Real-Time Multimodal Emotion Recognition in Conversation for Multi-Party Interactions

SANDRATRA RASENDRASOA[1**], University of Rouen Normandie, France
SEBASTIEN ADAM[1], University of Rouen Normandie, France
ALEXANDRE PAUCHET[2], Insa Rouen Normandie, France
JULIEN SAUNIER[2], Insa Rouen Normandie, France

In order to improve multi-party social interaction with artificial companions such as robots or virtual agents, real-time Emotion Recognition in Conversation (ERC) is required. In t his context, ERC is a challenging task which involves multiple challenges, such as processing multimodal data over time, taking into account the multi-party context with any number of participants, understanding implied relevant commonsense knowledge during interaction and taking into account each participant's emotional attitude. To deal with the aforementioned challenges, we design a multimodal off-the-shelf model that meets the requirements of real-life scenarios, specifically dyadic and multi-party interactions. We propose a Knowledge Aware Multi-Headed Network that integrates various sources including the dialog history and commonsense knowledge about the speaker and other participants. The weights of these pieces of information are modulated using a multi-head attention mechanism. The proposed model is learnt in a Multi-Task Learning framework which combines the ERC task with a Dialogue Act (DA) recognition task and an Emotion Shift (ES) detection task through a joint learning strategy. Our proposition obtains competitive and stable results on several benchmark datasets that vary in number of participants and length of conversations, and outperforms the state-of-the-art on one of these datasets. The importance of DA and ES prediction in determining the speaker's current emotional state is investigated.

CCS Concepts: • **Human-centered computing** → *Human computer interaction (HCI)*; • **Theory of computation** → **Machine learning theory**; • **Computer systems organization** → **Real-time system specification**; **Neural networks**.

Additional Key Words and Phrases: Emotion Recognition in Conversation, Multi-party Interaction, Multimodal Interaction, Affective computing

ACM Reference Format:
Sandratra Rasendrasoa[1*], Sebastien Adam[1], Alexandre Pauchet[2], and Julien Saunier[2]. 2022. Real-Time Multimodal Emotion Recognition in Conversation for Multi-Party Interactions. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3536221.3556601

1 INTRODUCTION

Emotions are a key element of human behavior, often defined as the mental state of an individual associated with thoughts, feelings and behavior. Identifying the emotional attitude of someone is crucial in the development of conversational agents, as affects have been shown to improve interaction with human users [15]. The increasing abundance in multimodal dialogues extracted from various sources, such as TV series [16] or scripted conversations [3] has opened the door to machine-learning contributions in the field of Emotion Recognition in Conversation (ERC). Multiple modalities such as visual and speech information can help disambiguate utterances, for example when irony

[**] : Contact Author, firstname.name[1]@univ-rouen.fr, [2]@insa-rouen.fr

is involved. Deep-learning approaches have become increasingly popular in this domain, due to their capability to compute and fuse representations from different modalities such as a video, sentence or audio clip. Our approach is part of the deep-learning field and tackles several ERC-related challenges. One of them is the ability to capture the contextual information associated with an utterance at a given time: for example, the sentence "*I got a call from the doctor*" can be either sad or happy depending on dialog history. Multiple propositions were made to inject contextual information representation either through recurrent or transformer based-networks [9, 13, 22, 27]. Recent improvements were made by combining the contextual and current utterance representation with additional source(s) of information. Knowledge bases such as ATOMIC [19] were exploited to convey commonsense knowledge to disambiguate the emotion associated to an utterance [7, 27].

A speaker's emotional state influences - and is influenced by - his communicative intent. The sentence "*I will think about it*" might have a different meaning, if the intent is to agree or disagree. Furthermore, someone who is angry is more likely to disagree with others, indicating that the dialogue act also depends on the user's emotional state. Thus, identifying the dialog act, which describes the communicative intent associated to an utterance, gives a hint to the user's emotional state. As stated in [18], most approaches in DAR focus on the textual modality. The work available in the literature is rather scarce when studying additional features such as the emotional state of the speaker, or facial expression. Another challenging task inherently associated to ERC is the emotion shift detection. It involves tracking the emotional state of a speaker throughout the conversation, and identifying the turns where the speaker's emotional state changes. Few approaches directly focus on that issue, improving instead the ERC general performance. ESD should be viewed as an auxiliary task for ERC, as ERC models have historically had issues identifying an abrupt change of emotion for a given party. Thus, being able to model label dependencies could help improve their performance. Injecting external commonsense knowledge has been observed [7] as helping in predicting emotion shift in a given speaker, and the use of Conditional Random Field has been proposed [14] as a possible benchmark.

In this article, we investigate the influence of dialogue act and emotion shift on the identification of a user's emotional state. Different modalities are combined with several sources of information to help predict the emotional attitude. The key contributions of this article are twofold : (1) we propose a Knowledge Aware Multi-Headed multimodal architecture which relies on multi-head attention to attend to multiple sources of information: dialog history, speaker's state and listener's state, and (2) through a joint learning strategy, we show the usefulness of considering Dialogue Act Recognition (DAR) and Emotion Shift Detection (ESD) to improve the identification of the emotional state of a user. Through these contributions, we are able to provide an off-the-shelf model which can be adapted to interactions where the number of participants and length of conversation vary, while holding true several criteria that are intrinsic to real-time ERC.

The remainder of this article is organized as follows. Section 2 introduces the criteria associated with real-time ERC and existing approaches. In Section 3, the overall architecture and the multi-task framework are described. Section 4 presents a series of experiments and results which validate our propositions. Finally, Section 5 concludes this article.

2 RELATED WORKS

Emotion recognition methods can be categorized into 4 types [2] [20] : (1) using physiological signals, which require extensive pre-processing in order to be usable, (2) using speech signals, by exploiting the variations and changes in audio signals, (3) using facial expression, where multiple neural networks are used together, and (4) using textual input, exploiting natural language processing. Deep-learning approaches have become popular in this domain, due to their capability to compute and fuse representations from different modalities such as a video, sentence, or audio clip [13] or handle contextual information [21].

In this section, we first define the criteria that enable real-time ERC in a data-oriented framework. Then, a description of the most recent deep-learning advances in ERC is given. Finally, we introduce the publicly available datasets used in ERC.

2.1 Real time emotion recognition in conversation

To be used during real-time interaction with artificial companions, emotion detection must meet several criteria:

(1) *Is the approach applicable in real time or not?* In the first case, learning is done from data received over time, thus only incorporating historical information to predict an emotional attitude. In the second case, both past and future utterances are used to predict the current emotion, *e.g.* to annotate automatically a corpus.
(2) *Does the approach consider the multi-party context?* ERC in a multi-party context is a complex problem compared to dyadic conversations, due to the difficulty of tracking speakers' mental states and processing co-references. Thus, being able to have stable performances while dealing with either dyadic or multi-party interactions is difficult. Furthermore, it also requires to consider whether the approach is speaker dependent or not. Creating a speaker representation, and by extension one for each participant has been discussed in [4] and [17].
(3) *Is the approach multimodal?* Depending on the tone of someone's voice, or his facial expression, the emotion associated to an utterance may vary. Thus, fusing multiple modalities can help distinguish emotional states.

2.2 Deep-learning in emotion recognition in conversation

Some of the existing approaches are only used in an offline setting, with applications in automatic annotation systems. BERT+MTL [10] is an offline uni-modal model founded on multi-task learning, with ERC as primary classification task and speaker identification (SI) as secondary. Multi-task learning has often been used in NLP, notably by [26] in question generation where the secondary task is to predict the next word in a sequence. The focus of ConGCN [25] is on the modeling of the context and of the speakers of a conversation. It models the whole conversation through a graph. The ERC task is performed offline with inferences on the graph nodes, via a Graph Convolutional Network. Although not applicable in real-time, those approaches provide relevant perspectives on social interaction representation and multi-task framework.

Speaker tracking can be omitted in multi-party interaction, thus being speaker-agnostic. It is the case for AGHMN [9], designed to perform the ERC task in real-time. It uses a custom recurrent attention cell and the notion of hierarchical memory. The model takes as input a sequence of statements from $t-k$ to t. The memory model relies on having the sequence from $t-k$ to $t-1$ as input for a bidirectional GRU and uses an attention function to attend important elements in the sequence. The memory model is then merged with the representation at time t to perform an inference. TODKAT [27] is also a speaker-agnostic model where the participants' states are not taken into account. It is a uni-modal model which employs commonsense knowledge extraction and topic representation to create a context and knowledge enriched embedding. The drawback of speaker agnostic models is that information about each participants might be lost while deducing the emotion focusing only on the utterance.

Some architectures rely on participants tracking, but only use the textual modality. For instance, COSMIC [7] is a uni-modal model which relies on the extraction of common sense notions from the participants' utterances and the representation of the context. The approach exploits COMET, a model which aims at generating new knowledge via a variant of the BERT architecture [1]. DialogXL [22] is also a uni-modal model founded on auto-regressive models such as XLNet [24]. XLNet is a Transformer [23] architecture that proposes an unsupervised learning via an objective function based on the permutations of the language tokens composing a given sequence. Dialog-XL proposes 4 attention mechanisms in parallel to model different information, by masking some inputs between two layers of the model. Thus, those models do not leverage the potential of using multiple modalities for real-time ERC.

Finally, while initially tested in a dyadic context, DialogueRNN [13] is the only work that completely answers the defined standards. It models via GRUs [5] 3 factors required for the recognition of emotional attitudes in real time: (1) the state of each party in the conversation, (2) the context given by past utterances, and (3) the emotion associated with previous utterances.

2.3 Datasets and summary of potential baselines

IEMOCAP, MELD and DailyDialog are the most widely used datasets to train and test models in the literature. They cover various situations, with different conversation length and number of participants.

- IEMOCAP [3] contains 12 hours of recordings of dyadic interactions between 10 actors. The actors play a script corresponding to a given situation, simulating a certain emotion. Each script usually amounts to a long sequence, averaging $\tilde{5}$0 utterances per conversation. Annotations for emotion are either angry, excited, frustrated, happy, neutral or sad.
- DailyDialog [11] records dyadic conversations about the daily life of the participants. Conversations are usually shorter that those in IEMOCAP, averaging 10 utterances per conversation. It provides dialogue act and emotion annotations among the following labels: anger, disgust, fear, joy, neutral, sadness or surprise.
- MELD is the most recent dataset for ERC: it contains multi-party dialogue recordings from the TV series Friends. Most conversations in the dataset are short, averaging 3 utterances per conversation, as they usually corresponds to short sequences from an episode. It is a multimodal corpus, which includes transcript associated with audio and video clips. The utterances were annotated with the same labels as DailyDialog.

Dialogue act annotations were also added on IEMOCAP and a subset of MELD by [18]. The different modalities and annotations makes those datasets versatile, as they cover different tasks such as DA recognition, sentiment analysis and ERC in real-time.

Table 1 summarizes the existing works according to the criteria required for real-time, multimodal, multi-party ERC, as well as the datasets they are tested on. To the extent of our knowledge, the state-of-the-art models (i.e Todkat and DialogXL) are unable to obtain stable results, relative to the conversation length and number of participants, on the aforementioned datasets: Todkat being the top performer on MELD and Dailydialog, which both have shorter sequences and DialogXL being better on longer sequences (IEMOCAP).

Model	Speaker representation	Offline vs online	Modalities	multi-party	Tested on
BERT+MTL [10]	✓	offline	Text	✗	MELD
ConGCN [25]	✓	offline	Text, audio	✓	MELD
AGHMN [9]	✗	online	Text	✗	IEMOCAP, MELD
TODKAT [27]	✗	online	Text	✗	IEMOCAP, MELD, DailyDialog*
COSMIC [7]	✓	online (variant)	Text	✓	IEMOCAP, MELD, DailyDialog
Dialog-XL [22]	✓	online	Text	✓	IEMOCAP*, MELD, DailyDialog
DialogueRNN [13]	✓	online (variant)	Text, audio, video	✓	IEMOCAP, MELD
KAMUH	✓	online (variant)	Text, audio, video	✓	IEMOCAP, MELD*, DailyDialog

Table 1. Existing approaches for ERC. * indicates that the approach is the state of the art on the dataset

3 PROBLEM SETUP AND METHODOLOGY

Given the transcript of a conversation as well as the associated audio and video clipsof each speaker/participant(s), the ERC task consists in identifying the emotion associated with each utterance from several predefined emotions. Formally, the input sequence of N utterances is $[(u_1, p_1), (u_2, p_2), .., (u_t, p_t), .., (u_N, p_N)]$, where u_t is the t-th utterance representation obtained from the combined textual, audio and video features associated to a sentence formulated by the speaker p_t. At a given turn t, the task is to predict the emotion e_t associated with each utterance u_t.

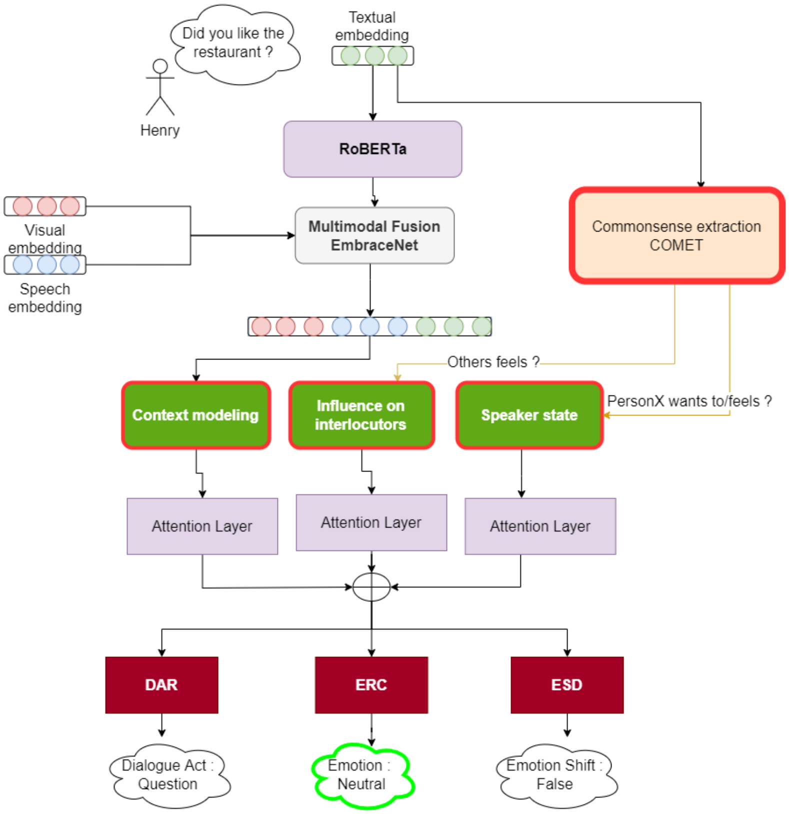

Fig. 1. Overview of the KAMUH architecture.

Figure 1 gives a general overview of the proposed Knowledge Aware Multi-Headed architecture (KAMUH), which is divided into several components:

- feature extraction from the text, video and audio inputs using pre-trained open source models;
- robust multimodal fusion through the EmbraceNet Network;
- commonsense knowledge and context modeling though memory banks and attention layers;
- joint-learning classifiers which provide several outputs : predicted emotion, dialogue act and the possible appearance of an emotion shift.

3.1 Feature extraction

At a given turn t, the model takes as input the following modalities: (1) the transcript of the current utterance, (2) the video clip associated, (3) the audio of the utterance. In order to extract relevant feature vectors, we employ open-source methods that use deep-learning based pre-trained models:

(a) RoBERTa [12] is a language model to extract a $e_t^{text} \in \mathbb{R}^{d_{text}}$ feature vector via the transcript, in the version from [7];

(b) COMET [1] is a commonsense model used to extract the speaker's intent $c_t^{intX} \in \mathbb{R}^{d_{intX}}$, the effect of the speaker's utterance on himself $c_t^{effX} \in \mathbb{R}^{d_{effX}}$ and the listeners' reaction to the utterance $c_t^{effO} \in \mathbb{R}^{d_{effO}}$, where $d_{text} = d_{intX} = d_{effX} = d_{effO}$;

(c) OpenCV toolkit enables to preprocess the video and extract a feature vector $e_t^{video} \in \mathbb{R}^{d_{video}}$; a pre-trained 3D-Convulotional Neural Network is used for feature extraction, as they were shown to have some success for video classification tasks;

(d) COVAREP and OpenSmile are used to extract an acoustic embedding $e_t^{audio} \in \mathbb{R}^{d_{audio}}$, based on features such as pitch tracking, speech polarity and spectral envelope. Specifically, we use the IS13-ComParE configuration file and a neural network to obtain the required feature.

3.2 Robust multimodal fusion

In a real life interaction, combining multiple inputs from different sources is a challenging task. Deficiency from a single sensor can lead to missing data from a modality, which is quite detrimental in inference if the approach cannot adapt to this constraint. We propose to use a robust multimodal fusion approach [6], that combines the audio, textual and video representations extracted from the steps described in 3.1. This component of our architecture is divided in two parts:

- the docking layer takes the modalities e_t^{audio}, e_t^{text} and e_t^{video} as inputs. The objective of this layer is to ensure that each embedding is converted to *dockable* vectors who share the same size. In practice, it applies a linear transformation to the incoming data. The output of this layer are d_t-dimensional vectors $dock^{(audio)}, dock^{(text)}, dock^{(video)}$, so that

$$dock^{(m)} = f_a(w^m e_t^m + b^m) \tag{1}$$

  with $dock^{(m)} \in \mathbb{R}^{d_t}$, $w^m \in \mathbb{R}^{d_t \times d_m}$ a learning parameter, b_m a bias, $m = \{audio, text, video\}$ a modality, d_m the size of the input embedding and f_a an activation function.
- the embracement layer then combines the 3 vectors from the docking layers in a so-called *embraced vector* $u_t \in \mathbb{R}^{d_t}$. The fusion mechanism is founded on a probabilistic approach:
  Let $\mathbf{r_i} = [r_i^{audio}, r_i^{text}, r_i^{video}]$ where $i \in 1, 2, .., d_t$ be a vector drawn from a multinomial distribution, i.e.,

$$\mathbf{r_i} \sim Multinomial(1, p), \tag{2}$$

  where $p = [p^{audio}, p^{text}, p^{video}]^T$ are the probability values which sum to 1. This effectively mean that when one value of $\mathbf{r_i}$ is equal to 1, the rest are 0. The Hadamard product, denoted by $\odot$ is employed in the following:

$$dock'^{(m)} = \mathbf{r^{(m)}} \odot dock^{(m)} \tag{3}$$

and we obtain the embraced vector $u_t = [u_t^1, u_t^2, ..u_t^{d_t}]$ from :

$$u_t^i = \sum_{k \in \{audio, text, video\}} dock_i'^{(m)} \tag{4}$$

This ensures that only one modality can influence the i^{th} component of u_t at a time. The final output still relays multimodal information since this process is done on each component of the vector independently. This process can also handle missing data at the embracement layer by adjusting the probabilities p, by making sure the chance of drawing a missing modality remains 0.

3.3 Memory representation

At a given turn, the objective of this component is to save in a memory cell the information associated to the current turn. Using the previously computed u_t and two vectors computed from COMET, u_t^{speak} and u_t^{listen}, memory banks are built to represent different information at each turn throughout the conversation. Built via deep neural recurrent networks, they are illustrated in Figure 1 by the three green boxes. As an example, a generic memory bank H sequentially encodes information in a memory cell h_t from a sequence $S = \{s_1, .., s_t, .., s_n\}$, s.t:

$$H = [h_1; ..; h_t; ..h_n] \tag{5}$$

$$h_t = GRU(h_{t-1}, s_t) \tag{6}$$

Thus, the content of a memory cell h_t inside H is the output of the recurrent network at a given turn t. In the remaining of the article, the notations associated with the generic memory bank H and sequence S will be used to illustrate the following operations. Our proposal leverages three memory banks, based on 3 elements:

- the conversation history, which encodes contextual information in the memory bank H_{hist} from the utterance representation u_t.
- the speakers' state history, which encodes commonsense knowledge about the intent and reaction of the speaker to its utterance, represented by u_t^{speak}. Its associated memory bank is named H_{speak}.
- the influence of the utterance on the interlocutors, represented by u_t^{listen}, is saved in the memory bank H_{listen}.

Encoding the context and commonsense knowledge about participants through recurrent networks is also done in DialogueRNN [13] and COSMIC [7]. However, these architectures are designed to have an entire recurrent network per participant in the interaction, which makes them progressively more complex as the number of participants increases. In our approach, instead of having a full representation of each listener, we group all the listeners information in a single memory bank. This allows us to both reduce the size of the memory banks and consider any number of participants. Thus, we are still taking into account the multi-party criterion, while also having a speaker representation at every turn.

3.4 Attention mechanism

As we require an aggregated representation for each memory bank, an attention mechanism is proposed to weight the importance of the memory cells in each bank. Thus, a multi-head attention mechanism [23] is exploited. The multi-head attention mechanism is based on the dot product attention with a softmax normalization and scaling factor. Given a memory bank $H = [h_1; ..; h_t; ..; h_N] \in \mathbb{R}^{N \times d_t}$ and an utterance $u_t \in \mathbb{R}^{d_t}$, an attention vector $\alpha_t \in \mathbb{R}^{d_{model}}$ and the attention score a_i associated to a memory cell $h_i \in \mathbb{R}^{d_t}$ can then be defined:

$$h_i^Q = tanh(W_\alpha h_i + b_\alpha) \tag{7}$$

$$a_i = \frac{e^{(h^Q)^T u_t}}{\sum_{i=1}^{t-1} e^{(h^Q)_i^T u_t}} \tag{8}$$

$$\alpha_t = \sum_{i=1}^{t-1} a_i h_i^V \tag{9}$$

where $W_\alpha \in \mathbb{R}^{d_{model} \times d_t}$ is the weight matrix and $b_\alpha \in \mathbb{R}^{d_{model}}$ the bias. We also define the Attention function s.t.:

$$\alpha_t = Attention(H^Q, s_t, H^V) \tag{10}$$

The rationale of using multi-head attention is that learning different linear projections of our inputs can be beneficial to the final representation. The idea is to map a query Q and a set of key-value pairs (K,V respectively) to an output.

$$MultiHead(H', s_t, H) = [head_1; ..; head_M]W^O \tag{11}$$

$$head_i = Attention(W_i^Q H^Q, W_i^K u_t, W_i^V H^V) \tag{12}$$

where M is the number of heads at a given layer, $W_i^Q \in \mathbb{R}^{d_{att} \times d_t}$, $W_i^K \in \mathbb{R}^{d_{att} \times d_t}$, $W_i^V \in \mathbb{R}^{d_{att} \times d_t}$ and $W^O \in \mathbb{R}^{d_{att} \times d_t}$ are projection matrices for each entry of the ith head, with $d_{att} = d_{model} \div M$.

3.5 Masked attention

As we distinguish the speaker and the listeners, not every cell in our memory banks can be attended by the attention mechanism. At a given turn t, the accessible content in each memory bank goes through a Masked Attention Layer, shown in Figure 1:

(a) For H_{hist}, the dialog history is considered.

(a) For H_{speak}, only the past memory cells where the current speaker was speaking are taken into account. That way, we ensure that only commonsense knowledge about that specific participant is the input of the attention function.

(a) For H_{listen}, the past memory cells from the other participants (speaker not included) are accounted for.

An attention mask with the same size as the attention weights matrix from the multi-head attention are summed to obtain a masked score matrix. This process enables real-time, as only information from dialog history is gathered.

3.6 Combining representations

The final representation r_t then incorporates each memory bank to our input, through a concatenation:

$$r_t = [u_t; MultiHead(H_{hist}^Q, u_t, H_{hist}^V); MultiHead(H_{speak}^Q, u_t^{spea}, H_{speak}^V); MultiHead(H_{listen}^Q, u_t^{listen}, H_{listen}^V)] \tag{13}$$

where H_{hist}, H_{speak} and H_{listen} are memory banks for conversation history, speaker and listeners respectively. Thus, r_t is the final representation.

3.7 Multi-task learning

We propose to employ a multi-task learning strategy to leverage the interdependence between ERC and two secondary tasks.

3.7.1 Dialogue Act Recognition (DAR). A dialogue act (DA) is a speech act which describes the communicative intent of a speaker's utterance : a question, statement or command are all examples of dialogue act. Communicative intents and emotional states are intuitively closely related as illustrated in the following example. *Utterance: I can't leave it! You gouged a hole in my dingy floor. DA: Disagreement, Emotion: Anger.* Here, the act of disagreeing is directly related to the speaker's angry emotional state, which supports the idea that identifying the correct DA is linked to predicting the speaker emotional state. Thus, we employ DAR as a secondary learning task where, given an utterance u_t at turn t, the objective is to identify the corresponding DA.

3.7.2 Emotion Shift Detection (ESD). Emotion Shift Detection focuses on identifying whether the emotional attitude of a speaker changes from her previous utterances. For instance, statistical analysis on the IEMOCAP and MELD datasets shows that emotion shifts occur on average 3.5 times and 17.5 times respectively, where the average conversation lengths are 9 and 52 respectively. A speaker is then very likely to shift from one emotion to another, even in short discussion. We take that into account through this joint learning framework. ESD is a binary classification task, where the labels are: 0 if the emotional attitude of the speaker do not change and 1 if it changes.

3.7.3 Loss Functions. As shown in Figure 1, we propose three classification layers in parallel which share the same representation r_t as input. A combined loss L of our 3 tasks is defined as the weighted sum of each individual loss:

$$L = L_{ERC} + aL_{DAR} + bL_{ESD} \tag{14}$$

where L_{ERC}, L_{DAR} and L_{ESD} are respectively the loss functions for ERC, DAR and ESD, a and b are hyper-parameters. The loss function that is used for ERC and DAR is the Negative Log-likelihood function and a Binary Crossentropy function is used for ESD. [8] proposed the use of ERC as a secondary task in order to solve the DAR task, but did not provide the performance of their approach on ERC. Thus, our approach allows to investigate the performance of ERC while coupled with DAR and ESD, which could lead to new perspectives, where datasets annotated either with DA or emotions are merged to improve the overall performance on both tasks.

4 RESULTS AND ANALYSIS

The experimental setup and results for KAMUH vs baselines are provided in this section. Source code is publicly available on github[1].

4.1 Experimental Setup

Our benchmark is made on multiple publicly available datasets, presented in Table 2: IEMOCAP [3], DailyDialog [11] and MELD [16] described in Section 2. In the case of DailyDialog, KAMUH and the baselines have been adapted for text-based emotion recognition. For training, we use Adam optimizer with learning rate = 0.0001 and l2 norm = 0.0004.

We have identified the best approaches which answer defined criteria. ConGCN and BERT+MTL are offline models, thus are not included as baselines. DialogueRNN and COSMIC are not natively used in online setting but can be easily

[1] https://github.com/tenihasina/ERC_Real_Time

adapted to. AGHMN and TODKAT do not consider the notion of speaker in ERC, and directly associates utterances with emotions, but can still be applied for real-time. Dialog-XL is a uni-modal model, but validates all other criteria.

Datasets	Train/dev	Test
Dialogues		
IEMOCAP	120/12	31
MELD (subset)	745/83	208
DailyDialog	11,118/1,000	1,000
Utterances		
IEMOCAP	4,810/1,000	1,523
MELD (subset)	6,740/748	2,500
DailyDialog	87,170/8,069	7,740

Table 2. Number of dialogues & utterances

AGHMN results are produced by running the code published by the authors. As COSMIC and DialogRNN were initially used in offline setting, adaptations were made, based on the published code. Thus, results in an online setting (no access to future utterances) are provided. Results from DialogXL could not be reproduced on IEMOCAP with the code published by the authors and, to the best of our knowledge, they have never been reproduced in the literature.

4.2 Comparison with Baselines

Experimental results of KAMUH are reported in Table 3. We evaluate KAMUH using weighted average-F1 as a metric for IEMOCAP and MELD and Macro and Micro F1 on DailyDialog, following the literature. Our proposed KAMUH architecture holds competitive results on the three datasets. Notably, it offers the most stable results compared to Dialog-XL and TODKAT, the respective state of the art on IEMOCAP and MELD respectively. Paired t-test is performed to test the significance of the difference between two approaches, with a default significant level of 0.05. Average improvements over the baseline approaches are all significant, $p - value < 0.05$.

- KAMUH outperforms Dialog-XL on MELD (4th in the ranking) and on DailyDialog (5th in the ranking) by a range of 2-4%, and ranks 2nd behind the non reproducible results on IEMOCAP, while still being inside the error margin.
- Our approach also outperforms TODKAT, state-of-the-art on MELD by 0.53% and on IEMOCAP (4th in the ranking), by 4%, while also ranking 2nd on DailyDialog, being below TODKAT by a margin of 1%.

As stated in [22], the XLNet architecture specializes on dealing with longer sequences (SoTA on IEMOCAP), while TODKAT [27] seems to perform better on shorter sequences (SoTA on MELD and DailyDialog).

- KAMUH offers an alternative to Dialog-XL, which mainly relies on the pre-trained architecture, by using additional sources of information (commonsense + dialog history) and annotations to obtain comparable performance.
- KAMUH is also viable on shorter conversation (MELD & DailyDialog), where emotion is harder to identify due to a shorter dialog history. Additional information from knowledge bases seems to provide an improvement, as commonsense models like COSMIC, KAMUH and TODKAT are the top-3 performers on those datasets. KAMUH can be distinguished from those two as it employs a joint-learning framework and a different attention mechanism, which hold competitive result on those datasets.

Datasets	IEMOCAP	MELD	DailyDialog	
Models	Weighted Avg-F1		Macro F1	Micro F1
AGHMN	61.05 ± 1.64	57.6 ± 0.8	47.8 ± 0.2	54.7 ± 0.2
$COSMIC_v$	63.7 ± 0.2	63.2 ± 0.3	50.5 ± 0.1	56.1 ± 0.2
$DialogueRNN_v$	59.9 ± 0.5	57.9 ± 0.6	48.5 ± 0.3	55.2 ± 0.2
Dialog-XL	_65.8_*	62.41	-	54.93
TODKAT	61.3 ± 0.1	**65.47**	**52.6±0.2**	**58.5±0.1**
KAMUH	**65.7 ± 0.3**	**66.0 ±0.1**	**51.5±0.2**	**57.3±0.2**

Table 3. Baseline comparison on ERC task. Performance of KAMUH on IEMOCAP, MELD and DailyDialog (average of ten runs). Top-2 values are highlighted. $COSMIC_v$ and $DialogueRNN_v$ are unidirectional variants to validate the real-time criteria.
* Taken from [22], as the results could not be reproduced.

- We also leverage multimodal information, which is not the case for the current state-of-the-art, and make large improvements on the only multimodal baseline, DialogueRNN.

We argue that KAMUH offers more flexibility in a real-world scenario as it provides a better compromise for an off-the-shelf usage, where the number of participants and the dialogue duration can significantly vary, which is not the case for the studied baselines.

4.3 Ablation Study

4.3.1 Number of heads on the Multi-head attention component. We have studied the impact of the number of head on KAMUH by varying the number of heads between 1 and 6. On IEMOCAP and MELD, best results are attained for $nb\_head = 4$ and on DailyDialog for $nb\_head = 3$.

4.3.2 Secondary tasks. A GridSearch was performed to identify the value of a and b in the Loss L equation. The value of a and b were in the following range: $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. Best performances are achieved for IEMOCAP with $a = 0.7, b = 0.5$, MELD with $a = 0.7, b = 0.7$ and DailyDialog with $a = 0.7, b = 0.7$.

An ablation study is then performed to evaluate the impact of multi-task learning, shown in Table 4. Three variants are provided for the ablation study: $KAMUH_{DAR}$, $KAMUH_{SER}$ and $KAMUH_{SER,DAR}$. They respectively correspond to the original KAMUH without DAR, SER and DAR+SER.

- On IEMOCAP, removing the DAR task leads to 0.4 drop in performance, while removing the SER only make the performance drop by 0.1 pts. Dropping both secondary tasks leads to a 0.9 pts drop in performance.
- On MELD, removing the DAR, SER and SER+DAR tasks lead respectively to a drop of 0.5, 0.3 0.9 in performance.
- On DailyDialog, removing the DAR, SER and SER+DAR respectively lead to a drop in MAcro-F1 of 1.4, 0.2 and 2.0, while it leads to a drop of 1.1, 0.3 and 1.3 respectively on Micro-F1.

All across the board, discarding the DAR task leads to a bigger drop in performance than discarding SER. Removing both tasks leads to the biggest drop in performance, which demonstrates the value of including both tasks.

4.4 Case Study

A case study on a test conversation instance from MELD is illustrated in Fig 2. The most attended memory cells in each memory banks are shown, alongside the predicted emotion, dialogue act and emotion shift for the current utterance. To

Datasets	IEMOCAP	MELD	DailyDialog	
Models	Weighted Avg-F1		Macro-F1	Micro-F1
KAMUH	65.7 ± 0.3	66.0 ± 0.1	51.5 ± 0.2	57.4 ± 0.2
✗*DAR*	65.3 ± 0.1	65.5 ± 0.2	50.1 ± 0.1	56.3 ± 0.2
✗*SER*	65.6 ± 0.3	65.7 ± 0.1	51.3 ± 0.2	57.1 ± 0.2
✗*SER, DAR*	64.8 ± 0.4	65.1 ± 0.2	49.5 ± 0.2	56.1 ± 0.2

Table 4. Ablation study for secondary tasks, average on 10 runs.
✗indicates that the model does not use the secondary task

facilitate the memory cells interpretation, discrete sequences for commonsense knowledge (top-k events) alongside most attended past utterances are presented.

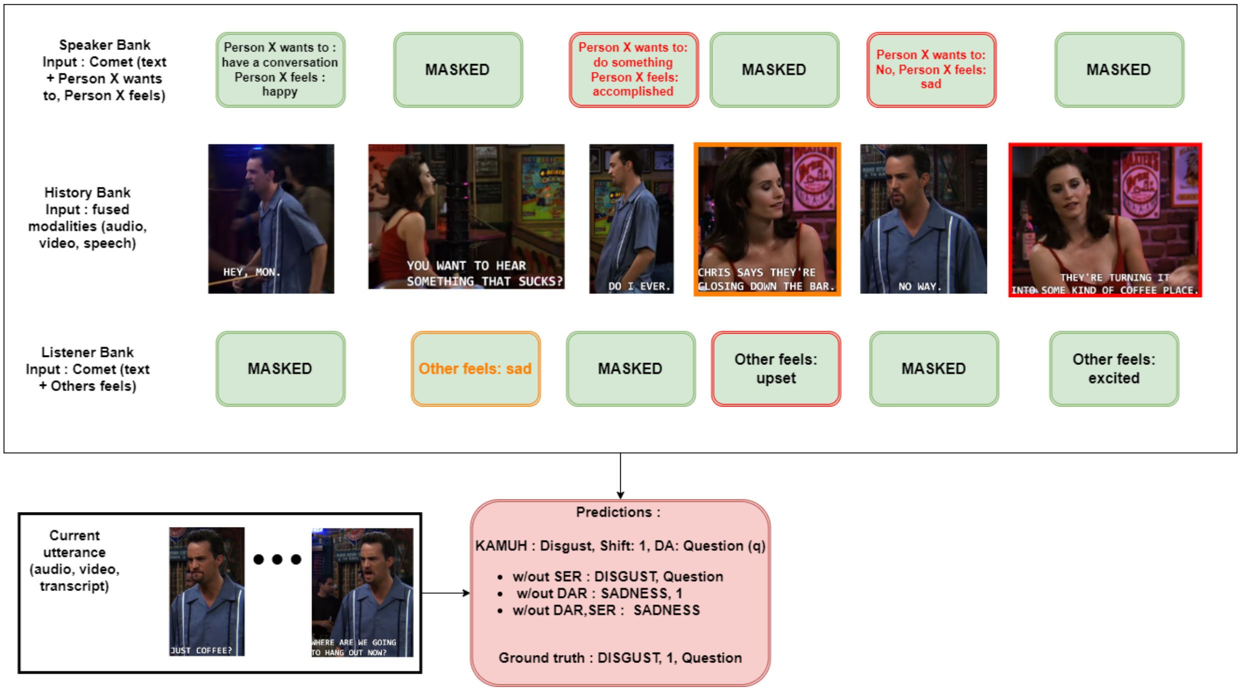
Speaker Bank
Input : Comet (text + Person X wants to, Person X feels)
Person X wants to : have a conversation Person X feels : happy
MASKED
Person X wants to: do something Person X feels: accomplished
MASKED
Person X wants to: No, Person X feels: sad
MASKED
History Bank
Input : fused modalities (audio, video, speech)
HEY, MON.
YOU WANT TO HEAR SOMETHING THAT SUCKS?
DO I EVER.
CHRIS SAYS THEY'RE CLOSING DOWN THE BAR.
NO WAY.
THEY'RE TURNING IT INTO SOME KIND OF COFFEE PLACE.
Listener Bank
Input : Comet (text + Others feels)
MASKED
Other feels: sad
MASKED
Other feels: upset
MASKED
Other feels: excited
Current utterance (audio, video, transcript)
JUST COFFEE?
WHERE ARE WE GOING TO HANG OUT NOW?
Predictions :
KAMUH : Disgust, Shift: 1, DA: Question (q)
w/out SER : DISGUST, Question
w/out DAR : SADNESS, 1
w/out DAR,SER : SADNESS
Ground truth : DISGUST, 1, Question

Fig. 2. Case study on a MELD instance. (1) In the top part, each line corresponds to a memory bank used by KAMUH (Speaker and listener) For the Speaker and Listener Bank, we show the output of COMET, given the corresponding utterance in a given row.
(2) In the bottom part, the objective is to show which memory cells are important to make the prediction on the current utterance. Attention scores inside memory banks are illustrated via a color gradient: orange being a higher score and red the highest score. MASKED cells corresponds to cells related to a different speaker in the Speaker bank and cells related to the current speaker in Listener bank.

This case study illustrates the ability of the model to focus on different turns depending on the memory bank. Notably, KAMUH focuses on perceived negative states in the speaker and listeners banks and predicts negative emotions for the current utterances. In the case of MELD, we have noticed that the "Question" DA is mostly associated with Neutral, Disgust, Surprise, and Frustration. This helps to predict disgust instead of sadness when the joint DAR task is included. The model also correctly predicts the emotion shift in Chandler's case in the current utterance, which can be explained by the commonsense knowledge from the Speaker and Listener banks.

5 CONCLUSION

In this article, we propose a novel architecture which investigates the importance of emotion shift and dialogue act for ERC by leveraging a multi-task framework. The proposed framework takes into account, by using multi-head attention, various sources of information, including the historical context and commonsense knowledge about the speaker and the other participants.

Real-time related criteria are considered in the design of the architecture. It can be observed that KAMUH natively provides the best compromise to the defined criteria: (1) learning is done from current and past utterances, (2) it is able to represent each speaker at a given turn through its memory bank, (3) a compromise is made for multi-party, as the participants who listen to the speaker share the same representation, (4) multimodal inputs are fused to help refine the utterance representation. While holding to those criteria, KAMUH has stable and competitive results on several benchmarks datasets, even outperforming the state-of-the-art on multi-party conversations.

As an off-the-shelf model, KAMUH offers more freedom of choice in the design of a real-life scenario, by removing the constraints of either switching to a better fit to a use case, or adapting the scenario to specifically work with a model. Further analysis on the fusion methods and a real-life scenario where our proposed model is embedded in a conversational agent are the next step in these studies.

REFERENCES

[1] Rashkin H. Sap M. Malaviya C. Celikyilmaz A. Choi Y. Bosselut, A. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *ACL 2019*. 4762–4779.

[2] Scott Brave and Cliff Nass. 2007. Emotion in human-computer interaction. In *The human-computer interaction handbook*. CRC Press, 103–118.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335–359.

[4] Hazarika D. Poria S. Hussain A. Subramanyam R. B. V. Cambria, E. 2017. Benchmarking multimodal sentiment analysis. In *CICLing*. Springer, 166–179.

[5] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[6] Jun-Ho Choi and Jong-Seok Lee. 2019. EmbraceNet: A robust deep learning architecture for multimodal classification. *Information Fusion* 51 (2019), 259–270.

[7] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. *arXiv preprint arXiv:2010.02795* (2020).

[8] A. Hosseinpanah G. 2020. *Investigating the Effect of User Variables on Perceiving the Emotional Nonverbal Behaviors of an Empathic Virtual Agent.* Ph. D. Dissertation.

[9] Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8002–8009.

[10] Jingye Li, Meishan Zhang, Donghong Ji, and Yijiang Liu. 2020. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *arXiv preprint arXiv:2003.01478* (2020).

[11] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957* (2017).

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

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

[14] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, and A. Gelbukh. 2019. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems* 34, 3 (2019), 38–43.

[15] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[16] S. Poria, D. Hazarika, N. Majumder, Gautam Naik, E. Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *ACL 2019*. 527–536.

[17] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access* 7 (2019), 100943–100953.

[18] Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4361–4372.

[19] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A Smith, and Y. Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI Conference on Artificial Intelligence*, Vol. 33. 3027–3035.

[20] Anvita Saxena, Ashish Khanna, and Deepak Gupta. 2020. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems* 2, 1 (2020), 53–79.

[21] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2020. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *arXiv preprint arXiv:2012.08695* (2020).

[22] W. Shen, J. Chen, X. Quan, and Z. Xie. 2021. DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition. In *AAAI Conference on Artificial Intelligence*, Vol. 35. 13789–13797.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

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

[25] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations.. In *IJCAI*. 5415–5421.

[26] W. Zhou, M. Zhang, and Y. Wu. 2019. Multi-Task Learning with Language Modeling for Question Generation. In *EMNLP-IJCNLP*. 3394–3399.

[27] L. Zhu, G. Pergola, L. Gui, D. Zhou, and Y. He. 2021. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In *ACL - IJCNLP*. Association for Computational Linguistics, Online, 1571–1582. https://doi.org/10.18653/v1/2021.acl-long.125