Cognitive Computation
https://doi.org/10.1007/s12559-019-09704-5

Emotion Aided Dialogue Act Classification for Task-Independent Conversations in a Multi-modal Framework

Tulika Saha[1] · **Dhawal Gupta**[1] · **Sriparna Saha**[1] · **Pushpak Bhattacharyya**[1]

Received: 30 August 2019 / Accepted: 20 November 2019

Abstract

Dialogue act classification (DAC) gives a significant insight into understanding the communicative intention of the user. Numerous machine learning (ML) and deep learning (DL) approaches have been proposed over the years in these regards for task-oriented/independent conversations in the form of texts. However, the affect of emotional state in determining the dialogue acts (DAs) has not been studied in depth in a multi-modal framework involving text, audio, and visual features. Conversations are intrinsically determined and regulated by direct, exquisite, and subtle emotions. The emotional state of a speaker has a considerable affect on its intentional or its pragmatic content. This paper thoroughly investigates the role of emotions in automatic identification of the DAs in task-independent conversations in a multi-modal framework (specifically audio and texts). A DL-based multi-tasking network for DAC and emotion recognition (ER) has been developed incorporating attention to facilitate the fusion of different modalities. An open source, benchmarked ER multi-modal dataset **IEMOCAP** has been manually annotated for its corresponding DAs to make it suitable for multi-task learning and further advance the research in multi-modal DAC. The proposed multi-task framework attains an improvement of **2.5%** against its single-task DAC counterpart for manually annotated **IEMOCAP** dataset. Results as compared with several baselines establish the efficacy of the proposed approach and the importance of incorporating emotion while identifying the DAs.

Keywords Dialogue act · Emotion · Multi-task · Classification · Multi-modal · Deep learning

Introduction

Dialogue act classification (DAC) forms one of the significant steps for understanding a user's utterance in any dialogue system. It provides a valuable insight into determining the communicative intention of the user and is represented as a function of the speaker's utterance. Thus, DA aims towards identifying the pragmatics of human conversation instead of just its literal meaning. The task of DAC is of exemplary importance as its ability to automatically detect discourse structure assists towards building intelligent dialogue systems, transcription of conversational speech, and so on. Extensive amount of works have been done for modeling the task of DAC employing ML [1–3] and DL [4–7] approaches for identifying the DAs with increased accuracy and precision.

However, any conversation involving humans necessitates a detailed analysis of its emotional state. *Emotion* is primarily defined to be the complex reaction of the brain to a stimulus, whether external (e.g., something I see or hear) or internal (e.g., thoughts, memories, imagination) [8]. Whereas *sentiment* is the combination of emotion and thought [8]. It is when we put a name to an emotion and decide how we react to it, e.g., positively and negatively. *Affect* on the other hand is considered to be the super-set in psychological literature that embodies emotion, sentiment, and feelings and is ontologenetically primitive than emotions [9]. So, whatever a human communicates (either through speech or text) essentially has an emotion implicitly attached to it. Language is primarily used as an indicator of communicative intention as well as emotion. Conversations are intrinsically determined and regulated by direct, exquisite, and subtle emotions. The emotional state of a speaker has a considerable affect on its intentional content or simply on its pragmatic content [10]. This hypothesis is

✉ Tulika Saha
sahatulika15@gmail.com

Dhawal Gupta
dhawal.gupta.iitp@gmail.com

[1] Indian Institute of Technology Patna, Bihta, Bihar, India

Published online: 22 January 2020

rather intuitive as emotions are extracted from and echoed through the pragmatic or intentional content of the speaker. Emotions affect intention or pragmatic and vice versa. For example, an utterance such as "*Okay* or "*Right*" can be interpreted as "agreement" or "disagreement" if implied sarcastically. But the speaker's emotional state may contain information that gives it a different meaning. In the case of expressive DAs such as "greeting", "thanking", "apologizing", emotion of the speaker can help in identifying correct communicative intentions and vice versa. Thus, it is essential to consider the emotional state of the speaker while determining its DA. Independently, there exists abundance of works which address the effect of emotion by recognizing the user's emotions [11–14], etc. and adapting the virtual agents (VAs) to behave accordingly [15–17], etc. Certainly, there exists very little work which addresses the affect of emotion while determining the communicative intention of the speaker [18, 19], as DAs primarily determine the flow of any conversation (either human-human or human-computer).

Several research investigations have established the advantage of exploiting the combination of text and non-verbal behaviors (acoustic and visual modalities) used by humans [11, 20, 21]. The primary benefit of incorporating other modalities with text is the inclusion of behavioral cues present in acoustic (vocal modulations) and visual (facial expression) modalities. The different modalities in conjunction provide crucial cues to better identify the communicative intention and emotional state of the speaker. Thus, a combination of different modalities will help in developing more robust DAC models.

In this paper, we investigate the role of emotions in determining the DA of the user utterance in task-independent conversations by exploiting the combination of vocal modulations and text. Thus, we implement a DL-based multi-tasking framework to model the identification of DAs with the help of emotion. The goal of this study is to analyze the affect of emotion in the automatic identification of the DAs. Thus, this work does not explicitly focus on improving the performance of the emotion recognition (ER) task from multiple modalities. Hence, the analysis and findings are reported only for the task of DAC. An-open source, benchmark ER multi-modal dataset **IEMOCAP** [22] has been used for this purpose. Since this dataset does not contain pre-defined tagged DAs, so, it has been developed manually for the joint tasks (DAC and ER). The proposed framework has been compared against several baselines and the results are reported accordingly.

The key contributions of this paper are the following:

- This paper carries a detailed investigation to analyze the affect and role of emotion for determining the communicative intention of the speaker.
- A DL-based multi-tasking network has been developed in a multi-modal framework (acoustic and text) incorporating attention to facilitate the fusion of different modalities. This is done in order to exploit the shared representation obtained by learning both the tasks (DAC and ER) jointly, solely for the automatic identification of DAs.
- A DA-annotated dataset has been developed manually by utilizing an open-access, standard ER based multi-modal corpus **IEMOCAP** which is now apt to perform both the tasks jointly and advance the research in the field of multi-modal DAC.

Related Works

This section provides a brief description of the works done so far on DAC and ER followed by the motivation behind solving this problem.

Background

Dialogue Act Frameworks Identification of DAs is an age-old task with some of its standard proposed approaches dating back to late 1990s [23, 24] and early 2000s [2, 25]. However, majority of the works done to date on DAC are based on the chat transcripts available for the dialogue conversation, leveraging from only the textual modality. This is primarily because of the unavailability of an open-source multi-modal dataset for the task. Some of the benchmark works include those of [2], wherein the authors employed a range of techniques such as decision trees, hidden Markov models, and neural networks on the switch-board (SWBD) [26] dataset. Similarly, authors of [25] used a naive Bayes approach to solve the task of dialogue act classification. In [27], authors applied a stacked long short-term memory (LSTM) network–based approach to identify speech acts in a dialogue conversation. In [28], a convolutional and recurrent neural network–based approach was developed to identify dialogue acts. Most recent works include that of [29], where the authors built a hierarchical neural network–based approach using bi-directional LSTM to capture word-, utterance-, and conversation-level features and passed it through a CRF layer to incorporate utterance-level dialogue acts for the task. The authors of [30] proposed a contextual self-attention framework fused with hierarchical recurrent units to formulate a sequence label classifier. In [31], the authors presented a mechanism to capture context of long-range interactions ranging over a sequence of utterances using convolutional recurrent neural network–based approach. Other significant works include those of [3, 4, 6, 32–34]. However, all these works identify communicative intentions of the speaker in a dialogue

conversation by exploiting only the textual modality without the use of emotional state.

Emotion-Aware DAs The affect of emotional state for determining the DAs has also not been studied in greater detail in a multi-modal framework with very little work existing in the literature. In [19], the authors particularly exploited emotions for disambiguating dialogue acts (in case of any confusion). They demonstrated their proposed work in a small-scale *Atomix* game particularly task-oriented conversation. Their approach used a Bayesian network to recognize user's emotion and then use that information to resolve ambiguities in dialogue acts (if any). However, their approach utilized only the physiological feedbacks of the user to model emotion and were used only to disambiguate DAs in selected scenarios. In [35], the authors studied the affect of incorporating facial features as means of recognizing emotion to identify DAs. They demonstrated their work for tutorial dialogue session (task-oriented) and employed logistic regression as the classifier to model DAs. However, they considered only the cognitive-affective states such as *confusion* and *flow* as the emotion categories to learn DAs for these particular emotional states. In [18], authors studied the role of affect analysis in DA identification for an unsupervised DAC model. They used lexicon-based features from *WordNet Affect* and *SentiWordNet* to map those with emotion categories and finally include those as features to model the DAs in their unsupervised LSA-based approach. They reported from their findings that a relationship does exist between the affective lexicon and the DA of an utterance; however, they could not establish whether affect analysis plays a role in DAC. The authors of [36] also studied the affect of emotions and intention mediated with DAs for an *in-game* Japanese conversation. Their aim was to establish DA-emotion pairings from the pre-annotated task oriented corpus. They used a *normalized pointwise mutual information* (npmi) score for each DA-emotion pair to find until what extent the occurrence of a given DA would indicate the observation of a given emotion and vice versa. They then employed k-means clustering with bootstrap resampling of the npmi score to find associated and disassociated DA-emotion pairs. However, such strict association or disassociation amongst DA-emotion pair may not truly generalize to real life conversations. In [37], the authors aimed to study the affect of sentiment in identifying DAs in Twitter-based dialogue conversations. They employed a hierarchical recurrent network–based multi-tasking approach to learn sentiment and DA jointly for a Twitter-like dataset collected from *Mastodon*. However, sentiment is a more coarse-grained reflection of the user's state of mind than emotion and conversation on twitter does not generalize well to real-time conversations because of the noisy and limited character length of the former platform [38].

Emotion-Based Frameworks Likewise, there also exist a wide range of works that identify the emotion of the user and utilize that information to help VAs or chat assistants behave or prepare strategies accordingly and generate emotion-aware responses. The authors of [39] proposed a novel method to generate word embeddings using an extreme learning machine approach. They highlighted the efficiency of the framework in sentiment analysis and sequence labeling task against widely used embeddings such as GloVe and Word2Vec. In [40], the authors examine the relationship between the optimal feature selection against the sentiment classification performance. For this, they proposed a chi-square-based feature selection algorithm using a preset score threshold. In [41], the authors proposed a DL model for emotion-aware response generation. Given a Chinese post and a user-specified emotion category, the task was to generate a response that was coherent with the emotion category. They proposed a long short-term memory–based encoder-decoder framework with emotion intelligence as an external knowledge to the model to produce content and affect enriched responses. In [42], the authors proposed a mobile-based conversational agent that focused on developing a context-aware multi-modal VA that dynamically incorporates users' requirements and preferences from the environment in order to provide personalized service. Their proposed framework was developed as an Android app in healthcare domain for older adults suffering from Alzheimer's disease. In [43], the authors reviewed various types of computational models of emotions from the perspective of their development as one of the challenges in the development of VAs is its capability of exhibiting very believable and human-like behaviors and emotions. They primarily investigated five design aspects that influence their development process: theoretical foundations, operating cycle, interaction between cognition and emotion, architectural design, and role in cognitive agent architectures. They discussed about the key issues and challenges in the development of emotion-aware VAs and suggested research directions that may lead to more robust and flexible designs for this type of computational model. However, all these works are focused on generating VAs' response based on the emotion of the user. The work described in this paper is focused on understanding the users' real pragmatic content/intention conditioned on the emotional state of mind of the user.

Motivation

It is clearly evident from the existing literature that majority of the works done in DAC do not incorporate emotional state

of the speaker and other non-verbal cues as means to model the task. These features/information provide significant knowledge to understand the real intention/pragmatic of the user content by delving into the state of the mind of the speaker. Also, there is clearly a dearth of an open-source dialogue dataset for the task of DAC in a multi-modal framework, let alone a multi-task, multi-modal dataset for DAC and ER. Also, the existing approaches to date, discussed above, propose algorithms for restricted scenarios and task-oriented conversations with various assumptions that do not generalize well to real-life conversations. Motivated by the inadequacy of the existing systems and approaches, this paper presents a DL-based multi-task model to study the affect and role of emotion for automatically determining the DA of the speaker in a multi-modal framework.

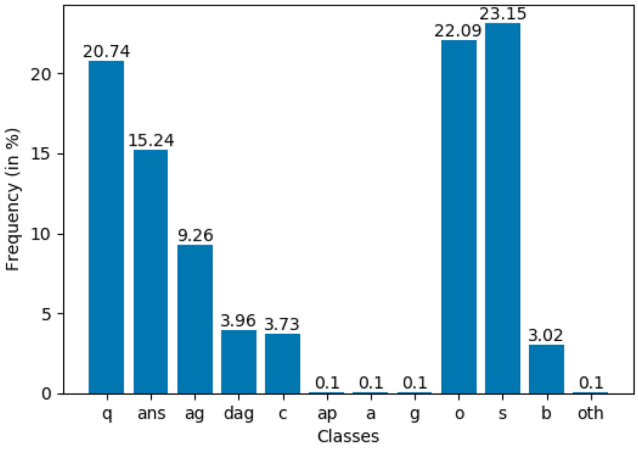

Fig. 1 Distribution of the DA tag-set in the IEMOCAP dataset

Dataset

IEMOCAP *Interactive Emotional Dyadic Motion Capture Database* [22] is an open-sourced, multi-modal ER dataset. It contains 151 videos of recorded dialogues, with 2 speakers per session for a total of 10 speakers in a two-way conversation segmented into utterances amounting to a total of 302 videos across the dataset. The dataset contains task-independent conversations and the medium of conversations in all the videos is English. Each utterance is annotated for the presence of 10 emotions namely, *fear*, *sad*, *angry*, *frustrated*, *excited*, *surprised*, *disgust*, *happy*, *neutral*, and *others*. However, this available dataset is not annotated for its corresponding DAs.

Benchmark-available datasets for DAC such as switchboard (SWBD) [26], ICSI meeting recorder [44], and TRAINS [45] contain only text-based conversations without any emotion tags. HCRC map task corpus [46] does contain audio modality along with the chat transcript but the corpus itself has task-oriented conversations and does not contain any emotion labels. To the best of our knowledge, we were unaware of any sizable and open-sourced DA and emotion annotated multi-modal dialogue data at the time of writing; thus, IEMOCAP dataset has been manually annotated for its corresponding DAs to make it suitable for developing a multi-task framework capable of learning DA and emotion for an utterance jointly and facilitate and advance research in the field of multi-modal DAC.

SWBD-DAMSL tag-set consisting of 42 DAs conceived by [47] for task-independent dialogue conversation such as SWBD corpus is used widely over the years for the task of DAC. Thus, SWBD-DAMSL tag-set has been used as the basis for conceptualizing tags for the IEMOCAP dataset since both the datasets contain task-independent conversations. Out of the 42 DAs of the SWBD-DAMSL tag-set, 12 most commonly occurring tags have been used to annotate utterances of the IEMOCAP dataset. The reason for choosing 12 tags is the limited length of the IEMOCAP dataset as compared with the SWBD corpus. It is highly likely that most of the tags of the SWBD-DAMSL tag-set will not appear in the IEMOCAP dataset because of less number of utterances and lower diversity in terms of occurrence of such fine-grained tags. The 12 frequently occurring chosen tags are *Greeting* (g), *Apology* (ap), *Command* (c), *Question* (q), *Answer* (ans), *Agreement* (ag), *Disagreement* (dag), *Statement-Opinion* (o), *Statement-Non-Opinion* (s), *Acknowledge* (a), *Backchannel* (b), and *Others* (oth). Three annotators graduate in English linguistics were assigned to annotate the utterances by only viewing the video available with the appropriate DAs out of the 12 chosen tags. The inter-annotator score with more than 80% was considered as reliable agreement. It was calculated based on the count that for a given utterance more than two annotators agreed on a particular tag. Figure 1 shows the distribution of the 12 DAs in the IEMOCAP dataset.[1] Some sample utterances with the corresponding DAs and emotion tags from the IEMOCAP dataset are shown in Table 1.

Proposed Methodology

In this section, we present the proposed approach describing the uni-modal feature extraction techniques and the developed architecture for the multi-task framework. The overall architectural diagram of the proposed network is shown in Fig. 2.

[1]The DA annotated dataset will be made publicly available for the research community.

Table 1 Sample utterances from the dataset with the corresponding DAs and emotion tags. F represents the female speaker and M represents the male speaker

Speaker	Utterance	DA	Emotion
F	*It was a hundred a sixty dollars. I'm sorry, uh, sixteen dollars.*	s	neu
M	*That's very amusing indeed.*	dag	ang
M	*Well, you know I appreciate you coming over and talking to me, I mean it definitely helps*	a	sad
F	*I know. I made such a mess of it, the entire time. The last three things I've done,*	ag	exc
	I've ruined everything, I think. Mm-hmm. Yeah.		
M	*Oh, it will, obviously you know. But I know it's something -*	ag	hap
	I don't know, you never really think about it happening, like before hand.		
M	*Well, I am thinking that way.*	dag	fru
F	*So let's go home.*	c	fru

Extracting Uni-modal Features

Here, we explain the textual and audio feature extraction methods. So, firstly, uni-modal features are extracted from each utterance separately.

Textual Features The transcripts of each video form the source of the textual modality. To extract textual features, a convolutional neural network (CNN) [48]–based approach is used. Pretrained GloVe [49] embeddings trained on the CommonCrawl corpus of dimension 300 have been used to represent words as word vectors. The resultant word embeddings are fed to the CNN layer having two kernels of size 3 and 4, with 64 feature maps each. The resultant outputs from individual channels are concatenated and passed through a fully connected layer of size 400 with activation function as *ReLU*. Thus, the CNN layer learns abstract representation of the phrases reflecting its semantic meaning which finally spans over to the entire sentence.

Audio Features To extract the audio features, *openSMILE* [50], an open-source software which automatically extracts audio features such as voice intensity and pitch has been used. The features extracted by openSMILE include 12 Mel-frequency cepstral coefficients, glottal source parameters [51], maxima dispersion quotients [52], several low-level descriptors (LLD) such as voice intensity, MFCC, voiced/unvoiced segmented features [53], perceptual linear predictive cepstral coefficients [54], psycho-acoustic sharpness [55], spectral harmonicity [56], pitch and their statistics (for example, root quadratic mean, mean), and voice quality (for example, jitter and shimmer).

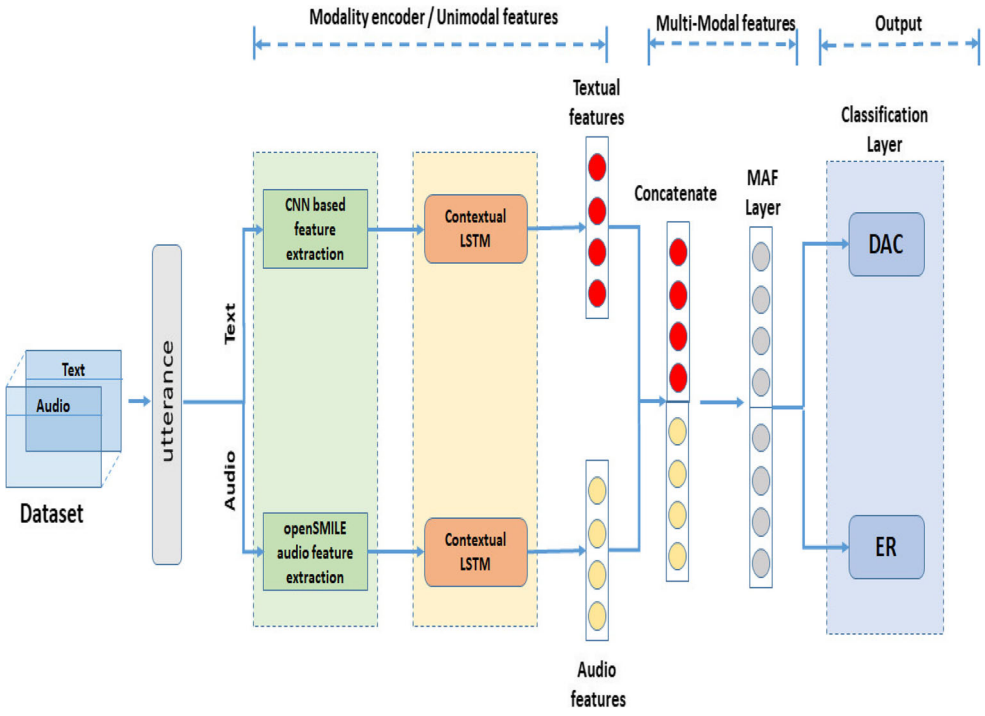

Fig. 2 Overall architectural diagram of the proposed network

Network Architecture

The proposed network architecture consists of three main components : (i) *modality enocoders* which typically takes as input the uni-modal features and outputs the modality encodings, (ii) *a modality attention fusion subnetwork* that fuses the individual modalities, and (iii) *classification layer* that encompasses outputs of both the tasks (DAC and ER) to be learned jointly conditioned on the output of the modality attention fusion subnetwork. The proposed architecture takes as input the elements of all the individual modalities of time series (TS) data where each modality is a two-dimensional matrix and the rows of each such matrix comprises of time-stamped feature vectors of that modality. Below, we explain each of the components of the proposed network. The detailed architectural diagram of the proposed network is shown in Fig. 3.

Modality Encoders The proposed network has n encoders for n individual modalities of TS data to be encoded. The encodings of the individual modalities are obtained from Eq. 1.

$$EC_i = MEC_i(TS_i : w_i) \tag{1}$$

where EC_i represents encoding obtained for modality i, TS_i is the time series data for modality i. MEC_i represents the modality encoding network for modality i and w_i is the set of weight parameters of the MEC_i network. The MEC network is basically a bi-directional long short-term memory (Bi-LSTM) network [57] with forget gate employed individually for all the different modalities. Here, Bi-LSTM is used as the modality encoder network because of its capability to capture long-term dependencies in TS data. Thus, the contextual LSTM captures the context of the words within a particular utterance along the TS data. The MEC vectors obtained from the individual modalities are then fed to the modality attention fusion subnetwork for further processing.

Modality Attention Fusion Subnetwork An attention mechanism typically guides the network to focus on the most important features of an object which are relevant for classification. Network with attention layer shows improvement in results compared with the non-attention-based counterparts. Not all the modalities are equally important for multi-task classification. This motivated us to introduce an attention mechanism in our network, termed as modality attention fusion (MAF) layer, which takes an audio and text modalities as input and returns attention score for each modality as an output.

At first, the outputs of the modality encoders are concatenated together vertically. Let $M = [M_t, M_a]$ be the concatenated feature set, where $M_a =$ output of the acoustic modality encoder and $M_t =$ output of the textual modality

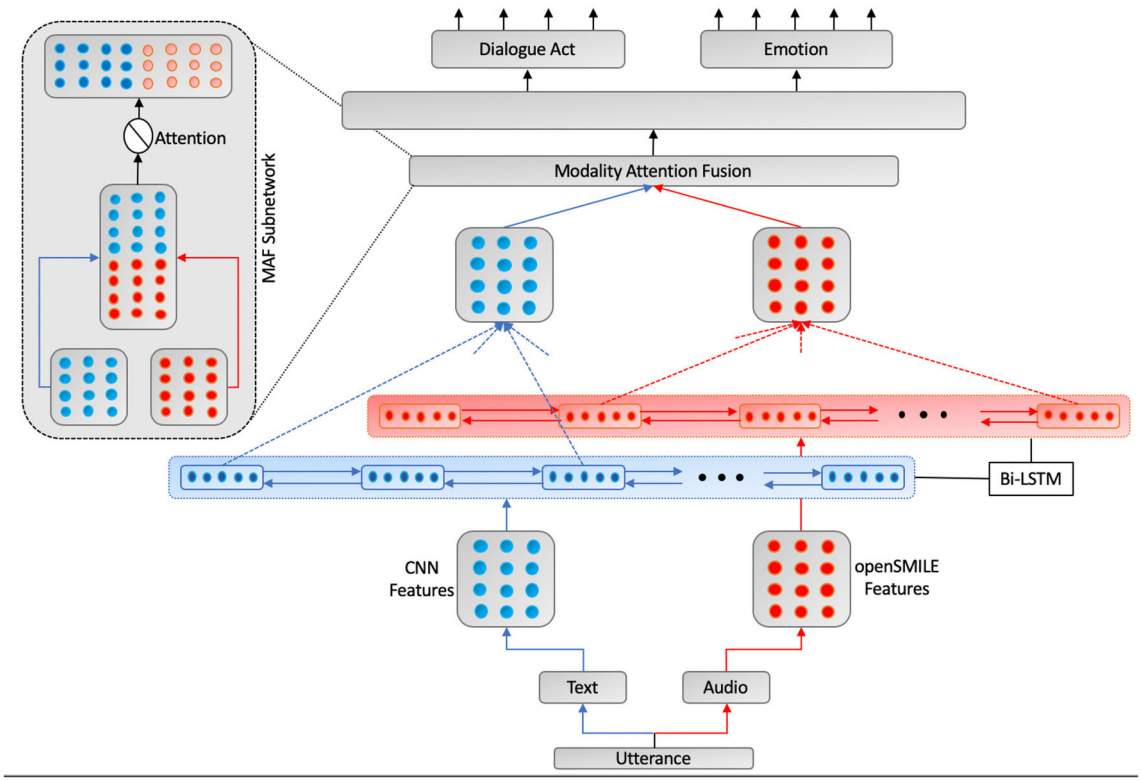

Fig. 3 Detailed architecture of the proposed network

encoder, each of size d dimensions and $M \in \mathbb{R}^{d \times 2}$. Here the value of d is 256.

The attention weight vector, β_{fuse}, and the final multi-modal vector, F, are obtained after attention is calculated as follows:

$$P_F = \tanh(W_F.M) \quad (2)$$

$$\beta_{\text{fuse}} = \text{softmax}(w_F^T.P_F) \quad (3)$$

$$F = M.\beta_{\text{fuse}} \quad (4)$$

Here, $W_F \in \mathbb{R}^{d \times d}$, $w_F \in \mathbb{R}^d$, $\beta_{fuse} \in \mathbb{R}^2$, and $F \in \mathbb{R}^d$. The output of the MAF layer is then fed to the classification layer. This MAF layer differs from [20] in the way that we let the multi-task network decide its parameters without any constraint for the generation of the attention scores.

Classification Layer The output from the MAF subnetwork is subsequently connected to the classification layer comprising of the output neurons for both the tasks. Thus, the classification layer contains two channels commonly sharing the MAF subnetwork layer. The two channels represent the output layers, one for each of the tasks (DAC and ER). The errors calculated from both these channels are jointly back-propagated to the subsequent previous layers for the proposed network to model and learn the features of both the tasks jointly thereby allowing both the tasks to benefit from the fully shared MAF subnetwork layer. Since, the focus of this study is to model DA with the help of emotion, the performance of the DAC task also banks on the quality of features learned by the ER task with better features aiding the joint learning process and vice versa. Algorithms 1, 2, and 3 show the pseudo-algorithm for each of the components of the proposed network.

Algorithm 1 Proposed multi-tasking algorithm.

begin
- Generate feature set from text as TS_t = `CNN(`$text$`)`;
- Generate feature set from audio as TS_a = `openSMILE(`$audio$`)`;
- M_a = `MEC(`TS_a`)` ▷ TS_a is time series audio feature set;
- M_t = `MEC(`TS_t`)` ▷ TS_t is time series text feature set;
- F = `MAF(`M_t, M_a`)`;
- $\hat{Y}_{DAC}, \hat{Y}_{ER}$ = `ClassificationLayer(`F`)` ▷ `ClassificationLayer()` has two output;
- $loss_{DAC}$ = `CrossEntropy(`$\hat{Y}_{DAC}, Y_{DAC}$`)` ▷ Y_{DAC}= true class label;
- $loss_{ER}$ = `CrossEntropy(`$\hat{Y}_{ER}, Y_{ER}$`)` ▷ Y_{ER}= true class label;
- $loss = loss_{DAC} + loss_{ER}$;
- Backpropagation to update the weights;

end

Algorithm 2 Modality encoders (MEC).

Procedure `MEC(`TS`)`
- M_i = `biLSTM(`TS`)`;
- **return :** M_i

Algorithm 3 Modality attention fusion (MAF) subnetwork.

Procedure `MAF(`M_t, M_a`)`
- M = `Concatenate(`M_t, M_a`)`;
- P_F = `tanh(`$W_F.M$`)` ▷ $W_F \in \mathbb{R}^{d \times d}$;
- β_{fuse} = `softmax(`$w_F^T.P_F$`)` ▷ $w_F \in \mathbb{R}^d$;
- F = M.β_{fuse};
- **return :** F

Implementation

This section presents the details of the training and testing data and the hyper-parameter details of each of the components of the proposed network.

The entire dataset was divided into two parts comprising of train and test set, hence a split of 80–20% was done. The statistics of the train and test set are shown in Table 2. The same train and test sets were used throughout all the experiments conducted, so as to have a fair comparison amongst all the approaches employed. For implementing the proposed and baseline models, Keras[2] has been used.

Experimental Details Section 1 clearly describes the details of the feature extraction process of the individual modalities. For encoding individual modalities, a Bi-LSTM layer with 256 memory cells was used followed by a dropout rate of 0.1. The classification layer comprises of two channels. The first channel contains 12 output neurons corresponding to the number of DA tags and the second channel contains 10 output neurons corresponding to the emotion categories. *Categorical crossentropy* is used as the loss function in both the channels. The following hyper-parameters were set to their default values and were then varied and tuned and optimal values were selected after the experiments on the proposed network and were used consistently to obtain results for all the baseline models:

- Learning rate : A learning rate of 0.01 was found to be optimum.
- Decay rate : A decay rate of 0.3 gave the best accuracy out of all.
- Dropout : A value of 0.1 was selected to be ideal for our setting.

[2] https://keras.io/

Table 2 Statistics of the train and test sets of the IEMOCAP dataset

	IEMOCAP	
	Utterance	Video
Train	7497	121
Test	1879	30

– Filter size : Filter sizes of 3 and 4 with 64 feature maps each in the CNN layer were used to obtain the textual features in our experimental setting.
– Other : 256 memory cells in the Bi-LSTM layer were found to give consistent results and *Adam* optimizer was used in the final setting.

Results and Analysis

A series of experiments were conducted to evaluate the performance of the proposed network for the multi-task framework along with different modalities in varying combinations of the network architecture, modalities and tasks. As stated earlier, the goal of this study is to analyze the role and affect of emotion while determining the DA of a speaker utterance from the multiple modalities. Thus, we do not put our focus on improving or analyzing the performance of the ER task and treat it as an auxiliary task aiding the primary task, i.e., DAC. In this regard, the results and findings are reported only for the task of DAC and its varying combinations.

Comparison with the Baselines For the baseline models, experiments were conducted in different categories. One set of categories involved experiments analyzing the contribution of different modalities. Next set analyzed the role of the MAF subnetwork for enhancing the contribution of different modalities. Another set examined the gain in fusion of modality features. All these categories were finally grouped to study the role of emotion in DAC by conducting experiments for the multi-task framework and its corresponding single task framework focused only on learning the automatic identification of the DAs. The baselines are listed as follows:

1. *Baseline:1* — Model is trained with only the textual features.
2. *Baseline:2* — Model is trained with the audio features solely.
3. *Baseline:3* — After extracting the audio and textual features, they are concatenated to pass through a single contextual LSTM (early fusion) without any attention layer.
4. *Baseline:4* — In addition to *Baseline:3*, MAF subnetwork is also incorporated.
5. *Baseline:5* — We follow our proposed approach without the attention layer to curate this baseline.

Table 3 shows the results of all the baselines and proposed models. As is evident from the table, as compared with the individual modalities, multi-modalities gives significantly better results in all the set-ups. Out of text and audio modalities, textual features are bound to give better results as for example, utterances belonging to "statement-opinion" and "non-opinion" might not have distinctive audio features but might have considerable semantic differences. However, in conjunction, audio features do add considerable differences to the aggregated

Table 3 Results of the baselines and the proposed models

	Task			
	DAC		DAC + ER	
Modality	Accuracy	F1 score	Accuracy	F1 score
Text	65.01%	0.6200	66.94%	0.6401
Audio	32.06%	0.2995	35.42%	0.3092
Text + audio (early fusion + without attention)	64.83%	0.6167	65.07%	0.6192
Text + audio (early fusion + with attention)	65.24%	0.6200	66.01%	0.6385
Text + audio (late fusion + without attention)	65.63%	0.6287	67.53%	0.6589
Text + audio (late fusion + with attention)	67.12%	0.6503	**69.63%** †	**0.6786** †

The higher the values of accuracy and F1 score, the better the performance of the corresponding model. †All the obtained results are statistically significant

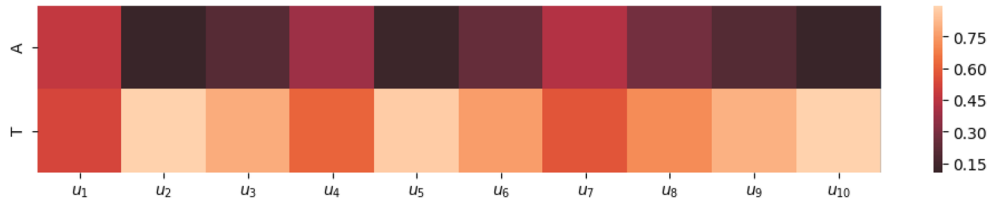

Fig. 4 The visualization of the attention scores for 10 sample utterances of the individual modalities. *A* and *T* represent attention scores of audio and textual features, respectively. Sample utterance—u_1: "Listen, shut your smug mouth, okay? I don't need that", u_2: "It was a hundred a sixty dollars. I'm sorry, uh, sixteen dollars", u_3: "You're quite insufferable", u_4: "Very well, if you insist on being boorish and idiotic.", u_5: "I thought you wanted to see them", u_6: "I mean look at the view of the moon we got from here", u_7: "It certainly is not. It's a slightly exaggerated scientific fact", u_8: "You felt something that far back?", u_9: "She's okay with it", u_{10}: "Hi. Thanks for waiting".

features in cases as, e.g., "agreement" and "disagreement" though spoken with similar linguistic content does add distinctive features in its audio counterpart because of the non-behavioral cues for agreeing or disagreeing to something. All these combinations, aid significantly with the incorporation of emotion into the framework as is evident from the results. All the multi-task experimental setups with emotion produce better results compared with their sole DAC counterparts. The proposed multi-task model with multi-modalities achieves the best accuracy of *69.63%* with F1 score of *0.6786* whereas the single-task DAC with multi-modalities achieves an accuracy of *67.12%* with F1 score of *0.6503*. The multi-task framework with multi-modalities attains an improvement of **2.5%** against its DAC counterpart with the proposed architecture (with attention). The baseline *without attention* refers to the model without the MAF subnetwork layer where the features from individual modalities are simply concatenated and passed through the classification layer. This helps in understanding that not all modalities contribute equally towards learning shared features to learn a model as is also evident from the individual modality counterpart. Figure 4 shows the heatmap visualization of the attention scores of the MAF subnetwork for individual modalities for sample utterances. It is evident from the visualization that attention scores were predominant for textual features. But there were couple of instances such as u_1 and u_4, where the audio modality indeed contributed distinctive features. The reason for the low F1 score in all the experiments can be attributed to the skewness in the dataset for the emotion as well as DA categories (as shown in Fig. 1) since not all the DAs have sufficient representations in the dataset. This is in sync with real time conversations where some DAs do not occur frequently than others. We also perform experiments in terms of when to fuse different modalities, i.e., early fusion or late fusion. As per our experiments, it is observed that late fusion performed better in our case.

As seen from the accuracy and F1 scores, it is noticeable that emotion indeed aided the joint learning process for the DAC to benefit from it. A thorough case study was conducted to investigate the role of emotion in classifying the DAs. For e.g., "*That's very amusing indeed*" was misinterpreted as "agreement" in the DAC model, but was correctly classified as "disagreement" in DAC+ER model as the emotion of the utterance was "angry" given the context that the speaker was disagreeing with the hearer in a sarcastic manner. It was seen that for longer utterances comprising of composite sentences, emotion did play significant role in correctly identifying the DA such as "Hey, I'm, uh. I'm really sorry about what happened. I don't um- I mean what you can you do?" was mis-classified as "question" in the DAC model but was correctly identified as "apologize" in the DAC+ER model given the "sad" emotion of the speaker that it is simply trying to sympathize with the sufferer. Emotion was seen to basically aid the performance of the expressive DAs such as "command", "apologize", "agreement", and "disagreement". For example, "All right. All right. Calm yourself" was wrongly identified as "agreement" but with DAC+ER model it was correctly classified as "command" with the help of the speaker's emotion which in this case was "frustrated". We also analyze the affect of emotion for the classification of DAs with the help of heatmap visualization of the weights

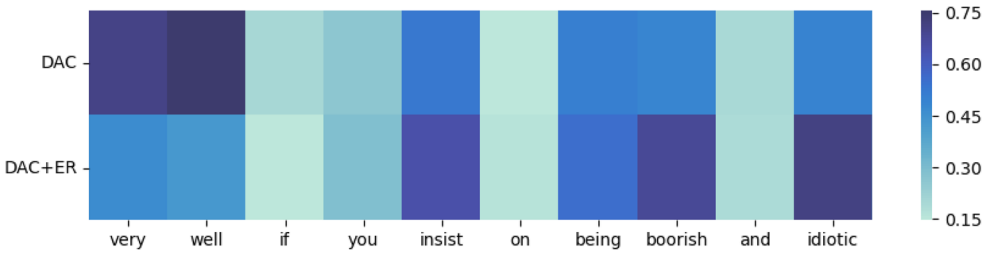

Fig. 5 The visualization of the weights for an utterance—u_1: "Very well, if you insist on being boorish and idiotic." for the multi-modal, single task DAC (baseline) and multi-task DAC+ER (proposed) model

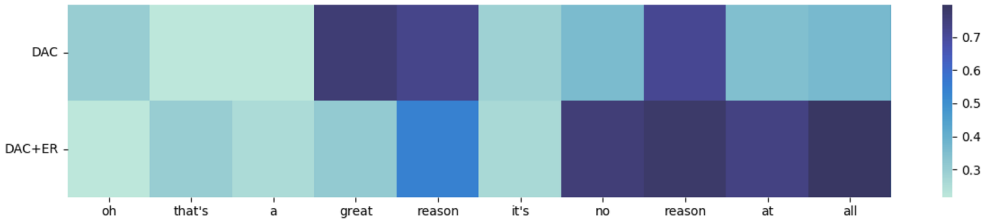

Fig. 6 The visualization of the weights for an utterance—u_2: "Oh that's a great reason. It's no reason at all" for the multi-modal, single-task DAC (baseline) and multi-task DAC+ER (proposed) model

learnt for the multi-task model against its single task DAC counterpart. Figure 5 shows the visualization of the weights for a sample utterance "Very well, if you insist on being boorish and idiotic" for the single-task DAC as well as the multi-task model. The DAC model misclassified the utterance as "agreement", whereas the proposed multi-task model correctly identified the utterance as "disagreement" by correctly recognizing the emotion of the speaker as "anger". As is seen from the figure, for the DAC model, much more emphasis was laid on words such as *very*, *well* which led to it being misclassified. Whereas for the proposed multi-task model, the joint representation learnt by the DAC and ER tasks laid much more emphasis on words such as *boorish*, *idiotic* by leveraging significantly from the multi-modalities. Similarly, Fig. 6 shows the heatmap visualization of one more sample utterance "Oh that's a great reason. It's no reason at all" for the analysis of affect of emotion. The DAC model misclassified the utterance as "acknowledge" whereas, the multi-task model correctly classified it as "disagreement" by correctly identifying the emotion as "frustrated". Thus, it is evident from Fig. 7 that the presence of emotion indeed aided the performance of the system in terms of correct identification of the DAs.

Statistical Significance Test Results produced by all our best performing models are statistically significant as we have performed Welch's t test [58] at 5% significance level. A statistical hypothesis test named Welch's t test (paired t test) is conducted at the 5% (0.05) significance level, i.e., 95% confidence to verify whether the performance improvement attained by our model is statistically significant or not. This is done to show that the best results obtained by our proposed method is statistically significant and has not occurred by chance. For statistical tests on the annotated corpus, the system was executed for a total of 20 times.

Error Analysis Investigation also revealed certain scenarios and reasons where the system falters and plausible reasons behind the same which are as follows: (i) As mentioned, one of the primary reasons for low F1 score is that the representation of most the tags in the dataset is very less; i.e., the dataset is skewed as shown in Fig. 1 with the maximum representation of "statement-non-opinion", "statement-opinion", "question", and "answer" tags. This is typically in sync with real-time task-independent conversations. (ii) Also, utterances of this dataset are of longer lengths and composite in nature encompassing multiple intentions in a single spoken utterance. Thus, it becomes difficult in those cases to learn features for discretizing DAs. For example, "Oh very, very interesting Amanda. How about the child of uh... four, six or maybe

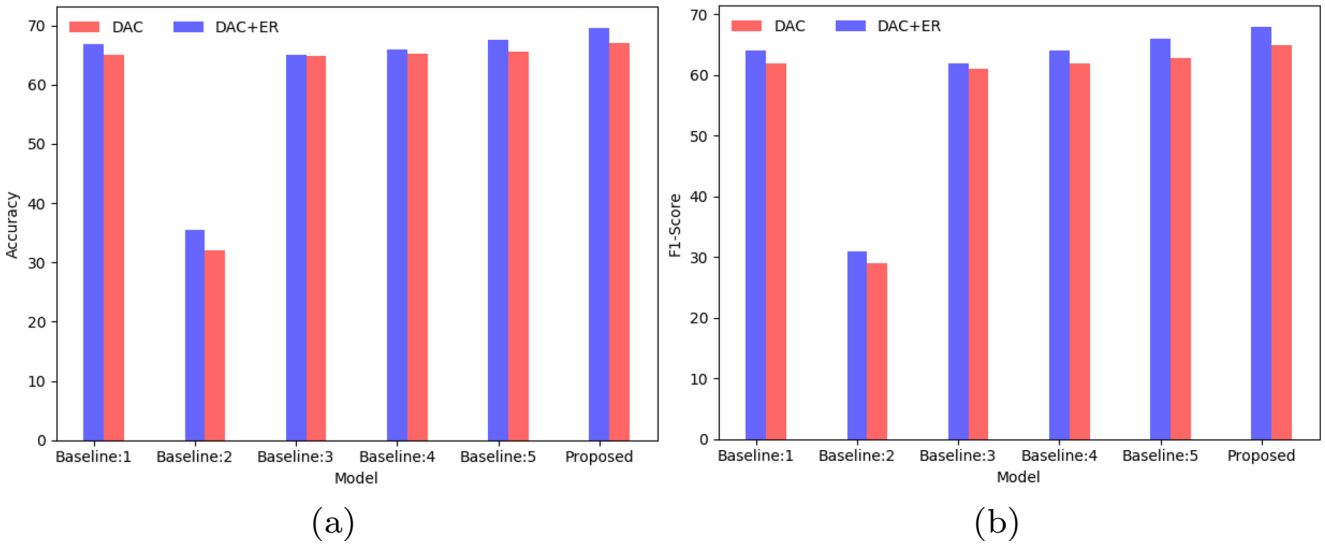

Fig. 7 Pair-wise comparison between the single-task (DAC) and multi-task (DAC+ER). **a** In terms of accuracy. **b** In terms of F1 score

nine—you know, we could work up a splendid little debate about it, you know? Intemperate tots" indicates "opinion", "question" and also "agreement" as its intention from the semantics making it extremely difficult and confusing to prioritize a distinct DA. (iii) One of the significant reasons for the misclassification of the DAs can be attributed to the misclassification of the emotions for that particular utterance. For example, "Look at this, my hairs are standing up my arm. I'm giving myself goose bumps" was misclassified as "fear" for the emotion tag as opposed to the correct "excited" tag. Similarly, "Really, Amanda" was wrongly classified as "excited" for the emotion tag whereas it should be tagged as "frustrated". In all these cases, their corresponding DA tag was wrongly identified.

Conclusions and Future Work

This paper presents an investigative study to analyze the role and affect of emotion in automatic identification of DAs in task-independent conversations as emotional state of a speaker has a considerable affect on its intentional or pragmatic content. A DL-based multi-task framework has been developed to jointly learn DAC and ER task in a multi-modal framework (specifically text and audio). The proposed network incorporates attention to facilitate the fusion of various modalities. IEMOCAP, an open-source benchmark ER multi-modal dataset, has been manually annotated for its corresponding DA to make it suitable for learning both the tasks jointly and boost the research in the field of multi-modal DAC. Several investigations and comparison with baselines varying in modalities, network architecture, and tasks were carried out. The proposed multi-task framework achieved a significant improvement of **2.5%** against its single-task DAC framework. It was seen that emotion indeed aided the task of DAC in several scenarios such as for expressive DAs, similar linguistic content, and composite DAs as reported.

Future work includes benchmarking this investigation and analysis in several other ER multi-modal datasets to increase its scope and advance the research in emotion-aided DAC, multi-modal DAC, etc. Currently, the proposed network exploits acoustic and textual modalities. Incorporating visual features such as motion capture of head, hands, and face to analyze the contribution of different modalities will be addressed in the future work.

Acknowledgments Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Jurafsky D, Bates R, Coccaro N, Martin R, Meteer M, Ries K, Shriberg E, Stolcke A, Taylor P, Van Ess-Dykema C. Automatic detection of discourse structure for speech recognition and understanding. In: 1997 IEEE workshop on automatic speech recognition and understanding proceedings, IEEE, pp 88–95. 1997.
2. Stolcke A, Ries K, Coccaro N, Shriberg E, Bates R, Jurafsky D, Taylor P, Martin R, Ess-Dykema CV, Meteer M. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational linguistics. 2000;26(3):339–373.
3. Verbree D, Rienks R, Heylen D. Dialogue-act tagging using smart feature selection; results on multiple corpora. In: Spoken Language Technology Workshop, 2006. IEEE, IEEE, pp 70–73. 2006.
4. Kalchbrenner N, Blunsom P. Recurrent convolutional neural networks for discourse compositionality. 2013. arXiv:13063584.
5. Papalampidi P, Iosif E, Potamianos A. Dialogue act semantic representation and classification using recurrent neural networks. SEMDIAL 2017 SaarDial, pp 104. 2017.
6. Liu Y, Han K, Tan Z, Lei Y. Using context information for dialog act classification in dnn framework. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 2170–2178. 2017.
7. Ribeiro E, Ribeiro R, de Matos DM. A multilingual and multidomain study on dialog act recognition using character-level tokenization. Information. 2019;10(3):94.
8. DeLamater JD, Ward A. Handbook of social psychology. Berlin: Springer; 2006.
9. Fleckenstein KS. Defining affect in relation to cognition: A response to susan mcleod . J Adv Comp. 1991;11:447–453.
10. Barrett LF, Lewis M, Haviland-Jones JM. Handbook of emotions. New York: The Guilford Press; 1993.
11. Zadeh AB, Liang PP, Poria S, Cambria E, Morency LP. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th annual meeting of the association for computational linguistics (vol 1: Long Papers), pp 2236–2246. 2018.
12. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG. Emotion recognition in human-computer interaction. IEEE Signal Proc Mag. 2001;18(1):32–80.
13. Jain N, Kumar S, Kumar A, Shamsolmoali P, Zareapoor M. Hybrid deep neural networks for face emotion recognition. Pattern Recogn Lett. 2018;115:101–106.
14. Zhang S, Zhang S, Huang T, Gao W, Tian Q. Learning affective features with a hybrid deep model for audio–visual emotion recognition. IEEE Trans Circuits Syst Video Technol. 2018;28(10):3030–3043.
15. Huang C, Zaiane O, Trabelsi A, Dziri N. Automatic dialogue generation with expressed emotions. In: Proceedings of the 2018 conference of the north american chapter of the association for

computational linguistics: Human language technologies, vol 2 (Short Papers), pp 49–54. 2018.
16. Zhou H, Huang M, Zhang T, Zhu X, Liu B. Emotional chatting machine: Emotional conversation generation with internal and external memory. In: 32nd AAAI conference on artificial intelligence. 2018.
17. Fung P, Bertero D, Xu P, Park JH, Wu CS, Madotto A. Empathetic dialog systems. In: The international conference on language resources and evaluation. European Language Resources Association. 2018.
18. Novielli N, Strapparava C. The role of affect analysis in dialogue act identification. IEEE Trans Affect Comput. 2013;4(4):439–451.
19. Bosma W, André E. Exploiting emotions to disambiguate dialogue acts. In: Proceedings of the 9th international conference on Intelligent user interfaces, ACM, pp 85–92. 2004.
20. Poria S, Cambria E, Hazarika D, Mazumder N, Zadeh A, Morency LP. Multi-level multiple attentions for contextual multimodal sentiment analysis. In: 2017 IEEE international conference on data mining (ICDM), IEEE, pp 1033–1038. 2017.
21. Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency LP. Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th annual meeting of the association for computational linguistics (vol 1: Long Papers), pp 873–883. 2017.
22. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS. Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation. 2008;42(4):335.
23. Reithinger N, Klesen M. Dialogue act classification using language models. In: 5th European conference on speech communication and technology. 1997.
24. Stolcke A, Shriberg E, Bates R, Coccaro N, Jurafsky D, Martin R, Meteer M, Ries K, Taylor P, Van Ess-Dykema C, et al. Dialog act modeling for conversational speech. In: AAAI spring symposium on applying machine learning to discourse processing, pp 98–105. 1998.
25. Grau S, Sanchis E, Castro MJ, Vilar D. Dialogue act classification using a bayesian approach. In: 9th Conference Speech and Computer. 2004.
26. Godfrey JJ, Holliman EC, McDaniel J. Switchboard: Telephone speech corpus for research and development. In: 1992 IEEE international conference on acoustics, speech, and signal processing, 1992. ICASSP-92, IEEE, vol 1, pp 517–520. 1992.
27. Khanpour H, Guntakandla N, Nielsen R. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In: Proceedings of COLING 2016, The 26th international conference on computational linguistics: Technical Papers, pp 2012–2021. 2016.
28. Lee JY, Dernoncourt F. Sequential short-text classification with recurrent and convolutional neural networks. In: Proceedings of the 2016 Conference of the North American chapter of the association for computational linguistics: Human language technologies, association for computational linguistics, pp 515–520. 2016. http://aclweb.org/anthology/N16-1062.
29. Kumar H, Agarwal A, Dasgupta R, Joshi S. Dialogue act sequence labeling using hierarchical encoder with crf. In: 32nd AAAI conference on artificial intelligence. 2018.
30. Raheja V, Tetreault J. Dialogue act classification with context-aware self-attention. 2019. arXiv:190402594.
31. Yu Y, Peng S, Yang GH. Modeling long-range context for concurrent dialogue acts recognition. 2019. arXiv:190900521.
32. Sitter S, Stein A. Modeling the illocutionary aspects of information-seeking dialogues. Inf Process Manag. 1992;28(2):165–180.
33. Ortega D, Li CY, Vallejo G, Denisov P, Vu NT. Context-aware neural-based dialog act classification on automatically generated transcriptions. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 7265–7269. 2019.
34. Saha T, Srivastava S, Firdaus M, Saha S, Ekbal A, Bhattacharyya P. Exploring machine learning and deep learning frameworks for task-oriented dialogue act classification. In: International joint conference on neural networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019, pp 1–8. 2019. https://doi.org/10.1109/IJCNN.2019.8851943.
35. Boyer KE, Grafsgaard JF, Ha EY, Phillips R, Lester JC. An affect-enriched dialogue act classification model for task-oriented dialogue. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies vol 1, Association for Computational Linguistics, pp 1190–1199. 2011.
36. Ihasz PL, Kryssanov V. Emotions and intentions mediated with dialogue acts. In: 2018 5th international conference on business and industrial research (ICBIR), IEEE, pp 125–130. 2018.
37. Cerisara C, Jafaritazehjani S, Oluokun A, Le H. Multi-task dialog act and sentiment recognition on mastodon. 2018. arXiv:180705013.
38. Vosoughi S, Roy D. Tweet acts: A speech act classifier for twitter. In: 10th international AAAI conference on web and social media. 2016.
39. Lauren P, Qu G, Yang J, Watta P, Huang GB, Lendasse A. Generating word embeddings from an extreme learning machine for sentiment analysis and sequence labeling tasks. Cogn Comput. 2018;10(4):625–638.
40. Wang Z, Lin Z. Optimal feature selection for learning-based algorithms for sentiment classification. Cognitive Computation pp 1–11. 2019.
41. Sun X, Peng X, Ding S. Emotional human-machine conversation generation based on long short-term memory. Cogn Comput. 2018;10(3):389–397. https://doi.org/10.1007/s12559-017-9539-4.
42. Griol D, Callejas Z. Mobile conversational agents for context-aware care applications. Cogn Comput. 2016;8(2):336–356. https://doi.org/10.1007/s12559-015-9352-x.
43. Rodríguez LF, Ramos F. Development of computational models of emotions for autonomous agents: A review. Cogn Comput. 2014;6(3):351–375. https://doi.org/10.1007/s12559-013-9244-x.
44. Shriberg E, Dhillon R, Bhagat S, Ang J, Carvey H. The icsi meeting recorder dialog act (mrda) corpus. In: Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL; 2004.
45. Heeman PA, Allen JF. The trains 93 dialogues. Tech. rep., Rochester Univ NY Dept of Computer Science. 1995.
46. Anderson AH, Bader M, Bard EG, Boyle E, Doherty G, Garrod S, Isard S, Kowtko J, McAllister J, Miller J, et al. The hcrc map task corpus. Language and speech. 1991;34(4):351–366.
47. Jurafsky D. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. Institute of Cognitive Science Technical Report. 1997.
48. LeCun Y, Bengio Y, et al. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks. 1995;3361(10):1995.
49. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543. 2014.
50. Eyben F, Wöllmer M, Schuller B. Opensmile: The munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia, ACM, pp 1459–1462. 2010.

51. Drugman T, Thomas M, Gudnason J, Naylor P, Dutoit T. Detection of glottal closure instants from speech signals: A quantitative review. IEEE Trans Audio, Speech, Language Process. 2011; 20(3):994–1006.
52. Kane J, Gobl C. Wavelet maxima dispersion for breathy to tense voice discrimination. IEEE Trans Audio, Speech, Language Process. 2013;21(6):1170–1179.
53. Drugman T, Alwan A. Joint robust voicing detection and pitch estimation based on residual harmonics. In: 12th annual conference of the international speech communication association. 2011.
54. Hermansky H. Perceptual linear predictive (plp) analysis of speech. The Journal of the Acoustical Society of America. 1990;87(4):1738–1752.
55. Fastl H. Psycho-acoustics and sound quality. In: Communication acoustics, Springer, pp 139–162. 2005.
56. Thomson DJ. Spectrum estimation and harmonic analysis. Proc IEEE. 1982;70(9):1055–1096.
57. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–1780.
58. Welch BL. The generalization ofstudent's' problem when several different population variances are involved. Biometrika. 1947;34(1/2):28–35.