IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 October 2022, accepted 6 November 2022, date of publication 11 November 2022, date of current version 16 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3221408

RESEARCH ARTICLE

Korean Drama Scene Transcript Dataset for Emotion Recognition in Conversations

SUDARSHAN PANT[1], EUNCHAE LIM[1], HYUNG-JEONG YANG[1], GUEE-SANG LEE[1], SOO-HYUNG KIM[1], YOUNG-SHIN KANG[2], AND HYERIM JANG[2]

[1]Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea
[2]Department of Psychology, Chonnam National University, Gwangju 61186, South Korea

Corresponding author: Hyung-Jeong Yang (hjyang@jnu.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) Grant through the Korean Government Ministry of Science and ICT (MSIT) under Grant NRF-2020R1A4A1019191.

ABSTRACT Understanding emotions in conversation is a challenging task, as the sentences often have an implied meaning that is not generally understood in isolation. Efficient use of contextual information is essential for emotion recognition in conversations. Many published datasets provide contextual information for situations such as text-based online messaging, chatbots, and movie dialogues. However, such dialogue-based datasets are collected by selecting ideal conversational situations and thus do not include many variations in dialogue length and number of participants. Therefore, such datasets may not be applicable for emotion recognition in text-based movie transcripts, where scenes contain variations in the number of speakers and length of spoken sentences. We present a conversation dataset based on the Korean television show transcripts to analyze the emotions in presence of scene context. The Korean Drama Scene Transcript dataset for Emotion Recognition (KD-EmoR) is a text-based conversation dataset. We analyze three classes of complex emotions: euphoria, dysphoria, and neutral, in the scenes of a television drama to build a publicly available dataset for further research. We developed a context-aware deep learning model to classify emotions using the speaker-level context and scene context and achieved an F1-score of 0.63 on the proposed dataset.

INDEX TERMS Emotion analysis, emotion recognition in conversation, scene transcripts.

I. INTRODUCTION

Machines fail to understand the implied meaning in human conversation because human dialogue often contains meaning that may not be understood in isolation. Therefore, contextual information is necessary for emotion recognition in conversation (ERC). Although ERC has been extensively researched using English conversations, very little research has been conducted on emotion recognition in Korean conversations. Moreover, the scarcity of labeled emotion datasets is one of the significant challenges in developing robust ERC models for Korean language data [1]. Several datasets [2], [3], [4] have been constructed using English language conversations; however, the ERC in the Korean language has not been sufficiently studied.

The ERC research deals with emotion recognition in conversational scenarios, such as instant messaging, chatbots, and movies. Despite having different types of dialogues, the conversation scripts of movies or TV dramas have not been explored to their full potential for emotion recognition. Unlike in ideal conversation scenarios, recognizing emotions in movie scripts require the analysis of various components such as scene context, inter-speaker dynamics, and contextual meaning of the spoken sentences.

Analyzing emotions using a conversation dataset can be beneficial for generating emotionally appropriate responses in chatbots [5], [6]. Other applications of ERC include analyzing movie dialogues, instant messaging, tweets [7], [8], [9], and news headlines [10], [11]. Emotion recognition in

The associate editor coordinating the review of this manuscript and approving it for publication was Renato Ferrero.

scenes is different from that in smaller texts, such as tweets or social media posts as they are context-independent, (i.e., the emotion in such smaller texts is solely related to the words used) whereas the conversational texts the conversation history and the turns of the speakers affect the emotions of other speakers for conversational texts [12].

Traditionally, ERC is studied based on categorical [13] or dimensional views of emotions [14]. Categorical emotions include different classes of discretized emotions, hierarchically arranged in wheel [15], or tree-shaped [16] shaped models. Dimensional emotions refer to the numeric representation in a fixed scale in particular emotional spaces, such as arousal, valence, and dominance. More complex emotional states, such as euphoria and dysphoria, have not gained much attention outside clinical trials for mental disorders [17].

Although positive, negative, and neutral sentiments have been widely studied, a few studies have highlighted the application of emotions related to behavioral states such as euphoria and dysphoria. Starcevic conceptualized dysphoria as a complex emotional state of dissatisfaction and irritation and reviewed the relationship between emotional states and psychiatric disorders associated with dysphoria [18]. Zillman et. al. [19] studied the emotions of the movie protagonists in terms of euphoria and dysphoria through feedback from the children watching the movies. Katsis et al [31] studied the emotional states of car racing drivers as a classification of euphoria and dysphoria representing positive and negative emotional states, respectively. The authors interpreted the valence as euphoric and dysphoric emotions to represent the behavioral states. Studying these emotions in real-life situations can facilitate the early diagnosis of morbid emotional conditions. As emotions, such as euphoria and dysphoria, describe the general behavioral states of the characters and the overall scene, they represent the emotional states more accurately. Therefore, we labeled the emotions based on the presence of mild euphoric and dysphoric emotional states in the conversations.

Several conversation datasets have been developed considering the contextual information in TV shows [2], [3]. However, traditional conversation datasets are labeled based on dialogue representing ideal conversation situations with dyadic or multiparty situations. Additionally, preparing input data for such an ideal multiparty situation may involve the loss of several crucial components in the movie scripts such as single-speaker scenes, monologue, and narrative scene descriptions. The exclusion of such information may not fully represent the conversation in scene transcripts. Contextual information can be obtained from previously spoken sentences in text-based conversations, such as movie transcripts. Therefore, constructing the dataset using selected dialogues may only partially represent the conversations partially. Emotion recognition in scene transcripts is still challenging due to the lack of datasets designed to represent real-world conversations. Therefore, we propose expanding the notion of conversation to include realistic conversation scripts in the domain of ERC.

In this work, we propose the conversation dataset for the Korean Drama Scene Transcript dataset for Emotion Recognition (KD-EmoR), which includes realistic scene transcripts from a Korean drama, where scenes with a varying number of sentences are labeled with three sentiment labels representing euphoric, dysphoric, or neutral expressions in the drama. This paper makes the following contributions:

- A large Korean language ERC dataset, based on the scene transcript of a TV show, is presented.
- Annotation is performed for sentences instead of utterances to make emotion recognition more relevant to text-based conversations. Utterances may be more relevant in emotion recognition using speech modality.
- The dataset is labeled with complex emotions such as euphoria, dysphoria and neutral, to determine the socio-behavioral emotional states in psychology.
- We present a classification model named cross-attention-based fusion of speaker and scene context (XAF-SS) as a baseline method to classify the sentiments into euphoric, dysphoric, and neutral.

The remainder of the paper is organized as follows. Section II discusses the prior work in the field of ERC related to this work. Next, Section III describes the building process, annotation methodology, and statistics of the proposed dataset. Section IV details the proposed baseline method for emotion recognition in scenes. Finally, Section V summarizes the baseline experiment results, and Section VI concludes the paper.

II. RELATED WORK

Several datasets for the study of emotion recognition have been developed in the past. For instance, the Emotion-Lines [2] dataset consists of 2772 utterances from 1000 dialogue examples from a TV show called *Friends*. The dialogue consists of utterances spoken by multiple speakers. Each utterance is annotated with six categorical emotions (joy, anger, sadness, fear, surprise, and disgust) and binary labels indicating neutral and non-neutral emotions. Five annotators did the annotation, and the final annotation was determined by majority voting.

The MELD [3] dataset is a multimodal extension of the Emotion-Lines dataset, where video clips were included, and the dataset was re-annotated by three annotators. The dataset contains 13,708 utterances from 1433 pieces of dialogue with audio, visual, and textual modalities. The utterances in the dialogue are annotated with seven emotion labels (anger, fear, disgust, surprise, joy, sadness, and neutral) and positive, neutral, and negative sentiments.

The IEMOCAP [4] dataset consists of the video dataset with scripted and spontaneous conversations with 10,039 turns. The emotion labels include happiness, neutral, anger, disgust, fear, sadness, frustration, excitement, and surprise. The EmoryNLP [20] is a conversation dataset with 12,606 utterances from 897 scenes annotated with seven emotions: sad, scared, mad, joyful, peaceful, powerful, and neutral.

TABLE 1. Summary of existing similar datasets for emotion recognition in conversations.

Dataset	Size	Annotations	Conversation	No. of annotators	Inter-rater agreement (Fleiss Kappa)
Ara-SenTi-Tweet [21]	17,573 tweets	positive, negative neutral, and mixed	No	3	0.60
Gold Standard [23]	8,868 tweets	positive, neutral, negative, and mixed classes.	No	2	0.82
ASAD [22]	95,000 tweets	positive, negative, and neutral	No	3	0.56
MELD [3]	13,708 utterances	positive, negative, and neutral anger, fear, disgust, surprise, joy, sadness, and neutral	Yes	5	0.34
Emotion-Lines [2]	2,772 utterances	anger, fear, surprise, sadness, disgust, and joy	Yes	3	-
IEMOCAP [4]	10,039 utterances	anger, disgust, frustration, happiness, sadness, surprise, fear, excitement, and neutral	Yes	-	-
Proposed Dataset	12,289 sentences	euphoria, dysphoria, neutral	Yes	64	0.27

The AraSenTi-Tweet [21] is an Arabic language dataset with 17, 573 tweets each labeled with one of four emotions: positive, negative, mixed, and neutral. Three Arabic native speakers annotated the tweets. Similarly, ASAD [22] is an Arabic dataset for sentiment analysis of tweets. This dataset consists of 95,000 tweets with positive, negative, and neutral labels annotated by three annotators. The summary of the existing datasets for ERC is presented in Table 1.

The above-mentioned datasets, constructed to include ideal conversation scenarios through the selection of dialogues, either focus on multiparty or dyadic conversations. Real-world conversations, the conversations are highly dynamic, with varying sentence lengths. Although conversations are generally broken down into utterances in emotion recognition datasets, the utterances may not be well defined when only the textual modality is present. The utterance-based conversation datasets [3], [4], [20] require an audio modality to annotate based on the utterances because the utterance is the unit of spoken sentences. The model trained on such an utterance-based dataset may not be efficient for tasks related to a textual conversation. Therefore, we adopt the turns in a conversation as a unit for annotation.

Recently, fusion techniques have been investigated in several studies. An attention mechanism is helps infer latent cross-modal relationships between various modalities. An attention network was introduced by [32] to align input and output sequences in natural language processing tasks. The attention mechanism has been applied in diverse areas, including computer vision [33]. In emotion recognition, cross-attention is applied to fuse various modalities. Xu et al. [34] used the attention mechanism to learn the alignment of audio and text modality for multimodal emotion recognition. Praveen et al. [35] improved the performance of the emotion recognition model using a cross-attention module to fuse audio-visual features. Shriwardhana et al. [40] demonstrated the efficiency of fusion speech and text modalities on emotion recognition tasks with a self-supervised learning model. Attention mechanisms have been found effective for multimodal fusion emotion recognition [41], [42]. We adopt cross-attention between different textual components in the conversation to learn the contextual information.

III. PROPOSED DATASET

The dataset was constructed using the script of the TV show *Three Brothers*. It was selected based on the availability of the complete script, national audience ratings, and its diverse content. It is a popular Korean family drama with a viewer rate of 30%-40% per episode, and it consists of conversations with emotions related to different age groups, thus representing the general Korean population.

The dataset consists of 1,513 scenes with a varying number of speakers and sentences. Scenes in a TV show transcript refer to conversations in a particular situation. Scenes generally consist of one or more individuals having a conversation. Scenes usually include the narrative scene description which provides additional information about the scene. Conversations include several incomplete sentences and short sentences with vague meanings. Although such short sentences may not represent any emotion in isolation, they play an essential role in understanding the overall emotional state of the characters. Identifying such implied sentiment is challenging requiring efficient context-aware emotion recognition models. Therefore, we propose a dataset with scene transcripts with realistic conversation data for context-aware emotion recognition research.

The corpus was preprocessed to exclude inappropriate sentences to ensure the quality of the dataset prior to the annotation experiments. As the corpus is based on a TV show transcript, some conversation turns are represented only through visual cues. In such cases, the transcripts are marked with multiple punctuation marks. Such meaningless turns in the conversations were removed during preprocessing. The dataset was screened by two graduate researchers from the Department of Psychology and the Department of Artificial Intelligence Convergence.

FIGURE 1. The annotation environment setup.

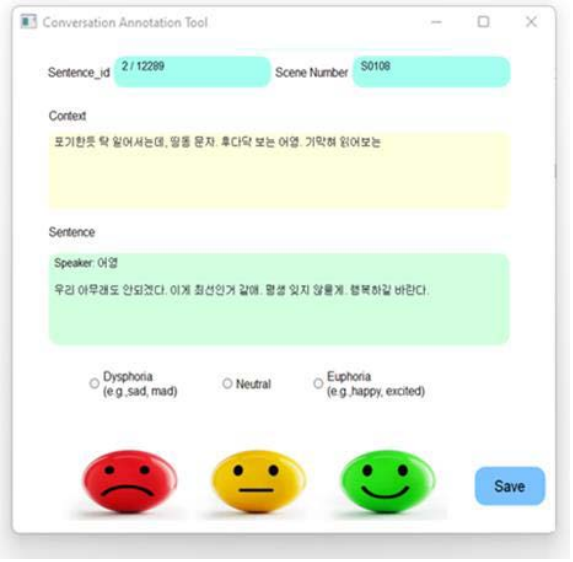

FIGURE 2. The user interface of the annotation tool with the sentence, scene description (context), and emotion label options.

A. ANNOTATION

Sixty-four annotators between 20 and 31 years of age (mean age: 24.31, 36 males and, 28 females) were recruited through Chonnam National University online job portal. Mental and emotional fitness was confirmed through self-declaration of the annotations to prevent faulty annotations due to abnormal judgment of the emotions in the conversations. Candidates with a history of brain injury, psychiatric disorder, or consumption of drugs, which could affect the emotional states, were disqualified for the annotation experiments.

An annotation experiment protocol was designed with the simulated prior simulation experiment under the guidance of psychologists. The dataset was developed to study the emotions defined as the behavioral states of the characters and the scene. To study emotions in context, we adopted euphoria, dysphoria, and neutral as emotion labels based on the terminologies used in psychological studies [43], [44]. In behavioral psychology euphoria and dysphoria have been used to study emotional states for several clinical [19] and non-clinical purposes [31]. Euphoria refers to intense excitement and happiness while dysphoria is a state of unease or generalized dissatisfaction with life, representing emotions with negative valence [14], [45]. In other words, euphoria describes the positive behavior of the characters as observed through an overall understanding of the scene. Similarly, dysphoric emotion is analogous to negative sentiment observed in the presence of contextual information.

The annotators were explained and provided with handouts describing the annotation process and the meaning of the labels. All the annotators confirmed that they understood the emotion labeling process before starting the annotation. The annotators were instructed to sit comfortably in front of a computer monitor to label the sentences. The annotation environment is presented in Fig. 1. We developed an annotation tool using the PyQT5 framework The participants were shown the scene context and the spoken sentence, and the interface for labeling the corresponding emotions. The scenes and sentences were displayed sequentially to the annotators allowing them to understand the contextual meaning of the sentences. Participants selected one of the three emotion labels using the annotation user interface, as illustrated in Fig. 2.

Sixty-four annotators were divided into two groups with 32 members each, and the annotation experiments were conducted for 3 days for each group. The annotators participated in the experiments for 5 hours every day. A 10-minute break every 40 to 50 minutes was provided to avoid fatigue during the experiment. The annotators were paid about 10 USD per hour for their time and effort in emotion annotation.

B. DATASET STATISTICS

The dataset consists of sequential information in the script from the Korean TV show *Three Brothers*. Table 2 presents an example of a scene with the scene description, the speakers, and the spoken sentences with emotion labels. It includes 1513 scenes with a total of 12289 sentences. The data characteristics are shown in Table 3. Each scene in the TV drama is considered a conversation consisting of sentences spoken by multiple speakers.

The reliability of the annotation is an essential factor in assessing the quality of the datasets. The reliability of the datasets is a measurement of the trustworthiness of the annotation instructions and the protocol [21]. The annotation quality of the datasets is assessed using a reliability measure called inter-rater reliability, which quantifies the degree of agreement among the annotators labeling the samples independently [24]. We adopted Fleiss Kappa [25] to evaluate the inter-annotator reliability because there are more than two annotators. A fair agreement [26] was observed among 64 annotators annotating 12,289 sentences for three

TABLE 2. Example annotations from the dataset with scene description, speaker information, and labeled sentences.

ID	Speaker	Sentence	Label
34	과자 (Kwa-Ja)	아들 표창장 받는 동안 아부진 뭐했대? (What was your dad doing when his son was rewarded?)	Dysphoria
35	순경 (Soon Kyung)	뭐? (What?)	Dysphoria
36	과자 (Kwa-Ja)	그냥 넘어가요. 입 심심해서 혼자 지껄여 봤네요 (nothing, I said just by myself.)	Dysphoria
37	현찰 (Hyun-Chal)	아부지, 저 먼저 가봐야겠는데요 (Dad, I must leave now.)	Neutral
38	순경 (Soon-Kyung)	오늘 중요한 일있다 그랬지? 얼른 가봐라., (Ah, you said you have something important … ok go ahead.)	Neutral
39	과자 (Kwa-Ja)	사진이라도 박고 가지. (It would have been better if you could leave after taking a photo.)	Dysphoria

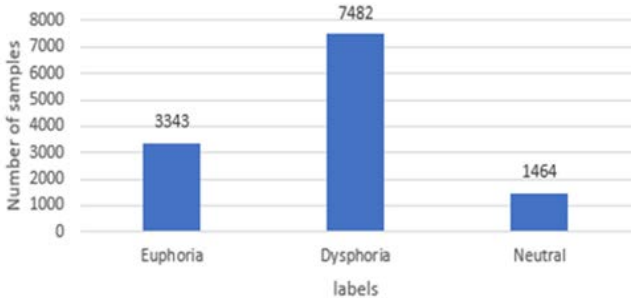

FIGURE 3. Distribution of three labels in the dataset.

TABLE 3. Dataset characteristics.

Property	Value
Total no. of sentences	12289
Length of sentences	2–274 (mean 27.82)
Total no. of speakers	118
Total no. of scenes	1513
Number of sentences in a scene	1–64 (mean: 8.12)

categories, with a Fleiss kappa of 0.27. The inter-rater agreement was comparable to that of the existing dataset despite having a higher number of annotators. The final labels for the sentences were selected based on majority voting on the emotions labeled by the annotators. Most sentences were labeled for dysphoria with 60.9% of total sentences, whereas neutral and euphoria constituted about 11.9% and 27.2%, respectively. The distribution of the labels in the dataset is depicted in Fig. 3.

The dataset will be made available to the participants of the Fourth Korean Emotion Recognition Challenge (KERC) 2022 during the competition duration (from August to October 2022). The dataset will also be publicly available for academic research at https://sites.google.com/view/KD-EmoR. As the proposed dataset includes scene transcripts from the drama with scenes representing various real-world situations, we believe the dataset will be useful for analyzing emotions in different application areas.

IV. EMOTION RECOGNITION IN SCENES

A. PROBLEM STATEMENT

The conversation in movie scripts is composed of several scenes containing N sequential tuples of sentences spoken by K speakers given as $[(s_1, p_1), (s_2, p_2) \ldots..(s_N, p_K)]$ where, s_N is the sentence spoken by a person p_K, and the number of speakers may be less than or equal to the number of sentences (i.e., $K \leq N$). The scene may have a narrative description D which provides its background information. Thus, scene C is represented as $(D, \left[(s_1, p_1), (s_2, p_2) \ldots . (s_N, p_K)\right])$. The goal is to classify s_N into one of the emotional labels considering the description and the sequence.

B. PROPOSED BASELINE MODEL

We developed a context-aware emotion recognition model for the classification of emotion labels. The overall architecture of the proposed classification model is presented in Fig. 4. The proposed model consists of a scene-level context module, a speaker-level context module, and context fusion. In a conversation, the emotion of a sentence is affected by the previously spoken sentences. Therefore, the scene-level context module is designed to capture the required information from the existing sentences during a scene. The speaker-level context captures the information from the past, distinguishing between the speakers from the speaker of the target sentence. The Bidirectional Encoder Representations from Transformers (BERT) [27] models have been proven to be effective in various tasks in natural language processing. For preprocessing we utilize the multilingual BERT model pre-trained on the Wikipedia corpus. The proposed baseline model consists of a speaker-level context module, scene-level context module, and attention-based fusion module.

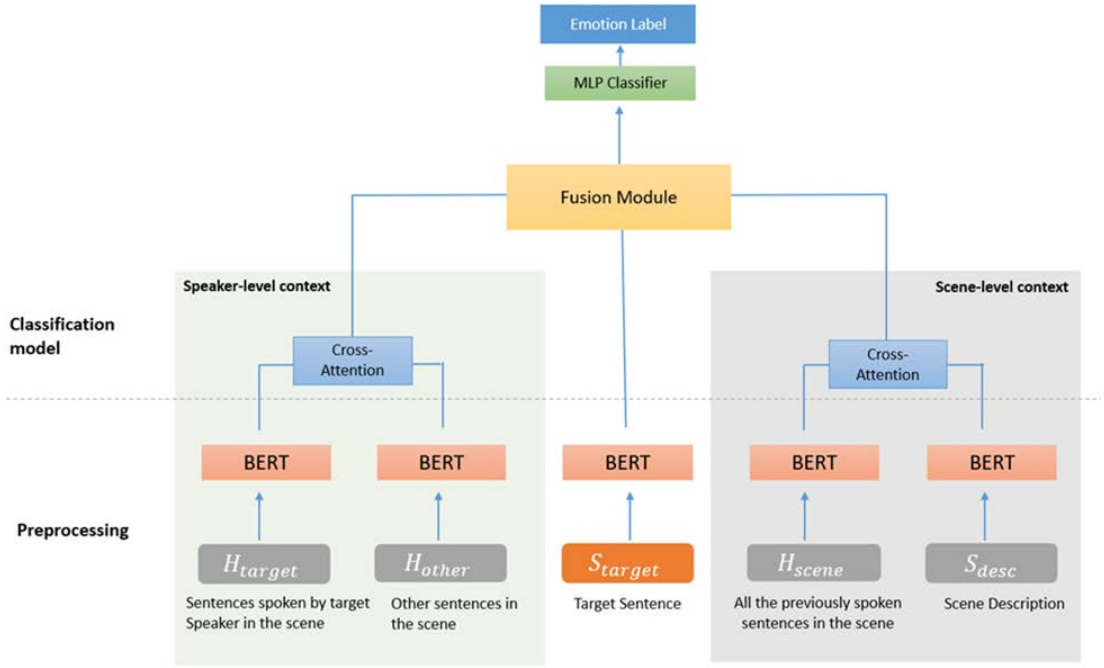

FIGURE 4. The architecture of the proposed model.

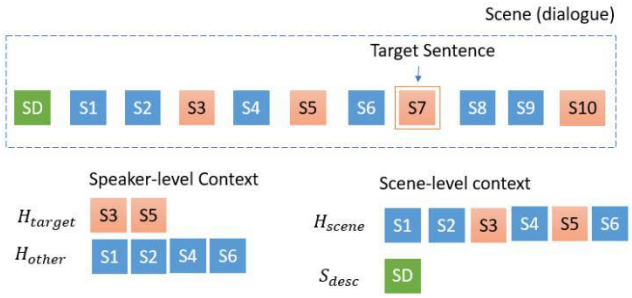

FIGURE 5. Distribution of three labels in the dataset.

1) SPEAKER_LEVEL CONTEXT MODULE

Speaker-level context refers to the information related to the speakers in the conversation. As the scene consists of sentences spoken by multiple speakers as a series of (S_N, P_K), the information flow among the speakers determines the emotional state of the sentences. To distinguish between the speakers' emotions, we consider the previously spoken sentences in the scene as contextual information. As shown in Fig. 5., we group the previous sentences into two groups: H_{target}, a set of sentences spoken by the same speaker and H_{other}, a set of sentences spoken by other speakers. The BERT [27] pre-trained on a multi-lingual corpus was used to extract features for two groups of sentences, which were concatenated to obtain the speaker-level features. For a target sentence (s_t, p_i), the speaker-level context $C_{speaker}$ is obtained as follows:

$$C_{target} = \text{BERT}(\left[(s_1, p_i), (s_2, p_i) \ldots . (s_{t-1}, p_i)\,;\right]) \quad (1)$$

$$C_{other} = \text{BERT}(\left[(s_1, p_1), (s_2, p_2) \ldots . (s_{t-1}, p_M)\right]) \quad (2)$$

$$C_{speaker} = X_{att}(C_{target}, C_{other}) \quad (3)$$

where C_{target} and C_{other} represent the features extracted from the sentences spoken before the target sentence by p_i, and the speakers other than p_i respectively. Moreover, X_{att} denotes a cross-attention operation.

2) SCENE_LEVEL CONTEXT MODULE

We used the narrative description of the scene and the historical conversational information in the scene to represent the scene-level context. The scene transcripts from the drama often include the scene information or the narrative description of the scene. Such information can be crucial to understanding the scene. As illustrated in Fig. 5., the scene context C_{scene} is represented by the fusion of the narrative scene description S_{desc} and all sentences H_{scene} spoken before the target sentence s7. All the previously spoken sentences in the scene are concatenated to form a scene history sequence H_{scene}. The features from S_{desc} and H_{scene} are extracted using a pre-trained BERT model. For a target sentence (s_t, p_i), the scene-level context C_{scene} is obtained as follows:

$$H_{scene} = \left[(s_1, p_1), (s_2, p_2) \ldots . (s_{t-1}, p_M)\right] \quad (4)$$

$$C_{scene} = X_{att}(\text{BERT}(S_{desc}), \text{BERT}(H_{scene})) \quad (5)$$

where X_{att} denotes a cross-attention operation of scene description features on scene history features. Using cross-attention, we combined the scene description features and previous sentences in the scene, selecting relevant features.

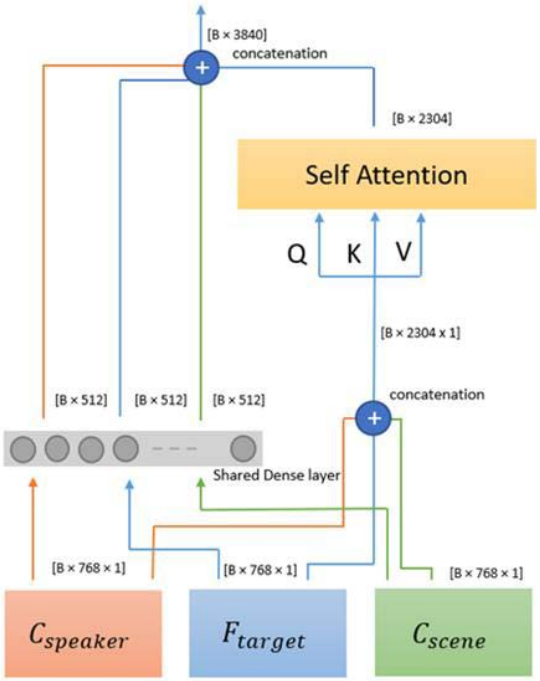

FIGURE 6. Attention-based fusion using self-attention and scaled residual features.

3) CONTEXT FUSION

We used an attention-based fusion method to combine different types of contexts with the features from a target sentence. The features F_{target} are extracted from the sentence (s_t, p_i), using a pre-trained BERT model to combine with the contextual features obtained from the speaker-level and scene-level context modules as indicated in Fig. 6.

The proposed fusion mechanism involves the fusion of the speaker-level and scene-level features with the target sentence features. The speaker-level context $C_{speaker}$, scene-level context C_{scene} and the target sentence features F_{target} are concatenated and the self-attention is computed using scaled dot-product attention [28] computed as $A(Q, K, V) = softmax((QK^T)/\sqrt{K}).V$. Simultaneously, the residual features are computed from the speaker-level context, scene-level context, and the target sentence features using a shared linear layer to reduce the size of the features. Then, the residual features are concatenated to the self-attended fused representation for classification as demonstrated in Fig. 6.

The fused representation is passed through a multilayer perceptron (MLP) classifier with a sigmoid layer to output three-class classification labels. The Adam optimizer [29] was used to train the network with a learning rate of 0.001. The hyperparameters used in the experiment are shown in Table 4. To alleviate the class imbalance problem in the dataset, we used focal loss as an objective function. We implemented the baseline model using the PyTorch framework and conducted experiments on Nvidia Ge-Force RTX 3080Ti GPU with 12 GB of memory.

TABLE 4. Hyperparameter settings.

Parameter	Settings
Loss function	Focal loss
Optimizer	ADAM
Batch size	32
No. of epochs	30
Learning rate	0.001

V. RESULTS AND DISCUSSION

The KD-EmoR dataset was divided into training, validation, and testing sets, each constituting 60%, 20%, and 20% of the total scenes, respectively. The training, validation, and testing sets included 7339, 2566, and 2384 sentences. The data set was not shuffled to preserve the sequential information and the test set was kept separate from the training process.

A. EXPERIMENT RESULTS

We evaluated the performance of the baseline model using micro-F1, macro-F1, weighted-F1, and Mathews Correlation Coefficient (MCC). Micro-F1 computes the F1 score by counting the total false positives, false negatives, and true positives globally. Macro-F1 calculates the unweighted average f1 score for each. The weighted F1 score is calculated by computing F1 metrics for each class and is weighted by the total number of samples for each class. As the dataset is unbalanced as shown in Fig. 3, we evaluated the results using MCC. Since F1-score does not consider the true negatives, MCC can additionally be used to evaluate classification accuracy [30].

The classification results using the proposed model, XAF-SS, are presented in Table 5. The proposed model demonstrated better performance than several baseline methods. We compared the results with TextCNN [36], TextRNN [37], TextRCNN [38], and BERT-based classifier as proposed in [39] for sentiment analysis. These models use a single sentence as input for sentiment classification without using any context. The TextCNN achieved the best classification results when no context was considered. However, spoken sentences without contextual information are insufficient for emotion classification as the meaning and the sentiment of a spoken sentence depend on previously spoken sentences. Therefore, we included ContextBERT [39] for comparison which uses historical information as context. With the cross-attention-based context fusion between speaker-level context and scene-level context, the proposed model exhibited improved results compared to the existing methods.

B. ABLATION STUDY

To evaluate the contribution of context and feature fusion modules in the proposed model, we conducted the ablation experiments with four variations, as shown in Table 6. The

TABLE 5. Experiment results.

Model	Macro F1	Weighted F1	Micro F1	MCC
TextCNN [36]	0.4538	0.5792	0.6284	0.2480
TextRNN [37]	0.2641	0.4417	0.5885	0.0527
TextRCNN [38]	0.4692	0.5859	0.6246	0.2486
BERT [39]	0.3645	0.4798	0.5864	0.1651
ContextBERT [39]	0.4681	0.5592	0.6028	0.1993
XAF-SS (proposed baseline)	**0.4989**	**0.5983**	**0.6288**	**0.2672**

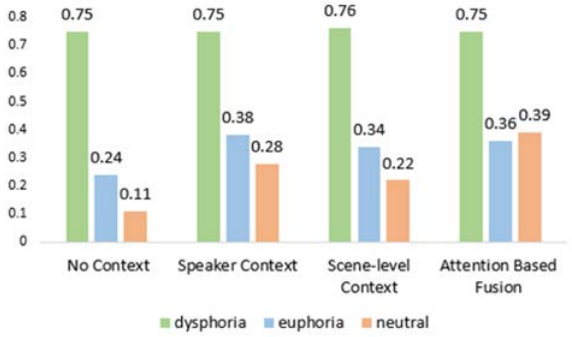

FIGURE 7. F1-scores for each emotion label with different experiment settings for selecting contextual information.

naïve model based only on the spoken sentences with no contextual information had a micro F1 score of 0.6053, whereas the use of each contextual module enhanced the emotion recognition performance. Although the scene-level context did not outperform the sentence-only approach, an efficient fusion with the speaker-level context enhanced the performance. The model with an attention-based fusion of the speaker-level context and scene-level context performed the best revealing the importance of both types of contexts. Fig. 7. illustrates the results for each emotion label when different types of contexts are used. Higher performance for dysphoria was observed due to the presence of a greater number of dysphoric samples. Although the attention-based fusion caused only a small increment in the overall F1-score, the classes with fewer samples (neutral and euphoria) displayed improved performance as shown in Fig. 7.

C. EXPERIMENT WITH EXTERNAL DATASET

We evaluated the proposed model on publicly available testing datasets. Due to the specific target domain, no similar datasets with scene-based conversations were found. Therefore, we compared the proposed model with closely related text-based datasets. The MELD dataset consists of multimodal modalities, including conversational texts labeled with emotions. We used the dialogues in the MELD dataset as analogous to the scenes in the proposed dataset. Therefore, the scene context was based on the conversation history due to the absence of narrative drama scene descriptions. Sentiment levels of positive, negative, and neutral were considered equivalent to euphoria, dysphoria, and neutral for this experiment.

We evaluated the model using the F1-score and MCC. As shown in Table 7, the baseline model had similar performance on both datasets using the weighted F1-score and micro F1-score whereas higher performance was reached using the macro F1-score and MCC. Despite the difference in semantics of the emotion labels and scene context information, high-performance results were observed across the datasets. The lack of datasets with behavioral emotion labels such as euphoria and dysphoria, indicates the need for more datasets to analyze contextual information.

D. DISCUSSION, LIMITATIONS, AND FUTURE WORK

The proposed baseline method involves embedding the spoken methods using a pre-trained BERT embedding module and learning the contextual information using attention-based fusion of different context representations. Using TV show scripts, we built the dataset close to real-world conversations. The utterances can be identified easily in spoken sentences due to recognizable pauses, therefore utterances are commonly used in emotion recognition. However, the utterances have no reasonable demarcation for text-based conversations, such as movie scripts. The conversation scripts usually contain one or more sentences spoken by the characters in their turn.

Therefore, we adopted the turn-based samples instead of arbitrarily splitting them into utterances. Traditionally, emotions in conversations are studied using basic categorical emotions or numeric representations in dimensions of arousal, valence, or dominance. However, more complex emotions, such as euphoria and dysphoria, may be experienced in various daily-life situations. Although these psychological states have been studied to describe the psychological abnormalities due to the influence of addictive substances, mild euphoric or dysphoric conditions in typical situations can be detected for early diagnosis of severe mental disorders. We annotate the dataset using these labels to identify early signs of such morbid emotions. A neutral class may not

TABLE 6. Ablation experiment results.

Feature sets	Macro F1	WeightedF1	Micro F1	MCC
Sentence only (No context)	0.3658	0.5024	0.6053	0.1643
With speaker context	0.4706	0.5699	0.6229	0.2405
With scene context	0.4388	0.5502	0.6195	0.2241
XAF-SS (Proposed baseline)	**0.4989**	**0.5983**	**0.6288**	**0.2672**

TABLE 7. Experiment with other datasets.

Dataset	Macro F1	Weighted F1	Micro F1	MCC
MELD [3]	**0.5968**	**0.6298**	**0.6342**	**0.4070**
KDEmoR (Ours)	0.4989	0.5983	0.6288	0.2672

necessarily represent the absence of emotions; instead, it represents the absence of euphoric or dysphoric psychological states.

The KD-EmoR dataset allows the analysis of emotions in real-world conversation scenarios. We hope that the proposed text-based emotion recognition dataset will be helpful in several tasks, such as enhancing text summarization and automatic synopsis generation of the TV shows through better representation of the emotional states of the characters' emotional states. Likewise, this dataset can be applied in various domains, such as suggesting more accurate sentence completion during online conversations and understanding social media posts to analyze the emotional state of the emotionally challenged or depressed populations who may require clinical support.

Despite being able to display real-world scenarios, conversations in TV shows often contain intensified emotions to attract viewers. In addition, due to the absence of video modality, some conversations may miss the conveyed message limiting the accurate representation of the contextual information. Moreover, the expression of emotions in the TV show scripts is limited by the genre and cultural aspects of the story. In future research, we plan to extend the dataset by including more dynamic cultural scenarios for cross-cultural analysis of emotions. We will explore more context fusion techniques to use inter-speaker and historical context more effectively.

VI. CONCLUSION

In this paper, we proposed a dataset for emotion recognition in the scene transcripts of TV shows. The dataset was annotated with three labels euphoria, dysphoria, and neutral to represent the emotional state of the characters on the TV show. The dataset was annotated by 64 annotators with a fair inter-rater agreement (Fleiss kappa = 0.27). We presented a context-aware deep learning method for the classification of emotion labels. The proposed model achieved an F1-score of 0.6288 on the testing set.

As a further study, we plan to extend the dataset with additional TV shows to include diverse real-life situations. We will improve the performance of the baseline method through improved context representation and fusion techniques. The dataset is publicly available for academic use to contribute to emotion recognition research.

REFERENCES

[1] Y.-J. Lee and H.-J. Choi, "Comparative study of emotion annotation approaches in Korean dialogue," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Jeju Island, South Korea, Jan. 2021, pp. 354–357, doi: 10.1109/BigComp51126.2021.00077.

[2] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "EmotionLines: An emotion corpus of multi-party conversations," 2018, *arXiv:1802.08379*.

[3] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," Jun. 2019, *arXiv:1810.02508*. Accessed: May 4, 2022.

[4] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008, doi: 10.1007/s10579-008-9076-6.

[5] D. Lee, K.-J. Oh, and H.-J. Choi, "The chatbot feels you–a counseling service using emotional response generation," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Jeju Island, South Korea, Feb. 2017, pp. 437–440, doi: 10.1109/BIGCOMP.2017.7881752.

[6] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," May 2018, *arXiv:1704.01074*. Accessed: Jul. 4, 2022.

[7] S. Mohammad, "Portable features for classifying emotional text," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2012, pp. 587–591.

[8] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, vol. 11, 2011, pp. 450–453.

[9] N. F. F. Da Silva, E. R. Hruschka, and E. R. Hruschka, Jr., "Tweet sentiment analysis with classifier ensembles," *Decis. Support Syst.*, vol. 66, pp. 170–179, Oct. 2014, doi: 10.1016/j.dss.2014.07.003.

[10] S. Taj, B. B. Shaikh, and A. F. Meghji, "Sentiment analysis of news articles: A lexicon based approach," in *Proc. 2nd Int. Conf. Comput., Math. Eng. Technol. (iCoMET)*, Sukkur, Pakistan, Jan. 2019, pp. 1–5, doi: 10.1109/ICOMET.2019.8673428.

[11] C. Strapparava and R. Mihalcea, "SemEval-2007 task 14: Affective text," in *Proc. 4th Int. Workshop Semantic Eval.*, 2007, pp. 70–74.

[12] D.-A. Phan, Y. Matsumoto, and H. Shindo, "Autoencoder for semisupervised multiple emotion detection of conversation transcripts," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 682–691, Jul. 2021, doi: 10.1109/TAFFC.2018.2885304.

[13] P. Ekman, "Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique," *Psychol. Bull.*, vol. 115, no. 2, pp. 268–287, 1994.

[14] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.
[15] R. Plutchik, "A general psychoevolutionary theory of emotion," *Emotion: Theory, research, and Experience: Theories of Emotion*, vol. 1, R. Plutchik and H. Kellerman, Eds. New York, NY, USA: Academic, 1980, pp. 3–33.
[16] W. G. Parrot, *Emotions in Social Psychology: Essential Readings*. Hove, U.K.: Psychology Press, 2001, pp. 30–40.
[17] R. Wilmot, "Euphoria," *J. Drug Issues*, vol. 15, no. 2, pp. 155–191, 1985.
[18] V. Starcevic, "Dysphoric about dysphoria: Towards a greater conceptual clarity of the term," *Australas. Psychiatry*, vol. 15, no. 1, pp. 9–13, 2007.
[19] D. Zillman and J. R. Cantor, "Affective responses to the emotions of a protagonist," *J. Exp. Social Psychol.*, vol. 13, no. 2, pp. 155–165, Jan. 1977, doi: 10.1016/S0022-1031(77)80008-5.
[20] S. M. Zahiri and J. D. Choi, "Emotion detection on TV show transcripts with sequence-based convolutional neural networks," in *Proc. AAAI Workshops*, 2018, pp. 44–52.
[21] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, "AraSenTi-tweet: A corpus for Arabic sentiment analysis of Saudi tweets," *Proc. Comput. Sci.*, vol. 117, pp. 63–72, Jan. 2017, doi: 10.1016/j.procs.2017.10.094.
[22] B. Alharbi, H. Alamro, M. Alshehri, Z. Khayyat, M. Kalkatawi, I. I. Jaber, and X. Zhang, "ASAD: A Twitter-based benchmark Arabic sentiment analysis dataset," Mar. 2021, *arXiv:2011.00578*. Accessed: Jul. 5, 2022.
[23] E. Refaee and V. Rieser, "An Arabic Twitter corpus for subjectivity and sentiment analysis," in *Proc. 9th Int. Conf. Lang. Resour. Eval.*, 2014, pp. 2268–2273.
[24] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tutor Quantum Methods Psychol.*, vol. 8, no. 1, pp. 23–34, Feb. 2012, doi: 10.20982/tqmp.08.1.p023.
[25] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.
[26] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 4171–4186.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Dec. 2017, *arXiv:1706.03762*. Accessed: Jun. 2, 2022.
[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[30] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp.1–13, Dec. 2020, doi: 10.1186/s12864-019-6413-7.
[31] C. D. Katsis, Y. Goletsis, G. Rigas, and D. Fotiadis, "A wearable system for the affective monitoring of car racing drivers during simulated conditions," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 3, pp. 541–551, Jun. 2011, doi: 10.1016/j.trc.2010.09.004.
[32] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015, pp. 1–15.
[33] Z. Shen, J. Wang, Z. Pan, Y. Li, and J. Wang, "Cross attention-guided dense network for images fusion," Aug. 2022, *arXiv:2109.11393*. Accessed: Oct. 18, 2022.
[34] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," in *Proc. Interspeech*, Sep. 2019, pp. 3569–3573.
[35] R. G. Praveen, W. C. de Melo, N. Ullah, H. Aslam, O. Zeeshan, T. Denorme, M. Pedersoli, A. L. Koerich, S. Bacon, P. Cardinal, and E. Granger, "A joint cross-attention model for audio-visual fusion in dimensional emotion recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2486–2495.
[36] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751, doi: 10.3115/v1/D14-1181.
[37] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*. Palo Alto, CA, USA: AAAI Press, 2016, pp. 2873–2879.
[38] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2267–2273. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/viewPaper/9745
[39] S. Feng, N. Lubis, C. Geishauser, H.-C. Lin, M. Heck, C. van Niekerk, and M. Gasic, "EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems," in *Proc. 13th Lang. Resour. Eval. Conf.*, Marseille, France, Jun. 2022, pp. 4096–4113. [Online]. Available: https://aclanthology.org/2022.lrec-1.436
[40] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning 'BERT-like' self supervised models to improve multimodal speech emotion recognition," in *Proc. Interspeech*, Oct. 2020, pp. 3755–3759, doi: 10.21437/Interspeech.2020-1212.
[41] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 6558–6569, doi: 10.18653/v1/P19-1656.
[42] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020, doi: 10.1109/ACCESS.2020.3026823.
[43] M. R. Basso, B. K. Schefft, and R. G. Hoffmann, "Mood-moderating effects of affect intensity on cognition: Sometimes euphoria is not beneficial and dysphoria is not detrimental," *J. Pers. Social Psychol.*, vol. 66, no. 2, p. 363, 1994.
[44] Y. I. Russell, R. I. Dunbar, and F. Gobet, "Euphoria versus dysphoria: Differential cognitive roles in religion?" in *Attention, Representation, and Human Performance: Integration of Cognition, Emotion, and Motivation*, S. Masmoudi, D. Y. Dai, and A. Naceur, Eds. Hove, U.K.: Psychology Press, 2012, pp. 147–165.
[45] J. A. Russell, A. Weiss, and G. A. Mendelsohn, "Affect grid: A single-item scale of pleasure and arousal," *J. Pers. Social Psychol.*, vol. 57, no. 3, pp. 493–502, Sep. 1989.

SUDARSHAN PANT received the B.S., M.S., and Ph.D. degrees from Mokpo National University, South Korea. He is currently a Postdoctoral Researcher with Chonnam National University. His research interests include machine learning, affective computing, and healthcare AI.

EUNCHAE LIM is currently pursuing the Ph.D. degree with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. She was a Researcher at the Department of Artificial Bio-Robot, Osong Medical Innovation Foundation, from 2019 to 2020. Her research interests include exploratory data analysis, emotion recognition, natural language processing, and the language model of the dialogue systems.

HYUNG-JEONG YANG received the B.S., M.S., and Ph.D. degrees from Chonbuk National University, South Korea. She is currently a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, South Korea. Her main research interests include multimedia data mining, medical data analysis, social network service data mining, and video data understanding.

GUEE-SANG LEE received the B.S. degree in electrical engineering and the M.S. degree in computer engineering from Seoul National University, South Korea, in 1980 and 1982, respectively, and the Ph.D. degree in computer science from Pennsylvania State University, in 1991. He is currently a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. His research interests include image processing, computer vision, and video technology.

SOO-HYUNG KIM received the B.S. degree in computer engineering from Seoul National University, in 1986, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology in 1988 and 1993, respectively. Since 1997, he has been a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. His research interests include pattern recognition, document image processing, medical image processing, and ubiquitous computing.

YOUNG-SHIN KANG received the B.A. degree in English literature and language and the M.A. degree in counseling psychology from Chonnam National University, South Korea, and the Ph.D. degree in counseling psychology from Northeastern University, USA. She is currently a Professor with the Department of Psychology, Chonnam National University. Her research interests include emotion recognition and contagion, affect regulation, traumatic stress, and posttraumatic growth.

HYERIM JANG is currently pursuing the Ph.D. degree in psychology with Chonnam National University, South Korea. She has been working as a Researcher with the Department of Psychology, since 2019, and as a Counselor, since 2021. Her research interests include loss, post-traumatic growth, and COVID-19.