COSMIC: COmmonSense knowledge for eMotion Identification in Conversations

Deepanway Ghosal[†], Navonil Majumder[†], Alexander Gelbukh[◇],
Rada Mihalcea[△], Soujanya Poria[†]

[†] Singapore University of Technology and Design, Singapore
[◇] CIC, Instituto Politécnico Nacional, Mexico
[△] University of Michigan, USA
{deepanway_ghosal@mymail.,navonil_majumder@, sporia@}sutd.edu.sg,
gelbukh@cic.ipn.mx, mihalcea@umich.edu

Abstract

In this paper, we address the task of utterance level emotion recognition in conversations using commonsense knowledge. We propose COSMIC, a new framework that incorporates different elements of commonsense such as mental states, events, and causal relations, and build upon them to learn interactions between interlocutors participating in a conversation. Current state-of-the-art methods often encounter difficulties in context propagation, emotion shift detection, and differentiating between related emotion classes. By learning distinct commonsense representations, COSMIC addresses these challenges and achieves new state-of-the-art results for emotion recognition on four different benchmark conversational datasets. Our code is available at https://github.com/declare-lab/conv-emotion.

1 Introduction

Emotion recognition is a long-standing research problem in Artificial Intelligence (AI). With the growing popularity of conversational AI research, the topic of emotion recognition in conversations has received significant attention from the research community (Li et al., 2020; Ghosal et al., 2019; Zhang et al., 2019). Identifying emotions in conversations is a core step toward fine-grained conversation understanding, which in turn is essential for downstream tasks such as emotion-aware chat agents (Lin et al., 2019; Rashkin et al., 2019), visual question answering (Tapaswi et al., 2016; Azab, 2019), health conversations (Althoff et al., 2016; Pérez-Rosas et al., 2017) and others.

Natural conversations are complex as they are governed by several distinct variables that affect the flow of a conversation and the emotional dynamics of the participants. These variables include
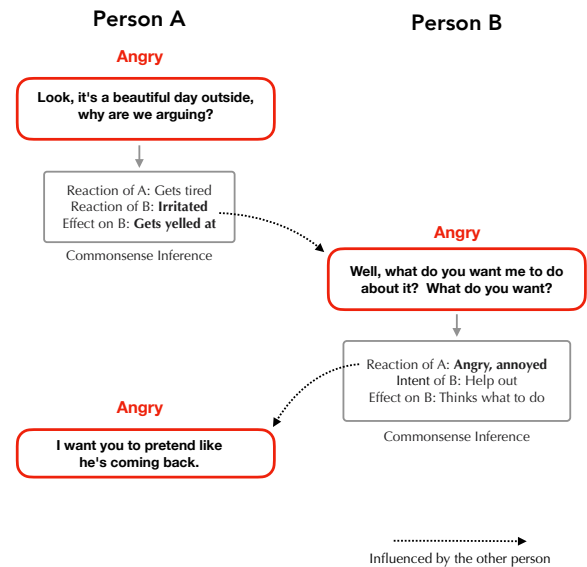

Figure 1: Commonsense knowledge can lead to explainable dialogue understanding. It will help models to understand, reason, and explain events and situations. In this particular example, commonsense inference is applied to a sequence of utterances in a two-party conversation. Person A's first utterance indicates that he/she is tired of arguing with person B. The tone of the utterance also implies that person B is getting yelled at by person A, which invokes a reaction of irritation in person B. Person B then asks what he/she can do to help and says this while being angry. This again makes person A annoyed and influences him/her to respond with anger. This kind of inferred commonsense knowledge about the reaction, effect, and intent of the speaker and the listener helps in predicting the emotional dynamics of the participants.

topic, viewpoint, speaker personality, argumentation logic, intent, and so on (Poria et al., 2019b). Additionally, individual utterances are also governed by the mental state, intent, and emotional state of the participants at the time when they are uttered. In this conversation model, only the utterances can be observed as the conversation unfolds,

while other variables such as speaker state and intent remain latent as they are not directly observed by the other participants. Similarly, the emotional state of the speakers cannot be directly observed, but it can be inferred from the utterances that are observable.[1]

The commonsense knowledge of the participants in a conversation plays a central role in inferring the latent variables of a conversation. It is used to guide the participants through their reasoning about the content of the conversation, dialog planning, decision making, and many other reasoning tasks. It is also used to recognize other finer-grained elements of a conversation, such as avoiding repetition, asking questions, refraining from giving unrelated responses, and so on — all of which control aspects of the conversation such as fluency, interestingness, inquisitiveness, or empathy. Commonsense knowledge is thus necessary to model the nature and flow of the dialogue and the emotional dynamics of the participants. In Figure 1, we illustrate one such scenario where commonsense knowledge is utilized to infer emotions of the utterances in a dialogue.

Natural language is often indicative of one's emotion. Hence, emotion recognition has been enjoying popularity in the field of NLP (Kratzwald et al., 2018; Colneriĉ and Demsar, 2018), due to its widespread applications in opinion mining, recommender systems, healthcare, and so on. Only in the past few years has emotion recognition in conversation (ERC) gained attention from the NLP community (Yeh et al., 2019; Chen et al., 2018; Majumder et al., 2019; Zhou et al., 2018) due to the growing availability of public conversational data. ERC can be used to analyze conversations that take place on social media. It can also aid in analyzing conversations in real time, which can be instrumental in legal trials, interviews, e-health services, and more. Unlike vanilla emotion recognition of sentences/utterances, ERC ideally requires context modeling of the individual utterances. This context can be attributed to the preceding utterances, and relies on the temporal sequence of utterances. Compared to the recently published works on ERC (Chen et al., 2018; Majumder et al., 2019; Zhou et al., 2018; Qin et al., 2020; Zhong et al., 2019; Zhang et al., 2019), both lexicon-based (Wu et al., 2006; Mohammad and Turney, 2010; Shaheen et al., 2014) and modern

deep learning-based (Kratzwald et al., 2018; Colneriĉ and Demsar, 2018) vanilla emotion recognition approaches fail to work well on ERC datasets as this work ignores the conversation specific factors such as the presence of contextual cues, the temporality in speakers' turns, or speaker-specific information.

In this paper, we introduce COSMIC, a commonsense-guided framework for emotion identification in conversations. By building upon a very large commonsense knowledge base, our proposed framework captures some of the complex interactions between personality, events, mental states, intents, and emotions leading towards a better understanding of the emotional dynamics and other aspects of conversation. Through extensive evaluations on four different conversation datasets and comparisons with several baselines and state-of-the-art models, we show the effectiveness of a model that explicitly accounts for commonsense. Moreover, feature ablation experiments highlight the role that such knowledge plays in identifying emotion in conversations.

2 Related Work

Emotion recognition has been an active area of research for many years and has been explored across inter-disciplinary fields such as machine learning, signal processing, social and cognitive psychology, etc (Picard, 2010). The seminal work from Ekman (1993) presented findings on facial expressions, methods to measure facial expression and their relation with human emotion. Acoustic information and visual cues were later used for emotion recognition by Datcu and Rothkrantz (2014).

However, emotion recognition in conversations has gained popularity only recently due to the emergence of publicly available conversational datasets collected from social media platforms and scripted situations such as movies and tv-shows (Poria et al., 2019a; Zahiri and Choi, 2018). The main approach towards conversational emotion recognition is to perform contextual modeling in either textual or multimodal setting with deep-learning based algorithms. Poria et al. (2017) used recurrent neural networks for multimodal emotion recognition followed by (Majumder et al., 2019), where party and global states were used for modeling the emotional dynamics. An external knowledge base was used in (Zhong et al., 2019) with transformer networks to perform emotion recognition. Some of the other

---

[1]In multimodal conversations, there are other variables that can be observed, such as facial expressions, gestures, pitch, and acoustic indicators.

important works include (Hazarika et al., 2018a,b; Zadeh et al., 2018b; Chen et al., 2017; Zadeh et al., 2018a).

3 Methodology

3.1 Task definition

Given the transcript of a conversation along with speaker information for each constituent utterance, the ERC task aims to identify the emotion of each utterance from a set of pre-defined emotions. Figure 1 illustrates one such conversation between two people, where each utterance is labeled by the underlying emotion. Formally, given an input sequence of N utterances $[(u_1, p_1), (u_2, p_2), \ldots, (u_N, p_N)]$, where each utterance $u_i = [u_{i,1}, u_{i,2}, \ldots, u_{i,T}]$ consists of T words $u_{i,j}$ spoken by party p_i, the task is to predict the emotion label e_i of each utterance u_i. In conversational emotion recognition, the task is to classify each of the constituting utterances into its appropriate emotion category. In literature, the main approach towards this problem has been to first produce context independent representations and then perform contextual modeling. We identify these two distinct modeling phases and aim to improve both of them through the proposed COSMIC framework. Our framework consists of three main stages:

1. Context independent feature extraction from pretrained transformer language models.

2. Commonsense feature extraction from a commonsense knowledge graph.

3. Incorporating commonsense knowledge to design better contextual representations and using it for the final emotion classification.

The overall architecture of the COSMIC framework is illustrated in Figure 2.

3.2 Context Independent Feature Extraction

We employ the RoBERTa model (Liu et al., 2019) to extract context independent utterance level feature vectors. We first fine-tune the RoBERTa Large model for emotion label prediction from the transcript of the utterances. RoBERTa Large follows the original BERT Large (Devlin et al., 2018) architecture having 24 layers, 16 self-attention heads in each block and a hidden dimension of 1024, resulting in a total of 355M parameters. Let an utterance

x consists of a sequence of BPE tokenized tokens $x_1, x_2, \ldots, x_N$, with emotion label E_x. In this setting, the fine-tuning of the pretrained RoBERTa model is realized through a sentence classification task. A special token $[CLS]$ is appended at the beginning of the utterance to create the input sequence for the model: $[CLS], x_1, x_2, \ldots, x_N$. This sequence is passed through the model, and the activation from the last layer corresponding to the $[CLS]$ token is then used in a small feedforward network to classify it into its emotion class E_x.

Once the model has been fine-tuned for emotion label classification, we pass the $[CLS]$ appended BPE tokenized utterances to it and extract out activations from the final four layers corresponding to the $[CLS]$ token. These four vectors are then averaged to obtain the context independent utterance feature vector with a dimension of 1024.

3.3 Commonsense Feature Extraction

Commonsense Feature	Notation	Nature	Causal Relation
Intent of speaker	$\mathcal{IS}_{cs}(.)$	Mental state	Cause
Effect on speaker	$\mathcal{ES}_{cs}(.)$	Mental state	Effect
Reaction of speaker	$\mathcal{RS}_{cs}(.)$	Event	Effect
Effect of listeners	$\mathcal{EL}_{cs}(.)$	Mental state	Effect
Reaction of listeners	$\mathcal{RL}_{cs}(.)$	Event	Effect

Table 1: Functional notations of commonsense knowledge used in COMET. The functions take as input the utterance u and returns the feature indicated in the leftmost column. Intent and effect on speaker and listeners can be categorized into *mental states*, whereas their reactions are *events*. Intent is also a *causal* variable whereas the rest are *effects*.

In this work, we use the commonsense transformer model COMET (Bosselut et al., 2019) to extract the commonsense features. COMET is trained on several commonsense knowledge graphs to perform automatic knowledge base construction. The model is given a triplet $\{s, r, o\}$ from the graph and is trained to generate the object phrase o from concatenated subject phrase s and relation phrase r. COMET is an encoder-decoder model that uses the pretrained autoregressive language model GPT (Radford et al., 2018) as the base generative model.

To perform the task of generative commonsense knowledge construction, COMET is trained on ATOMIC (The Atlas of Machine Commonsense) (Sap et al., 2019), a collection of everyday inferential *if-then* commonsense knowledge organized through textual descriptions. ATOMIC consists of nine different if-then relation types to distinguish

agents vs themes, causes vs effects, voluntary vs non-voluntary events, and actions vs mental states. Given an event in which X participates, the nine relation types (r) are inferred as follows: i) *intent of X*, ii) *need of X*, iii) *attribute of X*, iv) *effect on X*, v) *wanted by X*, vi) *reaction of X*, vii) *effect on others*, viii) *wanted by others*, and ix) *reaction of others*. As an example, given an event or subject phrase (s): "Person X gives Person Y a compliment", the inference from COMET for relation phrase (r): *intent of X* and *reaction of others* would be "X wanted to be nice" and "Y will feel flattered" respectively.

COMET is a generative model and as illustrated in the above example it produces a discrete sequence of commonsense knowledge conditioned on the subject and relation phrase. In our model however, we make use of continuous vectors of commonsense representations. For that, we take the pretrained COMET model on ATOMIC knowledge graph and discard the phrase generating decoder module. We treat utterance U as the subject phrase and concatenate it with the relation phrase r. Next, we pass the concatenated $\{U \oplus r\}$ through the encoder of COMET and extract out the activations from the final time-step. In particular we use the relations presented in Table 1: *intent of X*, *effect on X*, *reaction of X*, *effect on others* and *reaction of others* (where X is the speaker and $others$ are listeners). Performing this feature extraction operation results in five different vectors (respective to the five different relations) for each utterance in the conversation. These vectors are 768 dimensional.

The nature of the various relation types in ATOMIC allows us to extend it naturally to conversational frameworks. The relations enable the modeling of phenomenons such as content (event, persona, mental states) and causal relations (cause, effect, stative) which are essential elements for understanding conversational context. These different relations are of key importance because generally there is a major interplay between virtually all of them throughout the course of a conversation. For instance, the relations i) - vi) are all intrinsically related to the speaker and vii) - ix) are all akin to the listener. On a more fine-grained level, the *intent, effect* and *react* components of the speaker and listener are all elemental for understanding the nature of the conversation. We surmise that adopting these relational variables in a unified framework would be highly useful to create enhanced representations of the conversation.

State	Influenced By
Context State	Utterance, Internal state, External state
Internal State	Context state, *Effect on speaker, listener*
External State	Context state, Utterance, *Reaction of speaker, listener*
Intent State	Internal state, *Intent of speaker*
Emotion State	Utterance, Intent state Internal state, External state

Table 2: Different states and the respective variables they are influenced by. *Italic* variables are forms of commonsense knowledge from Table 1.

3.4 Commonsense Conversational Model

We first introduce our notations and present a high level view of the main architecture of our COSMIC model. A conversation consists of N utterances $u_1, u_2, \ldots, u_N$, in which M distinct speakers/participants $p_1, p_2, \ldots, p_M$ take part. Utterance u_t is spoken by participant $p_{s(u_t)}$. For every $t \in \{1, 2, \ldots, N\}$, we denote context independent RoBERTa vectors by x_t. Commonsense vectors corresponding to *intent of X, effect on X, reaction of X, effect on others* and *reaction of others* are denoted by $\mathcal{IS}_{cs}(u_t), \mathcal{ES}_{cs}(u_t), \mathcal{RS}_{cs}(u_t), \mathcal{EL}_{cs}(u_t)$, and $\mathcal{RL}_{cs}(u_t)$ respectively. X is assumed to be the speaker and $others$ are assumed to be the listeners.

Since conversations are highly sequential in nature and contextual information flows along a sequence, a *context state* c_t and *attention vector* a_t are formulated that model the sequential dependency between utterances. The *context state* and *attention vector* are always shared between all the participants of the conversation.

An *internal state, external state* and *intent state* are used to model different mental states, actions and events for the participants. These are represented by $q_{k,t}, r_{k,t}$ and $i_{k,t}$ for the participants $k \in [1, 2, \ldots, M]$. The *internal state* and the *external state* can be collectively considered as the speaker state. This states are necessary to capture the complex mental and emotional dynamics of the participants. The emotion state e_t is then modelled from a combination of the three states and the immediate preceding emotion state. Finally the appropriate emotion class for the utterance is inferred from the emotion state.

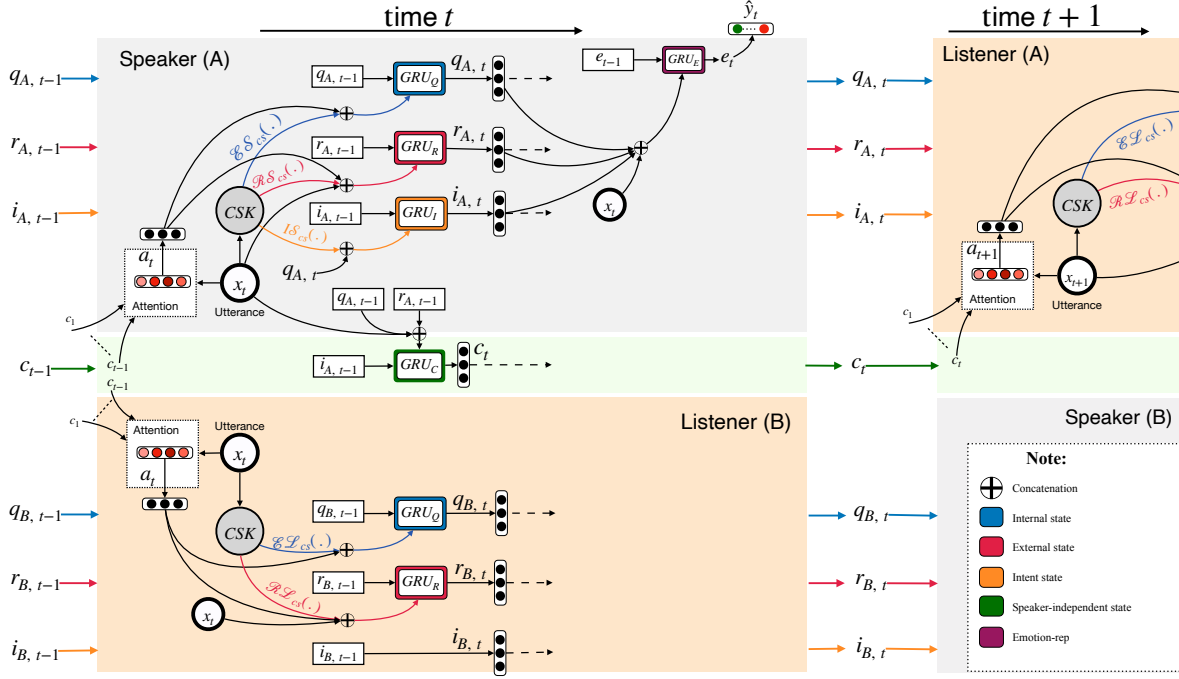In our framework, context and commonsense modeling is performed using GRU cells (Chung

Figure 2: Illustration of COSMIC framework. *CSK*: Commonsense knowledge from COMET. In practice we use Bidirectional GRU cells. However, for clarity unidirectional cells are shown in the sketch.

et al., 2014). GRU cells take as input y_t and update its hidden state from h_{t-1} to h_t using the transformation: $h_t = GRU(h_{t-1}, y_t)$. New hidden state h_t also serves as the output of the current step. The cell is parameterized by weights W and biases b of appropriate sizes depending upon the input y_t and output h_t. We use five Bidirectional GRU cells $GRU_C, GRU_Q, GRU_R, GRU_I$, and GRU_E for modeling context state, internal state, external state, intent state, and emotion state respectively. For ease of representation we formulate the different states with unidirectional GRU cells here.

Context State: The context state stores and propagates the overall utterance-level information along the sequence of the conversation flow. This state is updated using context GRU cell GRU_C after each time-step t when the utterance is uttered by some participant $p_{s(u_t)}$. RoBERTa feature vector x_t, internal state $q_{s(u_t),t-1}$, and external state $r_{s(u_t),t-1}$ of the speaker from the immediate previous time-step (just before uttering the utterance) are concatenated and serve as the input vector for GRU_C.

$$c_t = GRU_C(c_{t-1}, (x_t \oplus q_{s(u_t),t-1} \oplus r_{s(u_t),t-1}))$$
(1)

We also pool attention vector a_t from the history of context $[c_1, c_2, \ldots, c_{t-1}]$ using soft-attention. This attention vector is later used to perform updates on internal and external states.

$$u_i = tanh(W_s c_i + b_s), \ i \in [1, t-1]$$
$$\alpha_i = \frac{exp(u_i^T x_i)}{\sum_{i=1}^{t-1} exp(u_i^T x_i)}$$
$$a_t = \sum_{i=1}^{t-1} \alpha_i c_i$$
(2)

Internal State: The internal state of the participants is conditioned on how the individual is feeling and what is the effect perceived from other participants. This state may remain concealed, as participants may not always express explicitly their feeling or outlook through external stance or reactions. Apart from feelings, this state can also be considered to include aspects that the participant actively tries not to express or features that are considered common knowledge and don't require explicit communication. The effect on oneself is thus elemental to represent the internal state of the participants. We model the internal state of the participants using GRU_Q. For time-step t, the internal state of the speaker $p_{s(u_t)}$ is updated by taking into account the attention vector a_t and commonsense vector *effect on speaker* $\mathcal{ES}_{cs}(u_t)$

$$q_{s(u_t),t} = GRU_Q(q_{s(u_t),t-1}, (a_t \oplus \mathcal{ES}_{cs}(u_t)))$$
(3)

For all the other participants apart from the speaker, this update is performed using *effect on listeners* $\mathcal{EL}_{cs}(u_t)$.

$$q_{j,t} = GRU_Q(q_{j,t-1}, (a_t \oplus \mathcal{EL}_{cs}(u_t))); \forall j \neq s(u_t) \tag{4}$$

External State: Unlike the internal state, the external state of the participants is all about the expressions, reactions, and responses. Naturally, this state can be easily seen, felt, or understood by the other participants. For instance, the actual utterance, the manner of articulation, the speech, and other acoustic features, the visual expression, gestures, and stance can all be loosely considered to fall under the regime of external state. GRU_R updates the external state of the speaker $p_{s(u_t)}$ by taking as input the concatenation of attention vector a_t, utterance vector x_t and commonsense vector *reaction of speaker* $\mathcal{RS}_{cs}(u_t)$

$$r_{s(u_t),t} = GRU_R(r_{s(u_t),t-1}, (a_t \oplus x_t \oplus \mathcal{RS}_{cs}(u_t))) \tag{5}$$

For listeners, this update is performed using *reaction of listeners* $\mathcal{RL}_{cs}(u_t)$.

$$r_{j,t} = GRU_R(r_{j,t-1}, (a_t \oplus x_t \oplus \mathcal{RL}_{cs}(u_t)));$$
$$\forall j \neq s(u_t) \tag{6}$$

Intent State: Intent is a mental state that represents the commitment to carry out a particular set of actions. The intent of the speaker always plays a crucial role in determining the emotional dynamics of a conversation. The intent of the speaker changes from $i_{s(u_t),t-1}$ to $i_{s(u_t),t}$ at time-step t. This change is invoked by the commonsense *intent of speaker* vector $\mathcal{IS}_{cs}(u_t)$ and internal speaker state $q_{s(u_t),t}$ at that respective time-step t. The intent states are captured by GRU cell GRU_I:

$$i_{s(u_t),t} = GRU_I(i_{s(u_t),t-1}, (\mathcal{IS}_{cs}(u_t) \oplus q_{s(u_t),t})) \tag{7}$$

The intent of the listener(s), however, is kept unchanged. This is because the intent of a participant who is silent should not change. The change should occur only when the particular participant speaks again.

$$i_{j,t} = i_{j,t-1}; \forall j \neq s(u_t) \tag{8}$$

Emotion State: The emotional state determines the emotional mood of the speaker and the emotion class of the utterance. We posit that the emotional state depends upon the utterance and composite

state of the speaker that takes into account the internal, external, and intent state. Naturally the current emotion state also depends on the previous emotion state of the speaker. GRU_E captures the emotion state by combining all of the factors as following,

$$e_t = GRU_E(e_{t-1}, (x_t \oplus q_{s(u_t),t} \oplus r_{s(u_t),t} \oplus i_{s(u_t),t})) \tag{9}$$

Emotion Classification: Finally all the utterances in the conversation are classified with a fully connected network from e_t

$$P_t = softmax(W_{smax}e_t + b_{smax}); \forall t \in [1, N]$$
$$\hat{y}_t = \underset{k}{\mathrm{argmax}}(\mathcal{P}_t[k]) \tag{10}$$

4 Experimental Setup

4.1 Datasets

Dataset	# dialogues			# utterances		
	train	val	test	train	val	test
IEMOCAP	120	12	31	5810		1623
DailyDialog	11,118	1,000	1,000	87,832	7,912	7,863
MELD	1,039	114	280	9,989	1,109	2,610
EmoryNLP	659	89	79	7,551	954	984

Dataset	# classes	Metric
IEMOCAP	6	Weighted Avg. F1
DailyDialog	7*	Macro F1 and Micro F1
MELD	3 and 7	Weighted Avg. F1 over 3 and 7 classes
EmoryNLP	3 and 7	Weighted Avg. F1 over 3 and 7 classes

Table 3: Statistics of splits and evaluation metrics used in different datasets. In MELD and EmoryNLP evaluation is performed for 3 class (broad) and 7 class (fine-grained) classification. *Neutral** classes constitutes to 83% of the DailyDialog dataset. These are excluded when calculating the Micro F1 score.

We benchmark COSMIC on four different conversational emotion recognition datasets: i) **IEMO-CAP** (Busso et al., 2008) ii) **MELD** (Poria et al., 2019a) iii) **DailyDialog** (Li et al., 2017), and iv) **EmoryNLP** (Zahiri and Choi, 2018). IEMOCAP and DailyDialog are two-party datasets, whereas MELD and EmoryNLP are multi-party datasets. We report experimental results for conversational emotion recognition from the textual information for all four datasets. Information about the datasets is shown in Table 3.

IEMOCAP (Busso et al., 2008) is a dataset of two person conversations among ten different unique speakers. The train set dialogues come from the first eight speakers, whereas the test set dialogues are from the last two. Each utterance is

annotated with one of the following six emotions: *happy, sad, neutral, angry, excited,* and *frustrated.*

DailyDialog (Li et al., 2017) covers various topics about our daily life and follows the natural human communication approach. All utterances are labeled with both emotion categories and dialogue acts. The emotion can belong to one of the following seven labels: *anger, disgust, fear, joy, neutral, sadness*, and *surprise*. The dataset has over 83% *neutral* labels and these are excluded during Micro-F1 evaluation.

MELD (Poria et al., 2019a) is a multimodal dataset extended from the EmotionLines dataset (Chen et al., 2018). MELD is collected from the TV show *Friends* and has more than 1400 dialogues and 13000 utterances. Utterances are labeled with emotion and sentiment classes. The emotion classes belong to *anger, disgust, sadness, joy, surprise, fear*, or *neutral*, and the sentiment classes belong to *positive, negative* or *neutral*.

EmoryNLP (Zahiri and Choi, 2018) is another dataset also based on the show *Friends*. Utterances in this dataset are annotated on seven and three emotion classes. The seven emotion classes are *neutral, joyful, peaceful, powerful, scared, mad* and *sad*. To create three emotion classes: *joyful, peaceful*, and *powerful* are grouped together to form the *positive* class; *scared, mad* and *sad* are grouped together to form the *negative* class; and the *neutral* class is kept unchanged.

4.2 Training Setup

For context independent feature extraction, the RoBERTa model is fine-tuned on the set of all utterances and their emotion labels in the training data. We fine-tune the RoBERTa model for a batch size of 32 utterances with Adam optimizer with learning rate of 1e-5. In the case of MELD and EmoryNLP datasets, we use a residual connection between the first and the penultimate layer which brings more stability in the training in the emotion recognition model. The emotion recognition model is trained with Adam optimizer having a learning rate of 1e-4.

5 Results and Analysis

5.1 Baseline and State-of-the art Methods

For a comprehensive evaluation of COSMIC, we compare it against the following methods: **CNN** (Kim, 2014) is a convolutional neural network model trained on top of pretrained GloVe

embeddings. Standard configurations of filter sizes are used. The model is trained at the utterance level to predict the emotion classes. **ICON** (Hazarika et al., 2018b) uses two GRU networks to learn the utterance representations for dialogues between two-participants. The output of the two speaker GRUs is then connected using another GRU that helps in performing explicit inter-speaker modeling. ICON is limited to conversations with only two participants only. **KET** (Zhong et al., 2019) or Knowledge enriched transformers dynamically leverages external commonsense knowledge using hierarchical self-attention and context aware graph attention. **ConGCN** (Zhang et al., 2019) considers utterances and participants of a conversation as nodes of graph network and models both context and speaker sensitive dependence for emotion detection. **BERT DCR-Net** (Qin et al., 2020) is a deep co-interactive relation network that uses BERT based features for joint dialogue act recognition and emotion (sentiment) classification. A relation layer learns to explicitly model the relation and interaction between these two tasks in a multi-task setting. **BERT+MTL** (Li et al., 2020) is a multi-task learning framework where features extracted from BERT are used in a recurrent neural network for emotion recognition and speaker identification. **DialogueRNN** (Majumder et al., 2019) models the emotion of utterances in a conversation with speaker, context and emotion information from neighbour utterances. These factors are modeled using three separate GRU networks to keep track of the individual speaker states.

We report and compare the performance of COSMIC on test data in Table 4. State-of-the-art models use GloVe embeddings to extract context-independent features. As features extracted from transformer based networks such as BERT and RoBERTa generally outperform traditional word embeddings such as word2vec and GloVe, we also report results of the models when used with BERT or RoBERTa features.

5.2 Comparison with the State-of-the-Art Methods

IEMOCAP and DailyDialog: IEMOCAP and DailyDialog contain dyadic conversations with mostly natural and coherent utterances. We observe that RoBERTa features improve the DialogueRNN models, and other BERT based models perform similarly. COSMIC improves over all the models,

	Methods	IEMOCAP	DailyDialog		MELD		EmoryNLP	
		W-Avg F1	Macro F1	Micro F1	W-Avg F1 (3-cls)	W-Avg F1 (7-cls)	W-Avg F1 (3-cls)	W-Avg F1 (7-cls)
GloVe-based	CNN	52.04	36.87	50.32	64.25	55.02	38.05	32.59
	ICON	58.54	-	-	-	-	-	-
	KET	59.56	-	53.37	-	58.18	-	34.39
	ConGCN	-	-	-	-	57.40	-	-
	DialogueRNN	62.57	41.80	55.95	66.10	57.03	48.93	31.70
(Ro)BERT(a)-based	BERT DCR-Net	-	48.90	-	-	-	-	-
	BERT+MTL	-	-	-	-	61.90	-	35.92
	RoBERTa	54.55	48.20	55.16	72.12	62.02	55.28	37.29
	RoBERTa DialogueRNN	64.76	49.65	57.32	72.14	63.61	55.36	37.44
	COSMIC	**65.28**	**51.05**	**58.48**	**73.20**	**65.21**	**56.51**	**38.11**
	w/o Speaker CSK	63.27	50.18	57.45	72.94	64.41	55.46	37.35
	w/o Listener CSK	65.05	48.67	58.28	72.90	64.76	**56.57**	**38.15**
	w/o Speaker, Listener CSK	63.05	48.68	56.16	72.62	64.28	55.34	37.10

Table 4: Comparison of results against various methods. Scores are average of five runs. Test scores are computed at best validation scores. COSMIC achieves new state-of-the-art results across all the datasets. CSK refers to commonsense knowledge components from COMET. We report the average score of the 10 runs for RoBERTa DialogueRNN and COSMIC. The CNN and DialogueRNN scores using Glove embeddings are obtained from (Ghosal et al., 2020).

however the improvement on IEMOCAP is not as large as it is on DailyDialog. COSMIC achieves new state-of-the-art scores of 65.28 on IEMOCAP; 51.05 and 58.48 in DailyDialog for the two different evaluation metrics.

MELD and EmoryNLP: These two datasets have been annotated from the TV show Friends, and utterances are often very short. Although dialogues occasionally contain emotion specific words, this does not happen very often at the utterance level. Naturally, emotion dynamics are highly contextual in nature and almost always depend on surrounding utterances. It has been observed in previous work that emotion modeling in MELD is difficult because often there are a lot of speakers in each conversation but they utter only a small number of utterances. Sophisticated models such as DialogueRNN do not bring as much improvement over CNN as they do on IEMOCAP. We observe that, COSMIC brings a large improvement over other models on the fine-grained (7 class) classification setup for both datasets. It achieves new state-of-the-art weighted F1 scores of 73.20 and 56.51 on three class classification; 65.21 and 38.11 on seven class classification on MELD and EmoryNLP.

5.3 The Role of Commonsense

In Table 4, we also report results of ablation studies by removing listener-specific and speaker-specific commonsense components. For speaker ablation,

we discard $\mathcal{IS}_{cs}(u_t), \mathcal{ES}_{cs}(u_t), \mathcal{RS}_{cs}(u_t)$, and observe a sharp drop in performance in most cases. For listener ablation, we discard $\mathcal{EL}_{cs}(u_t)$, and $\mathcal{RL}_{cs}(u_t)$ and find that the performance also drops but not as much as the speaker ablation. In fact, listener ablation leads to slight improvement in performance in EmoryNLP. The results suggest that speaker-specific commonsense has a greater impact in the overall performance of COSMIC, which is expected because we are predicting the emotion class of the speaker at each utterance. Finally, ablation with respect to both components at the same time naturally leads to higher drop in overall performance.

5.4 Case Study

We illustrate a case study on a test conversation instance from the IEMOCAP dataset in Figure 3. The conversation begins with a couple of *neutral* utterances, but then the situation quickly escalates, and finally, it ends with a lot of *angry* and *frustrated* utterances from both the speakers. State-of-the-art models like DialogueRNN often find this kind of scenarios difficult, when there is a couple of sudden emotions shifts in between (*neutral* to *frustrated* and then *neutral* again). These models also tend to misclassify utterances that have subtle differences in emotion classes such as *frustrated* and *angry*. In COSMIC, the propagation of commonsense knowledge makes it easier for the model to handle the
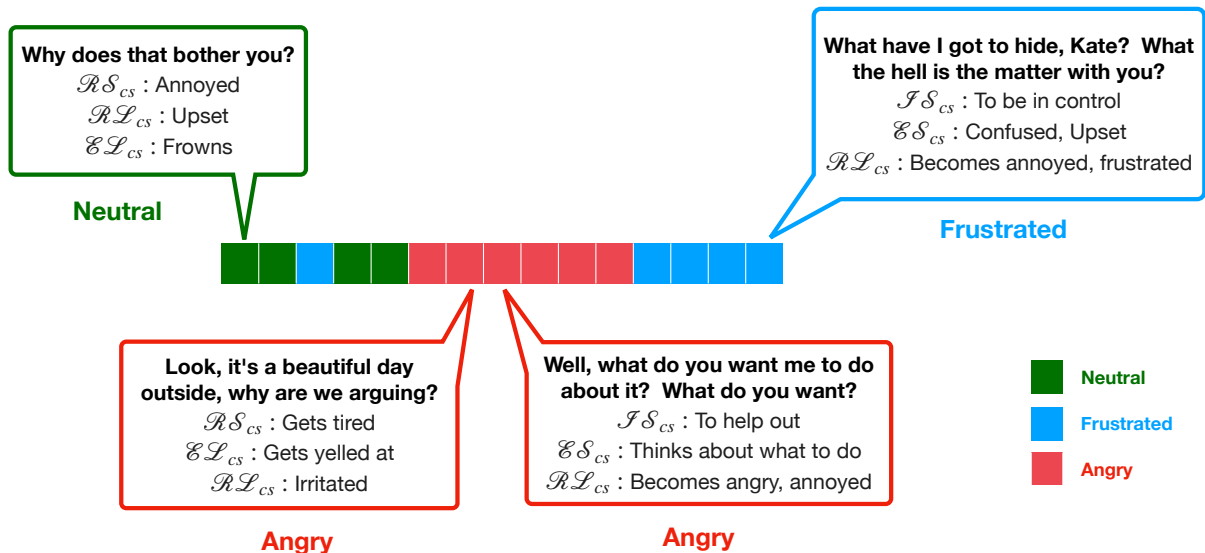
Figure 3: Case study from the IEMOCAP dataset. Discrete commonsense sequences are shown for more interpretability. Commonsense knowledge helps in predicting emotion shifts and understanding difference between closely related emotion classes such as *angry* and *frustrated*.

sudden transitions and to understand the subtle difference between closely related emotion classes. In Figure 3, for the first utterance, the commonsense model predicts that the *reaction of speaker* is *annoyed* and propagation of this information helps in predicting that the speaker's next utterance actually belongs to the *frustrated* class. Similarly for the rest of the illustrated utterances, the commonsense knowledge from *effect on speaker* and *reaction of listener* helps the model in distinguishing and predicting the *anger* and *frustrated* classes correctly.

5.5 Strategies to Incorporate Commonsense

Apart from the five commonsense features that we use in COSMIC (Table 1), there are four other features that can be extracted from COMET: *attribute of speaker*, *need of speaker*, *wanted by speaker*, and *wanted by listeners*. We incorporate them using different strategies that add extra complexity in our framework but ultimately do not improve the performance by a significant margin. We experimented along the following directions:

• *Attribute of speaker* is loosely considered as a personality trait. This latent variable influenced the *internal, external* and *intent states*. We find that the discrete *attribute* features from COMET are mostly a single word like 'stubborn', 'patient', 'argumentative', 'calm', etc and they change quite abruptly for the same participant in continuing utterances. Hence, we find that their vectorized representations do not help much.

• *Need of speaker*, *wanted by speaker*, and *wanted by listeners* are considered as output variables that are to be predicted from the input utterance and the five basic commonsense features (Table 1). We add auxiliary output functions and jointly optimize the emotion classification loss with mean-squared loss between predictions and reference commonsense vectors. This strategy also does not help much in improving the emotion classification performance.

Although the performance improvement is observed using commonsense knowledge across the datasets, this improvement is not very substantial. In the future, we plan to identify better commonsense knowledge sources and develop models that can infuse this knowledge into deep learning models more efficiently.

6 Conclusion

In this work, we presented COSMIC, a framework that models various aspects of commonsense knowledge by considering mental states, events, actions, and cause-effect relations for emotion recognition in conversations. Using commonsense representations, our model alleviates issues such as difficulty in detecting emotion shifts and misclassification between related emotion classes that are often present in current RNN and GCN based methods. COSMIC achieves new state-of-the-art results for emotion recognition across several benchmark datasets.

Acknowledgements

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4(0).

Mahmoud Azab. 2019. *Multimodal Character Representation for Visual Story Understanding*. Ph.D. thesis, University of Michigan.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

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):335–359.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171. ACM.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Niko Colnerĉ and Janez Demsar. 2018. Emotion recognition on twitter: comparative study and training a unison model. *IEEE Transactions on Affective Computing*.

Dragos Datcu and Leon JM Rothkrantz. 2014. Semantic audio-visual data fusion for automatic emotion recognition. *Emotion recognition: a pattern analysis approach*, pages 411–435.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2020. Utterance-level dialogue understanding: An empirical study. *arXiv preprint arXiv:2009.13902*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP 2014*, pages 1746–1751.

Bernhard Kratzwald, Suzana Ilic, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. 2018. Decision support with text-based emotion recognition: Deep learning for affective computing. *arXiv preprint arXiv:1803.06397*.

Jingye Li, Meishan Zhang, Dong-Hong Ji, and Yijiang Liu. 2020. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *ArXiv*, abs/2003.01478.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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*, volume 33, pages 6818–6825.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Veronica Pérez-Rosas, Rada Mihalcea, Ken Resnicow, Satinder Singh, Larry An, Kathy Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the European Association for Computational Linguistics (EACL 2017)*, Valencia, Spain.

Rosalind W Picard. 2010. Affective computing: from laughter to ieee. *IEEE Transactions on Affective Computing*, 1(1):11–17.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

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

Libo Qin, Wanxiang Che, Yangming Li, Mingheng Ni, and Ting Liu. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *Proceedings of the Association for Computatinal Linguistics*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, and Shady Elbassuoni. 2014. Emotion recognition from text based on automatically generated rules. In *2014 IEEE International Conference on Data Mining Workshop*, pages 383–392. IEEE.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. 2006. Emotion recognition from text using semantic labels and separable mixture models. *ACM transactions on Asian language information processing (TALIP)*, 5(2):165–183.

Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee. 2019. An interaction-aware attention network for speech emotion recognition in spoken dialogs. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6685–6689. IEEE.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory Fusion Network for Multi-view Sequential Learning. In *AAAI Conference on Artificial Intelligence*, pages 5634–5641.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5642–5649.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the*

Thirty-Second AAAI Conference on Artificial Intelligence.

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 *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5415–5421. AAAI Press.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.