-LSTM+Att	0.189	0.16	0.213	0.25	0.190	0.24	8.67	0.10
CMN	0.192	0.23	0.213	0.29	0.195	0.26	8.74	-0.02
ICON	0.181	0.245	0.19	0.317	0.185	0.272	8.45	0.244
DialogueRNN	0.168	0.35	0.165	0.59	0.175	0.37	7.90	0.37

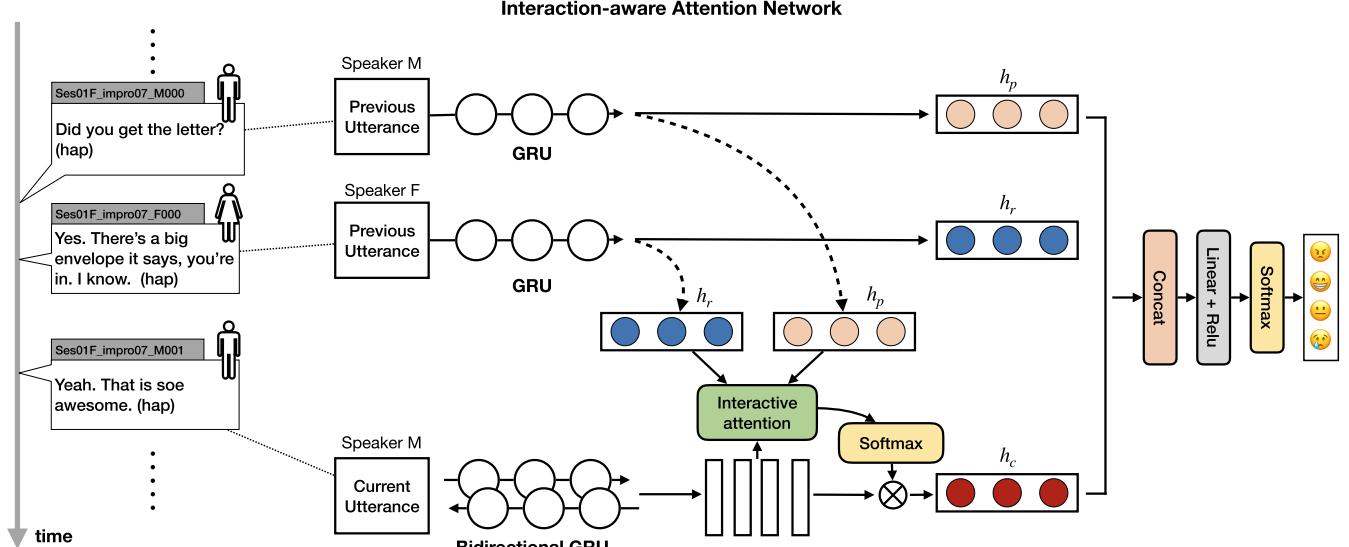


FIGURE 8. Interactive-aware Attention Network (IANN), proposed by Yeh et al. [9].

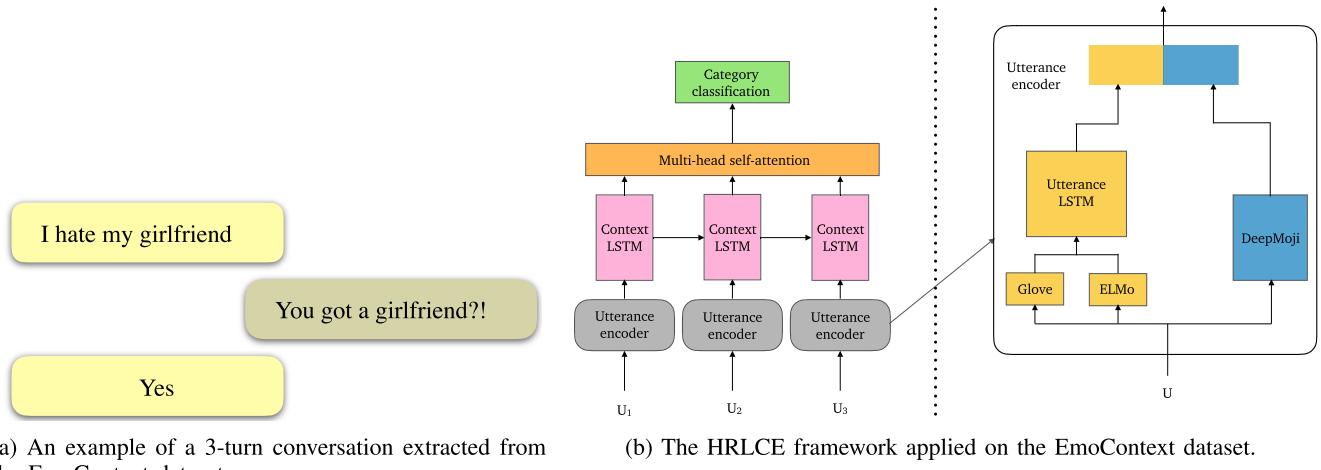
could only correctly predict 47.5% instances. This stands less as compared to the 69.2% success-rate that it achieves at the regions of no emotional-shift.

Among these three approaches, only DialogueRNN is capable of handling multiparty conversations on large scale. However, on the multiparty conversational dataset MELD, only a little performance improvement (shown in Table 5)

is observed by DialogueRNN compared to bc-LSTM which depicts a future research direction on multiparty ERC. ICON and CMN are designed to detect emotions in dyadic dialogues. Adapting ICON and CMN to apply on multiparty conversational dataset MELD can cause scalability issue in situations when number speakers participating in a conversation in the test data is more than the training data.

TABLE 5. Test-set F-score results of bc-LSTM and DialogueRNN for emotion classification in MELD. Note: *w-avg* denotes weighted-average. *text-CNN*: CNN applied on text, contextual information were not used.

Modality	Emotions								
	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	W-avg.	
CNN	34.49	8.22	3.74	49.39	74.88	21.05	45.45	55.02	
bc-LSTM	42.06	21.69	7.75	54.31	71.63	26.92	48.15	56.44	
DialogueRNN	40.59	2.04	8.93	50.27	75.75	24.19	49.38	57.03	



(a) An example of a 3-turn conversation extracted from the EmoContext dataset.

(b) The HRLCE framework applied on the EmoContext dataset.

FIGURE 9. The EmoContext dataset [37] and the HRLCE [40] framework.

TABLE 6. Recent works on the EmoContext dataset.

Methods	F1
CoAStaL [41]	73.40
TL-BLSTM [42]	74.50
LIRMM-Advanse [43]	75.82
DialogueRNN [11]	75.80
HRLCE [40]	76.66

Due to the sequential nature of the utterances in conversations, RNNs are used for context generation in the aforementioned models. However, there is ample room for improvement, as the RNN-based context representation methods perform poorly in grasping long distant contextual information.

Recently, two shared tasks — EmotionX¹ (co-located with SocialNLP workshop) and EmoContext² (co-located with Semeval 2019) have been organized to address the ERC problem. EmoContext shared task has garnered more than 500 participants, affirming the growing popularity of this research field. Compared to other datasets, EmoContext dataset [37] has very short conversations consisting only three utterances where the goal is to label the 3rd utterance as shown in Fig. 9a.

Emotion labels of the previous utterances are not present in the EmoContext dataset. The key works [24], [37], [40]–[44] on this dataset have mainly leveraged on context modeling

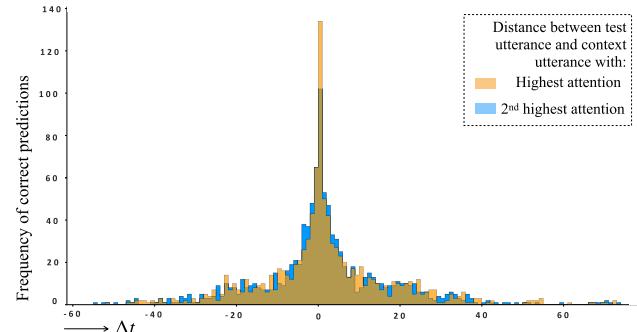


FIGURE 10. Histogram of Δt = distance between the target utterance and its context utterance based on DialogueRNN's attention scores. Note: Figure taken from Majumder et al. [11].

using bc-LSTM architecture [31] that encapsulates the temporal order of the utterances using an LSTM. A common trend can be noticed in these works, where traditional word embeddings, such as Glove [45], are combined with contextualized word embeddings, such as ELMo [29] to improve the performance. Most of these works use attention mechanism on top of the bc-LSTM to enrich context representation. In Fig. 9b, we depict the HRLCE framework, proposed by Huang et al. [40], that comprises of an utterance encoder and a context encoder that takes input from the utterance encoder. To represent each utterance, HRLCE utilizes ELMo [29], Glove [45], and Deepmoji [46].

The context encoder in HRLCE adapts the bc-LSTM framework followed by a multi-head attention layer. Huang et al. [40] applied HRLCE framework only on the

¹<https://sites.google.com/view/emotionx2019/>

²<https://www.humanizing-ai.com/emocontext.html>

EmoContext dataset. However, HRLCE can be easily adapted to apply on other ERC datasets. It should be noted that none of the works on the EmoContext dataset utilize speaker information. In fact, in our experiments, we found that DialogueRNN, which makes use of the speaker information, performs similar (Table 6) to Bae *et al.* [24], Huang *et al.* [40], and Chatterjee *et al.* [37] on EmoContext dataset. One possible reason for this could be the presence of very short context history in the dataset that renders speaker information inconsequential.

V. CONCLUSION

Emotion recognition in conversation has been gaining popularity among NLP researchers. In this paper, we summarized the recent advances in this task and highlight several key research challenges associated with this research area. Further, we pointed out how current work has partly addressed these challenges, while also presenting some shortcomings. Overall, we surmised that an effective emotion-shift recognition model and context encoder can yield significant performance improvement over chit-chat dialogue, and even improve some aspects of task-oriented dialogue. Moreover, challenges like topic-level speaker-specific emotion recognition, ERC on multiparty conversations, and conversational sarcasm detection can form new research directions. Additionally, fine-grained speaker-specific continuous emotion recognition may become of interest for the purpose of tracking emotions during long monologues. We believe that addressing each of the challenges outlined in this paper will not only enhance AI-enabled conversation understanding, but also improve the performance of dialogue systems by catering to affective information.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these agencies.

REFERENCES

- [1] C. Darwin and P. Prodg, *The Expression of the Emotions in Man and Animals*. Oxford, U.K.: Oxford Univ. Press, 1998.
- [2] R. Plutchik, “A psychoevolutionary theory of emotions,” *Social Sci. Inf.*, vol. 21, nos. 4–5, pp. 529–553, 1982.
- [3] P. Ekman, “Facial expression and emotion,” *Amer. Psychol.*, vol. 48, no. 4, p. 384, 1993.
- [4] B. Kratzwald, S. Ilic, M. Kraus, S. Feuerriegel, and H. Prendinger, “Deep learning for affective computing: Text-based emotion recognition in decision support,” 2018, *arXiv:1803.06397*. [Online]. Available: <https://arxiv.org/abs/1803.06397>
- [5] N. Colneri and J. Demsar, “Emotion recognition on Twitter: Comparative study and training a unison model,” *IEEE Trans. Affect. Comput.*, to be published.
- [6] C. Strapparava and R. Mihalcea, “Semeval-2007 task 14: Affective text,” in *Proc. 4th Int. Workshop Semantic Eval.*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 70–74. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1621474.1621487>
- [7] C. Strapparava and A. Valitutti, “Wordnet affect: An affective extension of WordNet,” in *Proc. LREC*, vol. 4, 2004, p. 40.
- [8] S. M. Mohammad and P. D. Turney, “Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon,” in *Proc. NAACL HLT Workshop Comput. Approaches Anal. Gener. Emotion Text*. Los Angeles, CA, USA: Association for Computational Linguistics, 2010, pp. 26–34.
- [9] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, “An interaction-aware attention network for speech emotion recognition in spoken dialogs,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6685–6689.
- [10] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, “Emotionlines: An emotion corpus of multi-party conversations,” 2018, *arXiv:1802.08379*. [Online]. Available: <https://arxiv.org/abs/1802.08379>
- [11] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, “DialogueRNN: An attentive RNN for emotion detection in conversations,” 2018, *arXiv:1811.00405*. [Online]. Available: <https://arxiv.org/abs/1811.00405>
- [12] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, “Emotional chatting machine: Emotional conversation generation with internal and external memory,” in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 730–738.
- [13] C.-H. Wu, Z.-J. Chuang, and Y.-C. Lin, “Emotion recognition from text using semantic labels and separable mixture models,” *ACM Trans. Asian Lang. Inf. Process.*, vol. 5, no. 2, pp. 165–183, 2006.
- [14] S. Shaheen, W. El-Hajj, H. Hajj, and S. Elbassuoni, “Emotion recognition from text based on automatically generated rules,” in *Proc. IEEE Int. Conf. Data Mining Workshop*, Dec. 2014, pp. 383–392.
- [15] E. Hovy, “Generating natural language under pragmatic constraints,” *J. Pragmatics*, vol. 11, no. 6, pp. 689–719, 1987.
- [16] J. J. Schlöder and R. Fernández, “Clarifying intentions in dialogue: A corpus study,” in *Proc. 11th Int. Conf. Comput. Semantics*, 2015, pp. 46–51.
- [17] B. J. Grosz and C. L. Sidner, “Attention, intentions, and the structure of discourse,” *Comput. Linguistics*, vol. 12, no. 3, pp. 175–204, 1986.
- [18] D. E. Appelt, *Planning English Sentences*. Cambridge, U.K.: Cambridge Univ. Press, 1992.
- [19] P. R. Cohen and H. J. Levesque, “Speech acts and rationality,” in *Proc. 23rd Annu. Meeting Assoc. Comput. Linguistics*. Chicago, IL, USA: Association for Computational Linguistics, 1985, pp. 49–60.
- [20] S. K. D’Mello, S. D. Craig, J. Sullins, and A. C. Graesser, “Predicting affective states expressed through an emote-aloud procedure from autotutor’s mixed-initiative dialogue,” *Int. J. Artif. Intell. Educ.*, vol. 16, no. 1, pp. 3–28, 2006.
- [21] C. Yang, K. H.-Y. Lin, and H.-H. Chen, “Emotion classification using Web blog corpora,” in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Nov. 2007, pp. 275–278.
- [22] D. A. Phan, H. Shindo, and Y. Matsumoto, “Multiple emotions detection in conversation transcripts,” in *Proc. 30th Pacific Asia Conf. Lang., Inf. Comput.*, Seoul, South Korea, Oct. 2016, pp. 85–94. [Online]. Available: <https://www.aclweb.org/anthology/Y16-2006>
- [23] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, “ICON: Interactive conversational memory network for multimodal emotion detection,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2594–2604.
- [24] J. Bae, J. Choi, and S.-G. Lee, “SNU_IDS at SemEval-2019 task 3: Addressing training-test class distribution mismatch in conversational classification,” 2019, *arXiv:1903.02163*. [Online]. Available: <https://arxiv.org/abs/1903.02163>
- [25] J. A. Russell, “A circumplex model of affect,” *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.
- [26] A. Mehrabian, “Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament,” *Current Psychol.*, vol. 14, no. 4, pp. 261–292, 1996. doi: [10.1007/BF02686918](https://doi.org/10.1007/BF02686918).
- [27] 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,” *Lang. Resour. Eval.*, vol. 42, no. 4, p. 335, 2008.
- [28] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, “Dailydialog: A manually labelled multi-turn dialogue dataset,” 2017, *arXiv:1710.03957*. [Online]. Available: <https://arxiv.org/abs/1710.03957>
- [29] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” 2018, *arXiv:1802.05365*. [Online]. Available: <https://arxiv.org/abs/1802.05365>
- [30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, *arXiv:1810.04805*. [Online]. Available: <https://arxiv.org/abs/1810.04805>

- [31] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. ACL*, vol. 1, 2017, pp. 873–883.
- [32] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Conf. Assoc. Comput. Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536.
- [33] P. Kuppens, N. B. Allen, and L. B. Sheeber, "Emotional inertia and psychological maladjustment," *Psychol. Sci.*, vol. 21, no. 7, pp. 984–991, 2010.
- [34] C. Navarretta, "Mirroring facial expressions and emotions in dyadic conversations," in *Proc. LREC*, 2016, pp. 469–474.
- [35] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, Jan./Mar. 2012.
- [36] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," 2018, *arXiv:1810.02508*. [Online]. Available: <https://arxiv.org/abs/1810.02508>
- [37] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions in text using deep learning and big data," *Comput. Hum. Behav.*, vol. 93, pp. 309–317, Apr. 2019.
- [38] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2018, pp. 2122–2132.
- [39] Y. Zhang, Q. Li, D. Song, P. Zhang, and P. Wang, "Quantum-inspired interactive networks for conversational sentiment analysis," in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2019, pp. 1–8.
- [40] C. Huang, A. Trabelsi, and O. R. Zaïane, "ANA at SemEval-2019 task 3: Contextual emotion detection in conversations through hierarchical LSTMs and BERT," 2019, *arXiv:1904.00132*. [Online]. Available: <https://arxiv.org/abs/1904.00132>
- [41] A. V. González, V. P. B. Hansen, J. Bingel, and A. Søgaard, "CoAStaL at SemEval-2019 Task 3: Affect classification in dialogue using attentive BiLSTMs," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 169–174.
- [42] A. Boucheik, P. Joshi, L. Boucheik, and H. Aflai, "EPITA-ADAPT at SemEval-2019 Task 3: Detecting emotions in textual conversations using deep learning models combination," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 215–219.
- [43] W. Ragheb, J. Azé, S. Bringay, and M. Servajean, "LIRMM-advanse at SemEval-2019 Task 3: Attentive conversation modeling for emotion detection and classification," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 251–255.
- [44] W. Ragheb, J. Azé, S. Bringay, and M. Servajean, "Attention-based modeling for emotion detection and classification in textual conversations," 2019, *arXiv:1906.07020*. [Online]. Available: <https://arxiv.org/abs/1906.07020>
- [45] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
- [46] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," 2017, *arXiv:1708.00524*. [Online]. Available: <https://arxiv.org/abs/1708.00524>



SOUJANYA PORIA received the Ph.D. degree in computer science from the University of Stirling, U.K. He is currently an Assistant Professor of computer science with the Singapore University of Technology and Design (SUTD), Singapore. He has coauthored more than 70 papers, and published in top-tier conferences and journals, such as ACL, EMNLP, AAAI, NAACL, *Neurocomputing* and *Computational Intelligence Magazine*. He was a recipient of the prestigious early career research award called NTU Presidential Postdoctoral Fellowship, in 2018. He has given several invited talks at venues like CICLing 2018, which is a large international NLP conference.



NAVONIL MAJUMDER received the M.Sc. degree from the CIC, Instituto Politécnico Nacional, Mexico, in 2017, where he is currently pursuing the Ph.D. degree. He is also a Visiting Researcher with Nanyang Technological University (NTU). He has more than 15 research publications in top-tier conferences and journals, such as ACL, EMNLP, AAAI, *Knowledge-Based Systems*, and the *IEEE Intelligent Systems*. His research interests include natural language processing, sentiment analysis, and multimodal language processing. His research works have occasionally been featured in news portals, such as KDDNuggets and Datanami. He received the Lázaro Cárdenas Gold Medal from the President of Mexico for his M.Sc. study.



RADA MIHALCEA is currently a Professor of computer science and engineering with the University of Michigan and the Director of the Michigan Artificial Intelligence Laboratory. Her research interests include lexical semantics, multilingual NLP, and computational social sciences. She serves or has served on the editorial boards of the *Journals of Computational Linguistics*, *Language Resources and Evaluations*, *Natural Language Engineering*, the *Journal of Artificial Intelligence Research*, the *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, and *Transactions of the Association for Computational Linguistics*. She was a recipient of the NSF CAREER Award, in 2008, and the Presidential Early Career Award for Scientists and Engineers from President Obama, in 2009. She was a Program Co-Chair of EMNLP 2009 and ACL 2011, and a General Chair of NAACL 2015 and *SEM 2019. She currently serves as the ACL Vice-President Elect. In 2013, she was made an honorary citizen of her hometown of Cluj-Napoca, Romania.



EDUARD HOVY received the Ph.D. degree in computer science from Yale University. He received the honorary doctorates from the National University of Distance Education (UNED) in Madrid, in 2013, and the University of Antwerp, in 2015. He is currently a Research Professor of the Language Technologies Institute, Carnegie Mellon University. He is one of the original 17 Fellows of the Association for Computational Linguistics. He has published more than 500 research papers. His research focuses on various topics, including aspects of the computational semantics of human language. He is a Fellow of the AAAI. He serves or has served on the editorial boards of several journals, such as the *ACM Transactions on Asian Language Information Processing*, *TALIP* and *Language Resources and Evaluation* (LRE).