Predicting Emotion Before an Utterance: Replacing Contextual Information with Knowledge-Based Prediction

Takumi Fujimoto and Takayuki Ito[a]
Department of Social Informatics, Kyoto University, Kyoto 6068317, Japan

Keywords: Emotion Prediction, Natural Language Processing.

Abstract: As textual conversation becomes more mainstream, emotional understanding that contributes to collective intelligence and team performance plays an essential role. However, most studies of emotion recognition in conversation (ERC) have focused on analyzing emotion after a participant speaks, and few have addressed predicting emotion in ongoing conversations before an utterance is made. In this paper, we tackle predicting emotion before a speaker expresses their emotion and propose utilizing knowledge-based prediction as an appropriate replacement for contextual information. We develop a context model to capture conversations and replace future conversational data that cannot be exploited in actual predictions with appropriate knowledge-based predictions. Our model outperformed existing models in predicting conversational emotions on a dataset composed of conversations between multiple participants.

1 INTRODUCTION

With the rapid development of Artificial Intelligence (AI), emotional understanding is an issue widely tackled in various fields, including Natural Language Processing (NLP) (Acheampong et al., 2021; Kumar and Garg, 2020; Poria et al., 2019b; Zhang et al., 2022; Zucco et al., 2020).

Thus, emotion has a wide range of potential applications (Yi and Wang, 2019; Wu et al., 2019; Khanpour and Caragea, 2018), and emotional understanding in communication, where people with different emotions interact with each other, plays a particularly significant role as there is a correlation between collective intelligence and empathy (Woolley et al., 2010) and a close relationship between team performance and psychological safety through emotional understanding (Edmondson, 2018).

The same is true for recent text-based communication, which has spread to many users with the proliferation of chatbots and social media. However, text-based conversations are problematic due to the loss of information such as the facial expressions, tone of voice, and mood of the speaker, that supports understanding of other participants' emotions in traditional communication. In terms of supporting the understanding of emotions in conversation, there is widespread research on Emotion Recognition in Conversation(ERC) (Gao et al., 2022; Ghosal et al., 2020; Lee and Lee, 2022; Li et al., 2022; Majumder et al., 2019; Saxena et al., 2022; Song et al., 2022; Tu et al., 2022; Zhong et al., 2019; Zhu et al., 2021). These studies aim to identify the emotions of the target utterance based on all statements made in the conversation. Therefore, this approach makes it possible to analyze emotions only after the communication is complete and makes it difficult to support the understanding of emotions in ongoing conversations. In this study, to support emotion understanding for textual communications in progress, we deal with Emotion Prediction in Conversation(EPC), which predicts the emotions of an utterance before the speaker makes that statement. The prediction has difficulties similar to recognition, like analyzing participants' emotions without emotional information such as facial expression, voice tone, conversational atmosphere, and so on. Moreover, in the prediction, future conversational information and utterance itself can not be used when the emotion of that utterance is predicted as opposed to the recognition studies. Though predicting has many challenges, it can contribute to understanding and considering other participants' emotions in conversation more than the recognition studies, because EPC focuses on predicting the emotions of participants in real time before a speaker makes an actual utterance. To handle the problem that future conver-

[a] https://orcid.org/0000-0001-5093-3886

Fujimoto, T. and Ito, T.
Predicting Emotion Before an Utterance: Replacing Contextual Information with Knowledge-Based Prediction.
DOI: 10.5220/0012307500003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 109-116
ISBN: 978-989-758-680-4; ISSN: 2184-433X

sational information cannot be used for prediction, we proposed to use knowledge-based emotion prediction as a replacement for future conversational information, and we developed a model to capture conversational contexts and replace future conversational data appropriately with knowledge-based predictions.

The remainder of this paper is organized as follows: Section 2 discusses related work; Section 3 provides a detailed description of our model; Section 4 presents and discusses the experimental results, and Section 5 concludes our paper.

2 RELATED WORK

Ekman's six basic emotions (surprise, fear, sadness, joy, disgust, and anger) (Ekman, 1992) are widely used in ERC (Chen et al., 2019; Poria et al., 2019a; Poria et al., 2019b). There are various other classifications of emotions (Zad et al., 2021), such as a method that shows the relationship between emotions in two dimensions, valence and arousal (Russell and Barrett, 1999), a way to categorize emotions in three dimensions according to complexity and intensity (Plutchik, 2001), and a means that classifies emotions into 20 categories defined by four independent dimensions (Cambria et al., 2012).

Recognizing emotions based on the conversational context is a habitual method in ERC (Majumder et al., 2019; Saxena et al., 2022), and using external knowledge has been undertaken recently to achieve more accurate recognition. For example, some studies recognize emotions by the commonsense knowledge graph in addition to the context and topic information of the conversation (Ghosal et al., 2020; Tu et al., 2022; Zhong et al., 2019; Zhu et al., 2021). Other studies utilize pre-trained language models as external knowledge and use them for classification with contextual information (Lee and Lee, 2022) and to generate conversation sequences (Li et al., 2022). Various other methods have been proposed in ERC, such as tracking emotional coherence by the Conditional Random Field (CRF) layer (Wang et al., 2020) and accounting for emotion shifts by focusing on the influence of other participants on a speaker (Gao et al., 2022; Song et al., 2022).

In many cases, the ERC targets the latest statement that is already made and is based on all statements made in a conversation. In contrast, EPC predicts the emotions of an utterance before the speaker makes that statement. Therefore, it cannot use the utterance text that is the subject of emotional analysis and future conversational information from that utterance onwards. Although there are few studies on EPC in contrast to ERC (Rong et al., 2019), some studies tackled predicting emotions by focusing on short- and long-term emotion propagation in conversation (Li et al., 2021a), using the commonsense knowledge graph (Li et al., 2021b), and distinguishing between internal emotional changes and external influences on the target user (Rong et al., 2019). In this study, we propose to utilize knowledge-based prediction as a substitution for the subject text and future conversational information that cannot be exploited in actual predictions. Through experiments to confirm the validation of two knowledge resources and the ratio between contextual and knowledge-based prediction, we attempt to predict emotion more accurately by replacing future conversational information with appropriate knowledge-based prediction.

3 EMOTION PREDICTION WITH CONTEXT AND KNOWLEDGE-BASED PREDICTION

3.1 Problem Statement

A tuple (u_1, s_1) is a part of a conversation and constructs sequential dialogue $[(u_1, s_1), (u_2, s_2), ..., (u_N, s_N)]$, which is annotated with sequential emotion labels $[y_1, y_2, ..., y_N]$. u_t is an utterance that a speaker s_t uttered at time t. Conventional ERC studies tackle recognizing emotion e_t of u_t and assigning the correct label similar to y_t given utterances until the t-th turn. On the other hand, our goal is to predict emotion e_t of u_t given utterances until the $t-1$-th turn, $[(u_1, s_1), (u_2, s_2), ..., (u_{t-1}, s_{t-1})]$, without the u_t itself (Figure 1).

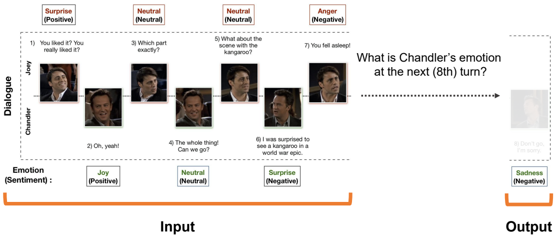

Figure 1: Example of the Emotion Prediction.

3.2 Model Overview

Figure 2 shows an overview of our proposed model, which consists of two main components, the bi-directional context modules (Section 3.3) and the knowledge-based modules (Section 3.4). In previous

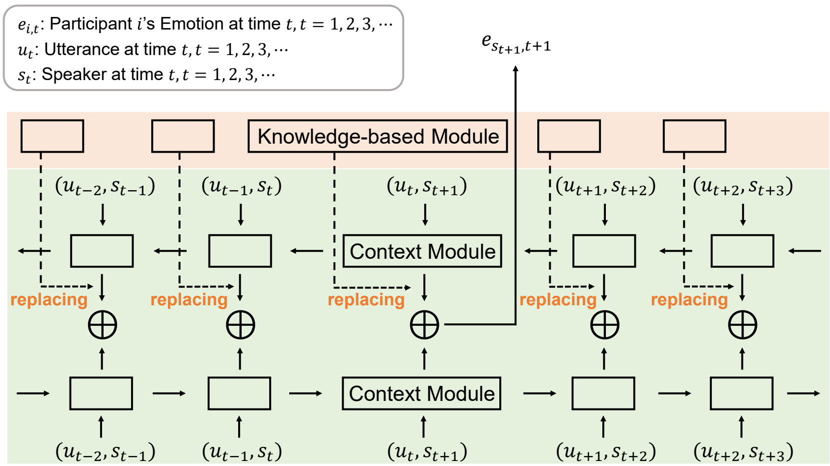

Figure 2: Overview of Proposed Model.

emotion recognition studies, the bi-directional structure is effective for tracking the conversation context (Ghosal et al., 2020; Majumder et al., 2019; Poria et al., 2019a). Although future context information $[u_t, u_{t+1}, \ldots, u_N]$ is considered to recognize emotion e_t in some ERC studies, this future information cannot be known and utilized to predict emotion e_t at the $t-1$-th turn in EPC. In addition, depending only on the conversation context makes it difficult to detect sudden emotional shifts. Therefore, we exploit the knowledge-based emotion prediction as information outside of the conversation. Specifically, we improve emotion prediction performance by using both backward context processing and knowledge-based results, in the learning phase, and completely substituting backward processing with knowledge-based predictions in the actual prediction. In this study, we use external knowledge sources, Transformer trained on a commonsense knowledge graph and a pre-trained Language Model, as the knowledge-based module.

3.3 Context Module

Context modules are connected bi-directionally along the time series and form a context model to predict emotions using information from the entire conversation. This context module consists of three components (Figure 3). First, context state c_t at time t is updated as follows:

$$c_t = GRU(c_{t-1}, x_t) \tag{1}$$

$$x_t = RoBERTa(u_t), \tag{2}$$

where GRU is a gated recurrent unit, RoBERTa is a pre-trained language model, and u_t is an utterance at time t.

Because a participant's emotions are influenced not only by utterances made by others but also by the participant's previous condition, understanding the state of the participant helps track and predict her emotion. We obtain the participant state $p_{i,t+1}$ of participant i at time $t+1$ as follows:

$$\begin{aligned} p_{i,t+1} = \lambda_i^{u_t} \cdot p_{i,t} + (1-\lambda_i^{u_t}) \\ \cdot GRU(p_{i,t}, (x_t \oplus Attention(x_t, [c_1, c_2, \ldots, c_t]))) \end{aligned} \tag{3}$$

$$\begin{aligned} Attention(x_t, [c_1, c_2, \ldots, c_t]) = \\ softmax(x_t^T \otimes [c_1, c_2, \ldots, c_t]) \otimes [c_1, c_2, \ldots, c_t]^T \end{aligned} \tag{4}$$

$$\lambda_i^{u_t} = \begin{cases} 0, \text{participant } i \text{ is a speaker of } u_t \\ 1, \text{participant } i \text{ is not a speaker of } u_t \end{cases}, \tag{5}$$

where $[c_1, c_2, \ldots, c_t]$ is the entire conversation history by time t.

Emotions in conversation are influenced by other participants' emotions. For example, when the previous speaker said something in anger, the emotion affects the next speaker, even if that anger is not directed at her. Thus, to account for this impact of the previous speaker's emotion, emotion state $e_{i,t+1}^{context}$ of participant i at time $t+1$ is defined as:

$$e_{i,t+1}^{context} = GRU(e_{s_t,t}^{context}, p_{i,t+1}). \tag{6}$$

3.4 Knowledge-Based Module

3.4.1 Commonsense Module

In this study, we utilize commonsense knowledge for knowledge-based prediction. As shown in Section 2, commonsense knowledge is widely used for ERC studies as external knowledge and contributes to the improvement of emotion analysis. The commonsense module is shown in Figure 4 as one of the knowledge-based modules. Com-

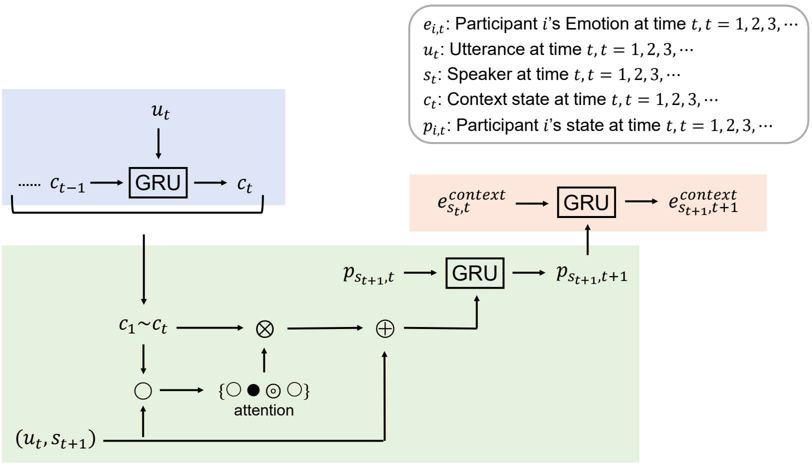
$e_{i,t}$: Participant i's Emotion at time $t, t = 1,2,3,\cdots$
u_t: Utterance at time $t, t = 1,2,3,\cdots$
s_t: Speaker at time $t, t = 1,2,3,\cdots$
c_t: Context state at time $t, t = 1,2,3,\cdots$
$p_{i,t}$: Participant i's state at time $t, t = 1,2,3,\cdots$
u_t
c_{t-1}
GRU
c_t
$e^{context}_{s_t,t}$
GRU
$e^{context}_{s_{t+1},t+1}$
$p_{s_{t+1},t}$
GRU
$p_{s_{t+1},t+1}$
$c_1 \sim c_t$
attention
(u_t, s_{t+1})

Figure 3: Overview of Context Module.

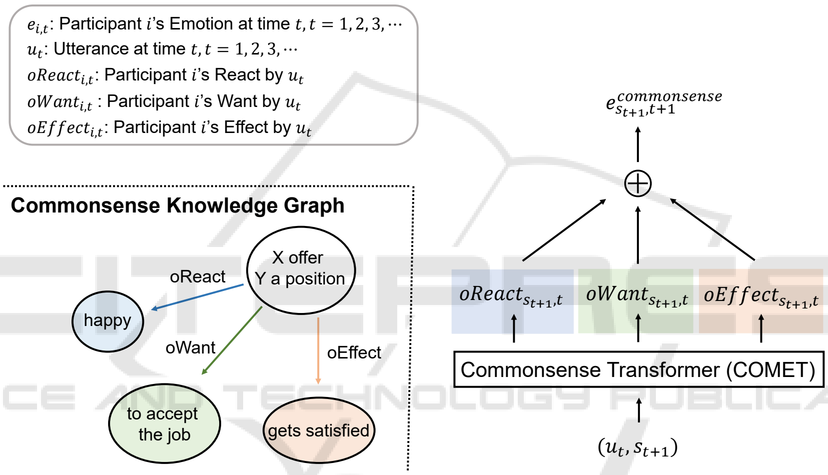
$e_{i,t}$: Participant i's Emotion at time $t, t = 1,2,3,\cdots$
u_t: Utterance at time $t, t = 1,2,3,\cdots$
$oReact_{i,t}$: Participant i's React by u_t
$oWant_{i,t}$: Participant i's Want by u_t
$oEffect_{i,t}$: Participant i's Effect by u_t
Commonsense Knowledge Graph
X offer Y a position
oReact
happy
oWant
oEffect
to accept the job
gets satisfied
$e^{commonsense}_{s_{t+1},t+1}$
$oReact_{s_{t+1},t}$
$oWant_{s_{t+1},t}$
$oEffect_{s_{t+1},t}$
Commonsense Transformer (COMET)
(u_t, s_{t+1})

Figure 4: Overview of Commonsense Module and Commonsense Knowledge Graph.

monsense Transformer (COMET) (Bosselut et al., 2019) is a commonsense generative model trained on ATOMIC (Sap et al., 2019), which is a knowledge graph organized as typed, if-then relations with variables $\{subject, relation, object\}$. COMET generates an $object$ from the $subject$ and the $relation$ (e.g., s = "take a nap," r = Causes, o = "have energy") as follows:

$$object = COMET(subject, relation). \tag{7}$$

In the commonsense module, emotion state $e^{knowledge}_{i,t+1}$ of participant i at time $t+1$ is decided using parameters $oReact$, $oWant$, and $oEffect$, which denote others' (non-speakers') parameters obtained from COMET, as follows:

$$e^{knowledge}_{i,t+1} = oReact_{i,t} \oplus oWant_{i,t} \oplus oEffect_{i,t} \tag{8}$$

$$\begin{aligned} oReact_{i,t}, oWant_{i,t}, oEffect_{i,t} &= \\ COMET(x_t, i, oReact, oWant, oEffect)' \end{aligned} \tag{9}$$

where $oReact$, $oWant$, and $oEffect$ are $relation$ parameters of COMET.

3.4.2 Pre-Trained LM Module

As another knowledge-based module, we use a pre-trained Language Model (LM) following previous ERC studies. The pre-trained LM, RoBERTa-large (Liu et al., 2019), is fine-tuned for emotion label prediction from transcripts of utterances and speakers. RoBERTa-large is an unsupervised pre-trained model with large-scale corpora of unlabeled text. It has 24 layers, 16 self attention heads in each block (24 blocks), a hidden dimension of 1024, and 355 M parameters. For RoBERTa-large fine-tuning, we set a batch size of 8, learning rate of 2e-6, weight decay of 0.1, training epoch of 3, and AdamW as the optimizer.

In the pre-trained LM module, emotion state $e^{knowledge}_{i,t+1}$ of participant i at time $t+1$ is decided as follows:

$$e^{knowledge}_{i,t+1} = RoBERTa(s_{t-1}, u_{t-1}, s_t, u_t, Q_{t+1}) \quad (10)$$

$$Q_{t+1} = \text{"How does } s_{t+1} \text{ feel after this?"} \quad (11)$$

3.5 Replacing Context with Knowledge-Based Predictions

We obtained the predicted emotion label $\hat{y}_{t+1}$ at time $t+1$ as follows:

$$\hat{y}_{t+1} = Attention(e_{i,t+1}, [e_{s_1,1}, e_{s_2,2}, ..., e_{i,t+1}]) \quad (12)$$

$$e_{i,t+1} = e^{context}_{i,t+1} \oplus e^{replaced}_{i,t+1} \quad (13)$$

$$e^{replaced}_{i,t+1} = \alpha \cdot e^{knowledge}_{i,t+1} + (1-\alpha) \cdot e^{backward}_{i,t+1} \quad (14)$$

$$\begin{aligned} &Attention(e_{i,t+1}, [e_{s_1,1}, e_{s_2,2}, ..., e_{i,t+1}]) = \\ &softmax(e^T_{i,t+1} \otimes [e_{s_1,1}, e_{s_2,2}, ..., e_{i,t+1}]) \quad , \quad (15) \\ &\otimes [e_{s_1,1}, e_{s_2,2}, ..., e_{i,t+1}]^T \end{aligned}$$

where $e^{backward}_{i,t+1}$ is an output of backward sequential context modules and, α takes 0 or 1 according to the replacement rate γ between 0.0 to 1.0

To determine the replacement rate γ, we conducted the preliminary experiments (Section 4.4.1)

4 EXPERIMENT

4.1 Dataset

We evaluated the performance of our proposed method with the Multimodal EmotionLines Dataset (MELD) (Poria et al., 2019a), a conversation dataset used in ERC. Conversations of this dataset were collected from the TV show *Friends*, and each utterance was labeled into seven emotion categories, based on Ekman's six basic emotions: surprise, fear, sadness, joy, disgust, anger, and neutral. As shown in Table 1, emotion labels are unbalanced like in actual conversations, e.g., neutral for over 46% in this dataset.

4.2 Comparison Methods

We evaluated the performance of our proposed method in comparison experiments with the following existing methods.

DialogueRNN (Majumder et al., 2019) recognizes the speaker's emotion by considering the conversation context of each participant using a Gated Recurrent Unit (GRU) network to keep track of individual speaker and listener states. This model is used as a baseline for ERC.

Table 1: Description of the dataset MELD. The dataset is modified to fit the emotion prediction task, and numbers in parentheses represent the amount of data contained in the original dataset.

	Train	Dev	Test
Neutral	4168(4710)	416(470)	1117(1256)
Surprise	1084(1205)	138(150)	253(281)
Fear	240(268)	37(40)	44(50)
Sadness	619(683)	104(111)	187(208)
Joy	1539(1743)	141(163)	358(402)
Disgust	247(271)	22(22)	63(68)
Anger	1024(1109)	137(153)	308(345)
Utterances	8951(9989)	995(1109)	2330(2610)
Dialogues	968(1039)	108(114)	268(280)

COSMIC (Ghosal et al., 2020) recognizes a speaker's emotions using the conversation context and commonsense knowledge graphs. Commonsense knowledge is used to track some speaker states, such as the internal, external, and intent state, and improves the emotion detection performance.

CoMPM (Lee and Lee, 2022) classifies a speaker's emotion considering the conversation context by combining the Transformer encoder and pre-trained LM. Using pre-trained knowledge makes this approach easy to extend to non-English languages.

TodKat (Zhu et al., 2021) classifies a speaker's emotion by topic detection of conversations using a fine-tuned pre-trained LM and knowledge awareness using a commonsense knowledge graph.

EmotionFlow (Song et al., 2022) recognizes a speaker's emotion considering the emotion shift of speakers in each turn within the conversation based on Transformer encoder and CRF layers.

DialogInfer-(S+G) (Li et al., 2021a) predicts a speaker's emotion by considering emotion propagation in conversation. Two types of emotion propagation are distinguished: long-term and short-term propagation. A sequence-based model captures the former and a graph-based model captures the latter.

DialogInfer-(S+G)+K (Li et al., 2021b) predicts a speaker's emotion by a hybrid emotion prediction model that combines DialogInfer-(S+G) and external knowledge as, a commonsense knowledge graph.

4.3 Training Setup

We used a batch size of 8, learning rate of 0.001, and dropout rate of 0.1 to train our models. We used weighted cross-entropy as the optimizer, and the Adam optimization algorithm. Our models were trained for 60 epochs and other hyperparameters were optimized using the grid search. All hyperparameters of all comparison methods were set to the same as reported in each paper. The input features for all models were unified into 1024-dimensional input using RoBERTa-large.

4.4 Results and Discussion

4.4.1 Preliminary Experiments

Preliminary experiments were conducted on the replacement rate γ that determines the weight between the backward sequential context model and the knowledge-based model, changing the value from 0.0 to 1.0 by 0.1 (Table 2). The replacement rate γ only works during the training phase, and it is set to 1.0 during the test phase. For instance, in a bi-Context model, the backward direction outputs are replaced by 0.0 according to γ in the training phase, and all backward outputs are replaced by 0.0 in the test phase. In the bi-Context+Commonsense and bi-Context+pre-trained LM, backward direction outputs are replaced by outputs of commonsense model and pre-trained LM respectively according to γ in the training phase. All backward outputs are replaced by outputs of the commonsense model and pre-trained LM appropriately in the test phase. All models show the best prediction scores at high values of the replacement rate γ. It confirms that using contextually informative backward directional outputs during the training phase does not improve the accuracy of emotion predictions, where the data cannot be used in the test phase. Though bi-Context and bi-Context+Commonsense show the best results when the replacement rate γ is 0.7, bi-Context+pre-trained LM shows the better score when the replacement rate γ is 0.9. Since it is verified that the pre-trained LM tends to be biased in emotion recognition (Mao et al., 2022), we conclude that bi-Context+pre-trained LM shows the highest prediction result at the high replacement rate, γ is 0.9, because it allows for strong consideration for biases during the training phase.

4.4.2 Comparing with Other Methods

Table 3 shows the performance of the comparison methods and our models. Among the comparison methods, CoMPM shows the best results in weighted-F1 and anger, and the methods show the best score differs by each emotion category. In contrast, our bi-Context+pre-trained LM method outperforms the existing methods in all emotion categories except neutral in EmotionFlow. It surpasses the highest comparison methods' score by 15.04%, 12.71%, 15.58%, 8.05%, and 9.23% in surprise, sadness, joy, anger, and weighted-F1, respectively. And while scores are still not high, fear and disgust can be predicted at about twice the rate of existing methods. In neutral, on the one hand, bi-Context+pre-trained LM is lower than EmotionFlow. On the other hand, the proposed method surpasses EmotionFlow by about 10% in all other categories. Thus, we think that this result is because our model can correctly predict each emotional category for utterances that existing methods incorrectly classify as neutral.

4.4.3 Ablation Study

The lower half of Table 3 presents the results of the comparison within the proposed method. As shown in the results for mono-Context and bi-Context, considering contextual information bi-directionally contributes to emotion prediction for all categories. As reported in a previous study (Mao et al., 2022), it can be confirmed that the prediction by the pre-trained LM alone is biased. In fine-tuning, more biased results were obtained by changing hyper-parameters such as the number of epochs and learning rate. On the other hand, the model combined with bi-Context improves prediction accuracy for all emotion labels except for surprise. In fear, the prediction score is doubled compared to the case where only conversational contexts are utilized. This result indicates that pre-trained LM can contribute to emotion prediction by simultaneously using contextual information. However, despite the result that independently utilizing commonsense knowledge contributes to performance improvements (Li et al., 2021b), few performance improvements are observed in bi-Context+Commonsense compared with bi-Context. We therefore conclude that, because the replacement of contextual information by commonsense does not include any conversation history, bi-Context+Commonsense shows few improvements from a comparison of bi-Context+Commonsense and bi-Context+pre-trained LM.

5 CONCLUSION

In this study, we tackled predicting emotion before a speaker expresses their emotion and proposed utilizing knowledge-based predictions as a replacement for contextual information that cannot be used in actual predictions appropriately. The model bi-Context+pre-trained LM, replacing context information with pre-trained LM, showed significant performance in the experiment, and we confirmed the effectiveness of knowledge-based results in predicting emotion. According to a comparison with the case using the commonsense knowledge graph, we verified that the knowledge-based predictions followed by conversa-

Table 2: Results of the preliminary experiments. Weighted-F1 is used as an evaluation metric, and all models are trained 5 times for each experiment. The median of the scores is evaluated (same in other tables). Bold text indicates the best performance in each part.

Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
bi-Context	30.05	44.10	45.39	44.60	45.88	46.87	46.32	**46.94**	45.10	45.43	34.07
bi-Context+Commonsense	19.96	33.27	38.76	39.37	41.15	39.94	41.41	**43.37**	42.02	41.38	42.69
bi-Context+pre-trained LM	23.31	38.26	43.99	43.55	45.97	45.99	47.41	46.63	46.71	**48.21**	43.64

Table 3: Comparison results of our methods with various previous methods and within our proposed methods' components on MELD. The first half is the performance of the previous methods and the second half is the performance of the proposed method and its components. Bold text indicates the best performance in each emotion category.

Method	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Weighted-F1
DialogueRNN	55.39	16.22	03.77	01.84	28.57	00.00	21.37	35.75
COSMIC	60.64	07.31	00.00	06.96	15.16	00.00	29.01	36.58
CoMPM	57.80	13.26	02.67	14.13	28.78	05.26	30.85	38.98
TodKat	57.40	10.55	03.45	15.86	29.43	04.82	30.60	38.70
EmotionFlow	**64.35**	07.87	00.00	02.94	18.99	02.99	29.42	38.78
DialogInfer-(S+G)	61.71	07.54	00.00	07.29	15.41	05.48	24.66	36.77
DialogInfer-(S+G)+K	61.08	12.99	00.00	07.33	16.29	04.76	23.39	37.04
mono-Context	55.36	11.93	00.00	02.56	20.13	00.00	21.49	34.07
bi-Context	61.91	**31.58**	03.03	23.70	43.29	10.20	37.40	46.94
Commonsense	37.06	15.35	02.02	16.79	21.33	05.13	21.14	27.03
pre-trained LM	44.85	18.09	00.00	10.04	22.74	00.00	31.04	31.87
bi-Context+Commonsense	59.42	21.05	03.70	24.56	34.68	06.45	38.25	43.37
bi-Context+pre-trained LM	62.61	31.26	**07.02**	**28.57**	**45.01**	**11.76**	**38.90**	**48.21**

tional context are more valid as a replacement for contextual information. In future studies, it should be pursued that real-time emotion prediction using chat tools' and other real-world applications' conversation histories that are not limited to data sets.

ACKNOWLEDGMENTS

This work was supported by JST CREST Grant Number JPMJCR20D1, Japan.

REFERENCES

Acheampong, F. A., Nunoo-Mensah, H., and Chen, W. (2021). Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, pages 1–41.

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.

Cambria, E., Livingstone, A., and Hussain, A. (2012). The hourglass of emotions. In *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*, pages 144–157. Springer.

Chen, S. Y., Hsu, C. C., Kuo, C. C., Huang, T. H. K., and Ku, L. W. (2019). Emotionlines: An emotion corpus of multi-party conversations. In *11th International Conference on Language Resources and Evaluation, LREC 2018*, pages 1597–1601. European Language Resources Association (ELRA).

Edmondson, A. C. (2018). *The fearless organization: Creating psychological safety in the workplace for learning, innovation, and growth.* John Wiley & Sons.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Gao, Q., Cao, B., Guan, X., Gu, T., Bao, X., Wu, J., Liu, B., and Cao, J. (2022). Emotion recognition in conversations with emotion shift detection based on multi-task learning. *Knowledge-Based Systems*, 248:108861.

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

Khanpour, H. and Caragea, C. (2018). Fine-grained emotion detection in health-related online posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166.

Kumar, A. and Garg, G. (2020). Systematic literature review on context-based sentiment analysis in social multimedia. *Multimedia tools and Applications*, 79:15349–15380.

Lee, J. and Lee, W. (2022). Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter* *of the Association for Computational Linguistics: Human Language Technologies*, pages 5669–5679.

Li, D., Zhu, X., Li, Y., Wang, S., Li, D., Liao, J., and Zheng, J. (2021a). Emotion inference in multi-turn conversations with addressee-aware module and ensemble strategy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3935–3941.

Li, D., Zhu, X., Li, Y., Wang, S., Li, D., Liao, J., and Zheng, J. (2021b). Enhancing emotion inference in conversations with commonsense knowledge. *Knowledge-Based Systems*, 232:107449.

Li, S., Yan, H., and Qiu, X. (2022). Contrast and generation make bart a good dialogue emotion recognizer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11002–11010.

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. *arXiv preprint arXiv:1907.11692*.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Mao, R., Liu, Q., He, K., Li, W., and Cambria, E. (2022). The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

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

Poria, S., Majumder, N., Mihalcea, R., and Hovy, E. (2019b). Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Rong, H., Ma, T., Cao, J., Tian, Y., Al-Dhelaan, A., and Al-Rodhaan, M. (2019). Deep rolling: A novel emotion prediction model for a multi-participant communication context. *Information Sciences*, 488:158–180.

Russell, J. A. and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805.

Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., and Choi, Y. (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.

Saxena, P., Huang, Y. J., and Kurohashi, S. (2022). Static and dynamic speaker modeling based on graph neural network for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 247–253.

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. IEEE.

Tu, G., Liang, B., Jiang, D., and Xu, R. (2022). Sentiment-emotion-and context-guided knowledge selection framework for emotion recognition in conversations. *IEEE Transactions on Affective Computing*.

Wang, Y., Zhang, J., Ma, J., Wang, S., and Xiao, J. (2020). Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue*, pages 186–195.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., and Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688.

Wu, L., Sun, P., Fu, Y., Hong, R., Wang, X., and Wang, M. (2019). A neural influence diffusion model for social recommendation. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 235–244.

Yi, Y. and Wang, H. (2019). Multi-modal learning for affective content analysis in movies. *Multimedia Tools and Applications*, 78:13331–13350.

Zad, S., Heidari, M., James Jr, H., and Uzuner, O. (2021). Emotion detection of textual data: An interdisciplinary survey. In *2021 IEEE World AI IoT Congress (AIIoT)*, pages 0255–0261. IEEE.

Zhang, W., Li, X., Deng, Y., Bing, L., and Lam, W. (2022). A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.

Zhong, P., Wang, D., and Miao, C. (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.

Zhu, L., Pergola, G., Gui, L., Zhou, D., and He, Y. (2021). Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582.

Zucco, C., Calabrese, B., Agapito, G., Guzzi, P. H., and Cannataro, M. (2020). Sentiment analysis for mining texts and social networks data: Methods and tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(1):e1333.