Emotions Relationship Modeling in the Conversation-Level Sentiment Analysis

Jieying Xue, Minh-Phuong Nguyen and Le-Minh Nguyen
Japan Advanced Institute of Science and Technology, 923-1292, 1-8 Asahidai, Nomi, Ishikawa, Japan

Keywords: Sentiment Analysis, Emotion Recognition in Conversation, COSMIC, Emotion Dependencies, Transformer.

Abstract: Sentiment analysis, also called opinion mining, is a task of Natural Language Processing (NLP) that aims to extract sentiments and opinions from texts. Among them, emotion recognition in conversation (ERC) is becoming increasingly popular as a new research topic in natural language processing (NLP). The current state-of-the-art models focus on injecting prior knowledge via an external commonsense extractor or applying pre-trained language models to construct the utterance vector representation that is fused with the surrounding context in a conversation. However, these architectures treat the emotional states as sequential inputs, thus omitting the strong relationship between emotional states of discontinuous utterances, especially in long conversations. To solve this problem, we propose a new architecture, ***L**ong-range dependenc**Y** emotion**S** **M**odel (LYSM)* to generalize the dependencies between emotional states using the self-attention mechanism, which reinforces the emotion vector representations in the conversational encoder. Our intuition is that the emotional states in a conversation can be influenced or transferred across speakers and sentences, independent of the length of the conversation. Our experimental results show that our proposed architecture improves the baseline model and achieves competitive performance with state-of-the-art methods on four well-known benchmark datasets in this domain: IEMOCAP, DailyDialog, Emory NLP, and MELD. Our code is available at https://github.com/phuongnm94/erc-sentiment.

1 INTRODUCTION

Emotion recognition in conversation, as a crucial research topic in natural language processing (NLP), it has received increasing attention (Poria et al., 2017; Zhang et al., 2019; Ghosal et al., 2020a; Guibon et al., 2021; Song et al., 2022). Unlike ordinary sentence or utterance emotion recognition, ERC ideally requires context modeling of individual utterance. This context can be attributed to the preceding utterances and relies on the temporal sequence of utterances.

Since ERC relies heavily on temporal order-based context, therefore, previous works (Poria et al., 2017; Majumder et al., 2019; Ghosal et al., 2019; Zhang et al., 2019) applied recurrent neural network (RNN) to process the constituent utterances of a conversation in sequence. Besides, with the success of pre-trained language models (Devlin et al., 2019; Liu et al., 2019), recent works (Guibon et al., 2021; Lee and Choi, 2021; Song et al., 2022) integrate contextual information by connecting surrounding utterances for the current utterance encoding process. Furthermore, many works (Ghosal et al., 2019; Lee and Choi, 2021) tend to leverage the relationships between speakers in a conversation and apply the graph neural network to improve the performance. In another aspect, a proposed framework, COSMIC (Ghosal et al., 2020a), applies a commonsense knowledge extractor to collect additional useful features of utterance representations, such as *the intent* or *reaction of speaker*, etc.

This work explores the contribution of dependencies between emotional states among utterances in a conversation. The intuition is that the emotional states can be affected or transferred between speakers in a conversation, regardless of the length of the conversation. The relationship among the emotional states is an essential aspect of the ERC system (Song et al., 2022; Guibon et al., 2021; Kim and Vossen, 2021). However, these previous approaches only consider the transfer of emotion between the adjacent utterances in a conversation and thus omit the dependencies of emotion states in long-range utterances. In another aspect, the emotional states in conversations are typically transferred between speakers. For example, funny people usually have positive emotions in their sentences and could transmit that to those who

Xue, J., Nguyen, M. and Nguyen, L.
Emotions Relationship Modeling in the Conversation-Level Sentiment Analysis.
DOI: 10.5220/0011658200003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 276-284
ISBN: 978-989-758-623-1; ISSN: 2184-433X

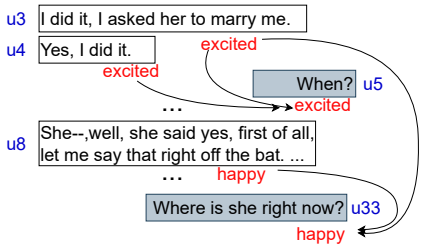

Figure 1: Example of the long-range emotions dependencies in a conversation (IEMOCAP dataset). The blue words following the template $u<id>$ with $<id>$ is the index of utterance in the conversation.

talk to them (Figure 1). The emotional states can be transferred between speakers in a conversation via adjacent utterances (e.g. `u3, u4, u5`) or long-range distance utterances (e.g. `u3, u8, u33`). The questions of the right speaker (`u5, u33`) are normal emotions if stand-alone, however, in particular contexts, these utterances are affected by the feelings of the other speaker regardless of their position in the conversation. Accordingly, our model can enhance the performance of the ERC system by learning the attention between the pairs of emotions in a conversation and achieving rich transcription.

Therefore, we propose a new architecture LYSM, to model the dependencies between the emotional states applying the self-attention mechanism to boost the robust baseline model based on the COSMIC framework. We also conducted experiments to evaluate the performance of our proposed model compared with previous methods on four popular benchmark datasets: IEMOCAP, DailyDialog, Emory NLP, and MELD. Experimental results showed that our proposed model works effectively and achieves competitive results with the current SOTA results, and outperform the baseline models on all experimental datasets.

2 RELATED WORK

Overview of ERC Task. Currently, most dialogue emotion recognition tasks are based on three major innovations: recurrent neural networks-based (RNNs) (Hochreiter and Schmidhuber, 1997), graph convolution network-based (GCN) (Defferrard et al., 2016) and self-attention-based (Devlin et al., 2019). There is a fact that contextual information plays an important role in understanding the meaning of utterances in a conversation, and RNNs architecture (like LSTMs and GRUs) have achieved great improvements in ERC (Poria et al., 2017; Ghosal et al., 2020a) because it can capture context as sequential information. Besides, in some works, the utterance content and speaker identity are encoded to capture sentence semantics better. On the other hand, GCNs has also attracted many recent works (Ghosal et al., 2019; Zhang et al., 2019; Lee and Choi, 2021) to accomplish this task by capturing the relationships between interlocutors and the dependence of utterance on the speakers and the listeners. However, these works have not considered the emotional dependencies between utterances, while this feature plays an important role in conversational sentiment detection.

Lastly, with the success of the pre-trained models in many NLP tasks (Vaswani et al., 2017; Devlin et al., 2019), the ERC tasks have also been applied widely in many recent works (Kim and Vossen, 2021; Ghosal et al., 2020a; Lee and Choi, 2021; Song et al., 2022). Most of these works use the self-attention mechanism at words-level to utterance encoding (Kim and Vossen, 2021; Song et al., 2022; Ghosal et al., 2020a) and capture the information in the whole context or localize context of each sentence in a conversation. Compared with these works, our work applies a self-attention mechanism over emotional states throughout a whole conversation to learn the strong effect of emotions between inter-speakers.

[Object Promise]. There have been a number of recently proposed models showing improvements in affective dependence in ERC tasks (Guibon et al., 2021; Lee and Choi, 2021; Song et al., 2022). Most of these works apply a CRF layer on the top of the deep learning model, which is typically applied to sequence labeling tasks in NLP. In another aspect, these works (Kim and Vossen, 2021; Lee and Choi, 2021; Song et al., 2022) use the self-attention mechanism to encode the dependencies between words in the limited context of current utterance, while our LYSM apply the self-attention mechanism to model the emotional states in the whole conversation. The closest model to our LYSM is the EmotionFlow model (Song et al., 2022). However, the EmotionFlow only considers the emotional relations of adjacent utterances, while our LYSM model can capture the dependencies among the emotional state of all utterances in a conversation. To this end, we also conduct experiments to compare with the approaches using the CRF to demonstrate the effectiveness of our proposed model.

3 METHODOLOGY

In this section, we detail our proposed model architecture, LYSM, based on the COSMIC framework. To model the strong relationship between emotions in the conversation, we proposed to use the self-attention

mechanism (Vaswani et al., 2017) and contrast it with the Conditional Random Field (CRF), which adapted from the idea of previous works (Song et al., 2022) on the top of the COSMIC framework. The whole system contains two main components: (1) the conversational encoding component to transform the utterances in conversation into the hidden vector representation, and (2) the emotional dependency encoding component to learn the effect of emotional relationships in conversation.

3.1 Task Definition

Given a conversation containing the sequence of utterances and corresponding speakers $[(u_t, p_t)]_{t=1}^{N}$ with N utterances, the target is to identify the emotion of each utterance (y_t) from the pre-defined set of emotions, such as *happy, sad,* etc. To represent the hidden vectors of each input sentences of speakers and their corresponding emotions simultaneously, and improve the sentiment analysis system based on the aforementioned information, we propose LYSM architecture to learn the dependencies between the utterances of the speakers in a conversation. Specifically, we leverage the current state-of-the-art model in this area as a strong baseline, COSMIC (Ghosal et al., 2020a), and build on it to model the layer relationships between emotion vectors via attention mechanism (Figure 2).

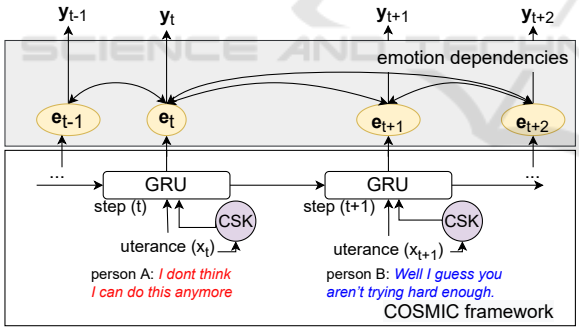

Figure 2: The architecture of the ***L**ong-range dependenc**Y** emotion**S** **M**odel (LYSM)* based on COSMIC framework in the *conversation-level emotion recognition* task. The *CSK* components indicate the external CommonSense Knowledge extractor module.

3.2 COSMIC Framework

This framework (Ghosal et al., 2020a) aims to inject prior commonsense knowledge into the emotion recognition process. The knowledge features were extracted by an external tool based on the commonsense knowledge graph, COMET (Bosselut et al., 2019). The main important part of this framework is that the authors reveal the contribution of special commonsense features to the emotion recognition task, which contains *the intent of speaker* $\mathcal{IS}_{cs}(u_t)$, *the effect of speaker* $\mathcal{ES}_{cs}(u_t)$, *the reaction of speaker* $\mathcal{RS}_{cs}(u_t)$, *the effect of listeners* $\mathcal{EL}_{cs}(u_t)$, and *the reaction of listeners* $\mathcal{RL}_{cs}(u_t)$. Furthermore, they also proposed a novel architecture using these features to identify the emotion of utterances effectively.

For representing sentences in the conversation, (Ghosal et al., 2020a) firstly fine-tune a pre-trained language model (e.g. RoBERTa) on the emotion classification task without considering the context. Then they use the fine-tuned model to generate the continuous vector of utterances. Similar to the conventional BERT architecture (Devlin et al., 2019), a token *[CLS]* is added at the beginning of the sentence to represent the meaning of the whole natural sentence. In addition, the authors use the average of four *[CLS]* hidden vectors in the last layers to get the final representation for an utterance ($\mathbf{x}_t$).

For modeling the sequential features in the conversation, this framework uses the GRU cells (Chung et al., 2014) to represent hidden states that affect the emotion of human sentences. There are five different features are constructed sequentially along with utterances in the conversation: *context state*, *internal state*, *external state*, *intent state*, and *emotion state*; these states are encoded by five separated GRU cells, GRU_C, GRU_Q, GRU_R, GRU_I, and GRU_E respectively. For mathematical operation, firstly, the context vector (c_t) is computed based on sentence vector (x_t), previous internal state ($q_{s(u_t),t-1}$) and previous external state ($r_{s(u_t),t-1}$):

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

where $\oplus$ is the concatenation function. Then, a soft attention vector (a_t) is introduced to update the internal and external hidden vectors:

$$u_i = tanh(W_s c_i + b_s), \quad i \in [1, t-1] \quad (2)$$

$$\alpha_i = \sum_{i=1}^{t-1} \left(\frac{exp(u_i^\intercal x_i)}{\sum_j^{t-1} exp(u_j^\intercal x_j)}\right) c_i \quad (3)$$

where W_s, b_s are learnable parameters. Then, the internal ($q_{s(u_t),t}$), external ($r_{s(u_t),t}$) intent ($i_{s(u_t),t}$) and emotion (e_t) states are computed based on the previous states incorporating with the commonsense knowledge and soft attention vectors:

$$q_{s(u_t),t} = GRU_Q(q_{s(u_t),t-1}, (a_t \oplus \mathcal{ES}_{cs}(u_t))) \quad (4)$$

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

$$i_{s(u_t),t} = GRU_I(i_{s(u_t),t-1}, (\mathcal{IS}_{cs}(u_t) \oplus q_{s(u_t),t})) \quad (6)$$

$$e_t = GRU_E(e_{t-1}, x_t \oplus q_{s(u_t),t} \oplus r_{s(u_t),t} \oplus i_{s(u_t),t}) \quad (7)$$

In addition, the states of listeners also are updated for each utterance (u_t) with the replacement of $\mathcal{ES}_{cs}(u_t)$, $\mathcal{RS}_{cs}(u_t)$ by $\mathcal{EL}_{cs}(u_t)$, $\mathcal{RL}_{cs}(u_t)$ in Equations 4, 5, respectively. Then, the emotion label probabilities of the current utterance are calculated via a *softmax* layer based on the emotion vector e_t:

$$p_t = \text{softmax}(W_e e_t + b_e) \tag{8}$$

where W_e, b_e are learnable parameters. Finally, the probabilities (p_t) of all sentences in the conversation are forwarded to compute the negative log likelihood loss, and the model is trained based on the back-propagation algorithm.

3.3 LYSM Architecture

In this work, we proposed a variant ***L**ong-range dependenc**Y** emotion**S** **M**odel (LYSM)* architecture based on COSMIC framework, which can learn the strong dependencies between emotional states in a conversation. We surmise that the emotional states in the conversation strongly affect each other. For example, the emotion of some certain utterances in the head conversation can affect the utterances in the middle or last position of the conversation. However, in the COSMIC framework, the emotional states in the long conversation are encoded by recurrent architecture that is not directly connected to each other. Therefore, in our proposed architecture, LYSM, we utilize the power of the COSMIC framework to get the utterance vector representation fused by commonsense knowledge and construct a new component to explore the strong relations between emotional states (e_t) via the Transformer Encoder layer with the Self-Attention mechanism (Vaswani et al., 2017).

For mathematical, the sequence of emotional states ($e = [e_t]_{t=1}^{N}$) taken from Equation 7 is fed to the Transformer Encoder layer to get new representation fused by emotion context dependencies.

$$\begin{aligned} g_j^q, g_j^k, g_j^v &= eW_j^q, eW_j^k, eW_j^v, \quad j \in [1, \#\text{heads}] \\ head_j &= \text{softmax}(\frac{g_j^q \cdot (g_j^k)^\intercal}{\sqrt{d_h}}) g_j^v \\ g_{mul} &= (head_1 \oplus head_2 \oplus ... \oplus head_{\#heads}) W^o \\ g_{norm} &= \text{LayerNorm}(g_{mul} + e) \\ e' &= \text{LayerNorm}(\text{FFW}(g_{norm}) + g_{norm}) \end{aligned} \tag{9}$$

where *#heads* is the number of heads in Multi-head layer, d_h is the dimension size of per head, LayerNorm and FeedForward(FFW) are the functions that are used similarly to (Vaswani et al., 2017). Finally, the new emotional state vector $e' = [e'_t]_{t=1}^{N}$ is used to compute the probabilities of emotion label by *softmax* layer, similar to Equation 8:

$$p_t = \text{softmax}(W_e e'_t + b_e) \tag{10}$$

3.4 Conditional Random Field

This architecture is typically applied for sequence labeling tasks such as POS tagging, Named Entity Recognition (Ma and Hovy, 2016). To model the dependencies between emotions in a conversation, previous works (Song et al., 2022; Guibon et al., 2021) built a CRF layer as the last layer of the Neural Network model. Therefore, it is potential to adapt this architecture to the COSMIC framework for comparison with our proposed model LYSM. In detail, we treat the sentiment vector representation (e_t) as the emission score of each utterance for all emotional labels, and the transmission score that is considered as the influence between emotions is random initial values. After that, these weights are learned end-to-end in the training process.

$$\begin{aligned} \text{score}(e, y) &= \sum_{t=1}^{N} (W_{em} e_t + b_{em})[y_t] + \sum_{t=0}^{N} (W_{tr}[y_t, y_{t+1}]) \\ p(y|e) &= \frac{\exp(\text{score}(e, y))}{\sum_{y'} \exp(\text{score}(e, y'))} \end{aligned} \tag{11}$$

where y_0, y_{N+1} is additional start and end of emotional labels; $[\cdot]$ is the matrix selection operator given row and column indexes; W_{em}, b_{em}, W_{tr} are the learnable weights for emission and transmission scores; y' is a candidate of emotion flow in the set of possible emotion flows. By using CRF layer, the model is trained to maximize the log-probability of gold emotion sequence labels.

4 EXPERIMENT

In this section, we describe the detail of the experiments to evaluate the performance of our LYSM.

Dataset. We conducted experiments to evaluate the performance of our proposed architecture, LYSM on four benchmark datasets (Table 1):

- **IEMOCAP:** (Busso et al., 2008) is the dataset of six different emotion categories collected from conversations of ten different speakers, each conversation contains utterances of two persons.
- **DailyDialog:** (Li et al., 2017) is the largest multi-utterance dialogue dataset collected in daily life conversations, including seven different emotion categories. Following the previous work experimental setup, we ignore the label *neutral* when

compute the evaluation score because this label is highly imbalanced in 83% of utterances across the whole dataset.

- **MELD:** (Poria et al., 2019) and **EmoryNLP** (Zahiri and Choi, 2018) are the datasets of seven different types of emotions scraped from TV shows.

The IEMOCAP is the dataset that contains long dialogues with an average of around 50 utterances per conversation, while DailyDialog is the dataset that contains many topics in conversation. By conducting experiments on these various kinds of datasets, we can evaluate the generalization ability and measure the improvement of our proposed model compared with the baseline COSMIC framework.

Table 1: Statistics information on all ERC datasets. The character # denote the size of the set.

Dataset	# dialogues			# utterances		
	train	dev	test	train	dev	test
IEMOCAP	108	12	31	5,163	647	1,623
DailyDialog	11,118	1,000	1,000	87,823	7,912	7,836
MELD	1,039	114	280	9,989	1,109	2,610
EmoryNLP	659	89	79	7,551	954	984

Experimental Setup. Since our proposed model is constructed based on COSMIC model, therefore, we conducted experiments using the results of the COSMIC framework with the following steps: fine-tune the pre-trained language model for utterance representation and commonsense knowledge feature extraction. We use these continuous feature vectors which are equal to input features of the COSMIC framework, as the input to our LYSM architecture.

In these experiments, we aim to evaluate the effectiveness of our LYSM architecture compared with the original COSMIC framework. In addition, to compare with the EmotionFlow model (Song et al., 2022) related to emotional transference, we also conducted experiments incorporating the CRF layer on our proposed model.

For each dataset mentioned above, we run it ten times with different random seeds and compute our proposed model performance using the Weighted Average F1 score (Ghosal et al., 2020a). The best model is selected based on the dev set of each dataset and used to get the evaluation score on the test set. Then, we report the mean value of performance compared with the previous works on these datasets.

5 RESULT ANALYSIS

5.1 Main Results

We conducted experiments on four aforementioned datasets and show the results in Table 2. Our proposed model improved the performance of the COSMIC framework on all datasets. On the IEMOCAP dataset, the LYSM architecture improved the F1 score by 0.19 compared to the COSMIC framework. On the DailyDialog dataset, our proposed model improved Macro F1 score by 0.27 and Micro F1 score by 0.21 compared with the baseline model. On the EmoryNLP and MELD datasets, we only conducted experiments on the setting of emotion recognition tasks with seven different emotion classes, and the result shows an improvement of 0.23 F1 score and 0.19 on F1 score, respectively. These results show that our LYSM architecture is generalized and the emotion dependencies component can works effectively when incorporated into the COSMIC framework.

Compared with EmotionFlow, which uses CRF to learn emotion transfer, our model has more advantages because it is supported by the commonsense knowledge information based on the COSMIC framework. However, for a fair comparison between the CRF layer and the self-attention mechanism, and we showed the ablation study in section 5.2. Compared with the EmoBerta, our LYSM architecture achieved competitive results on the MELD dataset, but lower than the results on the IEMOCAP dataset. We argue that the reason comes from the model size of EmoBerta. While EmoBerta fine-tuning on Roberta large (Liu et al., 2019) contains 355 million parameters with 9 minutes training per epoch (Kim and Vossen, 2021), our LYSM used a fixed fine-tuned Roberta to get an utterance encoding vector that only contains 17 million parameters for the training process with 30 seconds per epoch.

5.2 Ablation Study

In this experiment, we inspect the effectiveness of the component learning emotional dependencies. In our LYSM architecture, we used the self-attention mechanism whereas the previous works (Song et al., 2022; Guibon et al., 2021) suggest using the CRF layer to model the emotion transfer between sequences of utterances in a conversation. Therefore, we conducted additional experiments by applying the CRF layer (+CRF) on both COSMIC and our model (Table 3). In Table 3, although we used the same setting as (Ghosal et al., 2020a), the different results of our re-produced COSMIC framework with them because of the ex-

Table 2: Performance comparison between methods. This table contains two parts: previous works and our results. The values in the below part refer to the results of the experiments implemented in this work.

Methods	**IEMOCAP**	**DailyDialog**		**MELD**	**EmoryNLP**
	W-Avg F1	Macro F1	Micro F1	W-Avg F1	W-Avg F1
DialogueRNN re-product (Ghosal et al., 2020b)	62.57	41.80	55.95	57.03	31.70
EmoBerta (Kim and Vossen, 2021)	**67.42**	-	-	**65.61**	-
EmotionFlow (Song et al., 2022)	65.05	-	-	-	-
COSMIC (Ghosal et al., 2020a)	65.28	51.05	58.48	65.21	38.11
LYSM (ours)	65.47	**51.32**	**58.69**	65.40	**38.34**

Table 3: Ablation study on the IEMOCAP dataset.

Methods	**IEMOCAP**
EmotionFlow (Song et al., 2022)	65.05
COSMIC	64.50
COSMIC +CRF	64.82
LYSM	**65.47**
LYSM +CRF	65.43

perimental environment such as libraries or computing servers. Similar to previous works, our experimental results confirmed the CRF layer also supports the COSMIC model with 0.32 F1 score improvement. However, when compared with our proposed model using the self-attention layer, the performance improvement is larger than with a 0.97 F1 score. We argue that self-attention can learn the influence between pairs of discontinuous utterances in a conversation, not just adjacent utterances like the CRF layer. In addition, we also apply the CRF layer to the LYSM architecture, but the results did not improve because the emotional dependencies information was already captured by the self-attention layer. These results demonstrated the effectiveness of our LYSM architecture and the importance of emotional dependencies in a conversation.

5.3 Result Analysis

Conversation Length. We conducted an analytica experiment to inspect the effect of the conversation length on the performance of the ERC system (Figure 3). We found that the performance of both the COSMIC framework and our LYSM tends to decrease as conversation length increases. In addition, these results also show that our LYSM obviously outperforms the baseline model across all groups of conversation length, which prove the generalization of our model. In particular,the improvement was noticeable during the long conversations. This result is the evidence for the effectiveness of our LYSM in capturing long-range emotion dependencies.

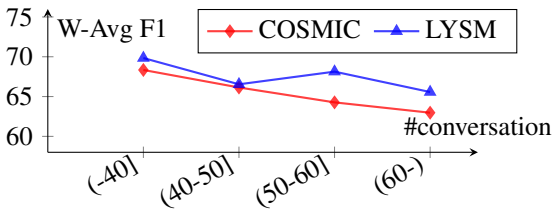

Figure 3: Performance comparison between our LYSM and COSMIC on IEMOCAP dataset with respect to the number of utterances in a conversation (#conversation) .

Improvement Example. Based on our observations on the output prediction results of the IEMOCAP dataset, we found that LYSM architecture generally works more effectively in strong emotional conversations. For better understanding the improvement of LYSM architecture, we present examples in Table 4 of different predictions on our LYSM architecture and the COSMIC model. In this conversation, there are two speakers who are in a negative mood with many `angry` labels in their sentences. We found that in the sentences containing strongly emotional words (*hell* in utterance `u30`), both LYSM architecture and COSMIC showed the correct predictions. However, the COSMIC framework can predict the utterances containing normal words (*u23*, *u24*, *u26*) as slight negative emotion (`frustrated`), but our LYSM based on the strong context relationships is more accurate for sentiment label prediction.

Learning Process. We reproduced the COSMIC framework and conducted experiments to inspect the weighted average F1 values of this model compared with our LYSM architecture on IEMOCAP dataset (Figure 5).

The weighted average F1 values of our proposed architecture on the development set are higher than the COSMIC framework in most epochs. These results demonstrated the generality and effectiveness of the LYSM architecture.

Table 4: Improvement example collected in IEMOCAP dataset. The green and red labels indicate the correct and incorrect prediction of the models, respectively.

Id	Utterance	Label	LYSM	COSMIC
u23	**S1:** You infuriate me sometimes. Do you know that? God.	angry	**angry**	**frustrated**
u24	**S1:** Isn't it your business, too, if dad – if I tell dad and he throws a fit about it? I mean, you have such a talent for ignoring things.	angry	**angry**	**frustrated**
u25	**S2:** I ignore what I got to ignore. I mean, the girl is Larry's girl.	angry	**frustrated**	**frustrated**
u26	**S1:** She is not Larry's girl!	angry	**angry**	**frustrated**
u27	**S2:** From your father's point of view he's not dead and she's still his girl. Now, you can go on from there if you know where to go, Chris, but I don't know. So what can I do for you?	angry	**frustrated**	**frustrated**
u28	**S1:** I don't know why it is but everytime I reach out for something I- that I want, I have to pull back because I might hurt somebody else. My whole bloody life; time after time after time.	frustrated	**frustrated**	**angry**
u29	**S2:** Well, you're a considerate fella, there's nothing wrong in that	neutral	**neutral**	**neutral**
u30	**S1:** To hell with that!	angry	**angry**	**angry**

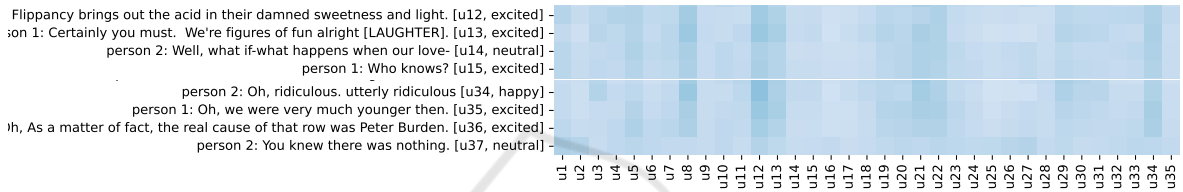

Figure 4: Heatmap visualization of dependencies between pairs of emotional states in a conversation. This figure shows the scaled Self-Attention in the LYSM architecture, computed in Equation 9. The title each row or column of this heatmap is an utterance (u_i) in a conversation. The bolder colors show higher attention scores.

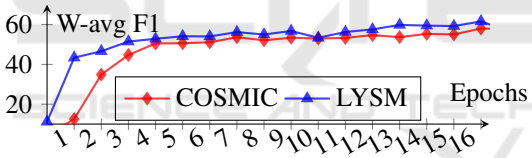

Figure 5: Weighted average F1 values of COSMIC and our LYSM architecture on IEMOCAP development set .

Emotional Dependency. In our LYSM architecture, we aim to model the dependencies between pairs of the emotional state of utterances to improve the performance of sentiment analysis system. Therefore, in this experiment, we depict the dependencies of emotional states pairs constructed by utterances in a conversation (Figure 4). We found that the dependencies among emotional states affect not only the adjacent sentences but also the remote sentences. For example, the emotion *happy* in utterance 34 (column u34) is affected by the utterances in the whole conversation including the beginning sentences. Besides, the special emotional states which are different from others in a conversation typically get more attention than the remainder, such as utterances u12, u13, u34. This clearly evidences emotion dependencies are important for emotion recognition systems.

6 CONCLUSION

In this work, we explore the importance of emotion dependency features in conversation-level emotion recognition tasks. We also proposed an effective model, LYSM, which incorporates a self-attention mechanism into the COSMIC framework to improve performance and achieve competitive results with the SOTA on four benchmark datasets IEMOCAP, DailyDialog, EmornyNLP, and MELD. Our model is simple yet effective that can be widely applied to other architectures in the sentiment recognition domain. In future work, we would like to apply the self-attention mechanism to model the emotional dependencies of each individual speaker in the conversation and explore the contribution of personality features in sentiment analysis.

ACKNOWLEDGEMENT

This work was supported by JSPS Kakenhi Grant Number 20H04295, 20K20406, and 20K20625.

REFERENCES

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. (2019). COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ghosal, D., Majumder, N., Gelbukh, A., Mihalcea, R., and Poria, S. (2020a). COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Ghosal, D., Majumder, N., Mihalcea, R., and Poria, S. (2020b). Utterance-level dialogue understanding: An empirical study. *CoRR*, abs/2009.13902.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. (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, Hong Kong, China. Association for Computational Linguistics.

Guibon, G., Labeau, M., Flamein, H., Lefeuvre, L., and Clavel, C. (2021). Few-shot emotion recognition in conversation with sequential prototypical networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6858–6870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Kim, T. and Vossen, P. (2021). Emoberta: Speaker-aware emotion recognition in conversation with roberta. *CoRR*, abs/2108.12009.

Lee, B. and Choi, Y. S. (2021). Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (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, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

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

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L.-P. (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.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). 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.

Song, X., Zang, L., Zhang, R., Hu, S., and Huang, L. (2022). Emotionflow: Capture the dialogue level emotion transitions. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8542–8546.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Zahiri, S. and Choi, J. D. (2018). Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. In *Proceedings of the*

AAAI Workshop on Affective Content Analysis, AFFCON'18, pages 44–51, New Orleans, LA.

Zhang, D., Wu, L., Sun, C., Li, S., Zhu, Q., and Zhou, G. (2019). Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.