Transfer Learning Model for Contextual Feature Extraction and Emotion Analysis in Dialogues

Seon-Haeng Kim*, So-Young Jun**, and Jong-Woo Kim***

Abstract

This paper proposes a KoGPT2-based transfer learning model, Context-KoGPT2, designed to identify and analyze contextual characteristics within Korean dialogues. By assigning specific weights, it aims to reflect the cumulative atmosphere of conversations and was tested utilizing a Korean dialogues dataset across seven emotional categories. The results present that the model, particularly with a weight of 0.9, significantly enhances emotion classification performance by taking into account the contextual cues within the dialogue. Furthermore, it is confirmed that the model effectively diminishes the influence of previous sentences based on a thorough understanding of time gaps and the earlier context in conversation. It offers practical implications for real-time systems requiring swift emotional recognition.

요 약

본 논문에서는 한국어 대화에서 문맥적 특성을 파악하고 분석하기 위해 고안된 KoGPT2 기반 전이 학습 모델인 Context-KoGPT2를 제안한다. 특정 가중치를 부여하여 대화의 누적된 분위기를 반영하는 것을 목표로 하며, 7가지 감정 범주에 걸친 한국어 대화 데이터 세트를 활용하여 검증되었다. 실험 결과, 이 모델은 대화 내 문맥적 단서를 고려함으로써 감정 분류 성능을 크게 향상시켰다. 특히 가중치를 0.9로 설정한 모델은 대화의 시간적 간격과 이전 문맥에 대한 철저한 이해를 바탕으로 이전 문장의 영향을 효과적으로 감소시키는 것으로 확인되었다. 이는 신속한 감정 인식이 필요한 실시간 시스템에 실용적인 시사점을 제공한다.

Keywords

contextual information, dialogue utterances, emotion analysis, feature extraction, natural language processing, transfer learning

* Associate, Production AIX Task Team, LG Electronics
 - ORCID: https://orcid.org/0009-0002-5366-6626
** Intern, Data Team, Storelink
 - ORCID: https://orcid.org/0000-0001-8187-5199
*** Professor, School of Business, Hanynag University
 - ORCID: https://orcid.org/0000-0001-9484-5968

Ⅰ. Introduction

Recently, the volume of text data has been exponentially increasing due to the development of social network services and the widespread use of smartphone[1]. This has led to abundant data, including comments, news articles, and product reviews.

Extracting valuable insights from this vast amount of data, known as text mining, has become increasingly important[2][3]. Sentiment analysis, a text mining technique, plays a crucial role in obtaining valuable insights by analyzing subjective information related to opinions, emotions, and attitudes expressed in texts.

Sentiments can be classified into binary categories, "positive or negative," or multiple categories. With the emergence of chatbot systems that interact with users seeking assistance, research in this field has become more active, including sentiment analysis of conversational texts[4]-[6].

Sentiment analysis and emotion recognition are highly significant for user satisfaction and can be applied in various scenarios, such as collecting speaker opinions during conversations and incorporating feedback into robotic agents[7].

Real-time dialogue analysis can enhance human-computer interactions by training conversations with specific emotions. Given that dialogue typically involves multiple speakers, understanding contextual features by referring to previous utterances and analyzing sentences are essential, as shown in existing studies.

Dialogue-based sentiment analysis generally involves three steps: (1) obtaining contextual information, (2) understanding how situations influence utterances, and (3) extracting emotional features for classification.

Previous models like cLSTM[8], DialogueRNN[9], and DialogueGCN[10] employ complex deep neural network structures to comprehend dialogue situations and predict the sentiment of speakers based on preceding conversations.

However, research on sentiment analysis specifically targeting Korean conversational text datasets has not been as actively pursued as with English datasets [11]-[13].

Due to linguistic differences, natural language processing for Korean requires a different approach than English.

For instance, the Bidirectional Encoder Representations from Transformers(BERT) model[14], a widely used transfer learning model for multiple languages, employs a WordPiece tokenizer that creates a corpus based on word spacing.

While this approach is applicable to languages with word segmentation, it is inadequate for agglutinative languages like Korean that use postpositions, endings, and affixes.

To process Korean text effectively, it is necessary to decompose it into smaller units, such as morphemes or characters, rather than relying on spacing[15].

Character-level analysis proves to be suitable for resolving out-of-vocabulary(OOV) issues and demonstrates improved accuracy and decomposition bias[16]. Hence, this study proposes using a language model that reflects the characteristics of Korean and applies Byte Pair Encoding(BPE) at the character level.

Despite the differences in natural language processing between Korean and English, there is a lack of research focusing on Korean dialogue analysis.

Previous studies utilizing Korean conversation datasets predominantly employed deep learning technologies to analyze speech acts rather than conducting multi-category sentiment classification. Therefore, this study showed discrete emotion analysis that incorporates dialogue context by training on a substantial amount of Korean conversation datasets available as open sources.

Furthermore, this study analyzed the performance of the latest Korean natural language processing technologies using the Korean dialogue dataset, aiming to enhance sentiment analysis performance through deep learning techniques.

In this paper, we propose using the Context-KoGPT2 model, which integrates previous contextual information with the results of individual sentence analysis. This model effectively captures the influence of contextual information on utterances and provides a simplified structure for extracting features used in emotion classification.

Additionally, we assign lower weight values to utterances farther away in the conversation, ensuring that the most recent speech has a more decisive influence on classifying the present emotion.

The contributions of this study are as follows. Firstly, the proposed model incorporates contextual information from preceding utterances to predict the sentiment of the current statement. The model's effectiveness in capturing contextual information in emotion classification holds practical implications for various real-time systems.

By considering previous contexts, the proposed model helps chatbots achieve a better understanding and responsiveness in dynamic conversations with improved results in emotion analysis.

It can be effectively applied in chatbot systems, counseling systems, and AI speakers, enabling more accurate and context-aware emotional recognition in interactive agent technology.

Additionally, the study suggests the potential of KoGPT2, a Korean transfer learning model, in analyzing emotions in dialogues. By leveraging the contextual cues captured, the model surpasses traditional approaches that solely focus on individual sentences. This finding contributes to the advancement of techniques specifically designed for analyzing emotions in Korean dialogues.

The remainder of this paper is organized as follows: In Section 2, we review related research and discuss the researches in the field of sentiment analysis. Section 3 presents our proposed model for emotion analysis in dialogue. We describe its architecture and methodology in detail.

In Section 4, we provide information about the experimental setup and evaluate the performance of our model by comparing it with other sentiment analysis models. Furthermore, Section 5 presents the conclusions drawn from our study and discusses implications for future research.

Ⅱ. Related Works

Sentiment analysis, or opinion mining, is one of the primary tasks in natural language processing(NLP). It involves analyzing subjective information and determining sentiment scores at various levels, including document, sentence, phrase, and aspect levels [17]-[21].

In related studies, lexicon-based approaches have been used to calculate positive/negative scores of sentences based on sentiment dictionaries[22].

However, while these approaches, along with machine learning models such as naive Bayesian classifiers, support vector machines(SVM), and artificial neural networks, can classify sentences with binary or multi-sentiment polarity, their performance heavily relies on lexical resources. Due to this dependency, lexicon-based sentiment analysis has faced challenges in achieving higher performance.

Moreover, machine learning-based methods have shown vulnerability to domain adaptation, as they are highly influenced by the characteristics of the training data.

Subsequently, advances in deep learning models have been employed to tackle the existing challenges [23]. The combination of word embeddings and deep learning architectures, such as convolutional neural networks(CNN) and recurrent neural networks(RNN), has exhibited excellent performance in sentiment analysis[24]-[26].

This success stems from not only the extraction of feature sets from training data but also the automatic extraction of complex hierarchical features.

Furthermore, the development of transfer learning models like ELMo[4], GPT/GPT2[27], and BERT[14] has overcome the limitations posed by insufficient training data and further improved the performance of NLP tasks.

In recent developments in natural language processing, there has been a notable trend of extracting contextual features from dialogues. Analyzing emotions within a conversation composed of multiple sentences has become feasible.

The process begins by segmenting the dialogue into sentences and extracting feature values for each sentence using syllable-based CNNs. Subsequently, contextualized features are extracted for the entire dialogue using RNNs[28][29].

Previous studies on Korean conversation primarily focused on unveiling the speaker's intention in the dialogue, mainly through speech act analysis, to enhance the practicality of the research(refer to Table 1).

While numerous studies have employed various deep learning techniques for emotion analysis or dialogue act classification, the latest models that can enhance NLP performance, such as KoBERT or KoGPT2, have not been utilized with dialogue datasets.

Hence, our approach aimed to extract contextual features and classify the speaker's emotions by combining deep learning and transfer learning models.

III. Proposed Approach

3.1 Emotional flow model

In the context of sentiment analysis in dialogues, an emotional analysis approach is essential to identify contextual features by taking into account previous utterances alongside existing sentiment analysis techniques.

For instance, consider a dialogue consisting of two utterances: "I am happy" and "Me too." Merely examining the second utterance, "Me too," makes it challenging to comprehend its sentiment. However, considering the contextual situation and the previous utterance, one can understand that the sentiment expressed is also "happiness."

Understanding contextual features is crucial in sentiment analysis within a dialogue. This study proposes the sentiment analysis model, which reflects the emotional flow and context to capture these contextual features effectively.

Table 1. Comparison of studies using Korean dataset

Study	Task		Model		Data	
	Sentiment analysis	Speech act analysis	Machine learning	Transfer learning	Short sentence	Dialogue
[10]	○		SVM			○
[11]	○	○	LSTM			○
[12]		○	CNN, LSTM			○
[13]	○	○		KoBERT	○	
[30]		○	RNN-CNN			○
[31]		○	SVM, CNN CNN-RNN			○
This paper	○		CNN-Attn. CNN-LSTM CNN-LSTM-Attn.	KoBERT KoGPT2		○

As illustrated in Fig. 1, the model is designed to predict the speaker's sentiment by capturing the characteristics of individual short sentences and the overall emotional flow within the dialogue.

To extract characteristic values from short sentences, we employ a character-level CNN(CharCNN) as commonly used in existing approaches. To comprehend the emotional flow of the dialogue, we utilize a Long Short-Term Memory(LSTM) model that considers the sentiment expressed in previous sentences.

For the first sentence, which does not have a previous utterance, the model cannot determine the previous sentiment flow, so it uses a CharCNN to extract the information represented by u_1 to predict the sentiment of that utterance.

Starting from the second sentence where the previous statement exists, we also characterize the emotional flow. Using the value of y_1, which is the value predicted by the CharCNN model, we use LSTM to concatenate s_1, the emotional flow feature, and u_2, the information of the utterance, to predict the emotion of the utterance.

This model extracts information from each utterance, represented as u_n, and captures the emotional flow through s_n obtained from the LSTM, which connects these two information sources.

3.2 Contextual features model

3.2.1 Input embedding

All utterances in the dialogue are tokenized using the BPE algorithm. The detailed process is illustrated in Fig. 2, where the statements are translated into English.

Speaker	Utterance	Emotion
A	You're fine without me.	Sadness
B	No. Do not leave me.	Sadness
A	I'm sorry, but I have to go.	Sadness

Representation	Label
[BOS] You're fine without me. [SEP1]	Sadness
[BOS] You're fine without me. [SEP1] No. Do not leave me. [SEP2]	Sadness
[BOS] You're fine without me. [SEP1] No. Do not leave me. [SEP2] I'm sorry, but I have to go. [SEP3]	Sadness

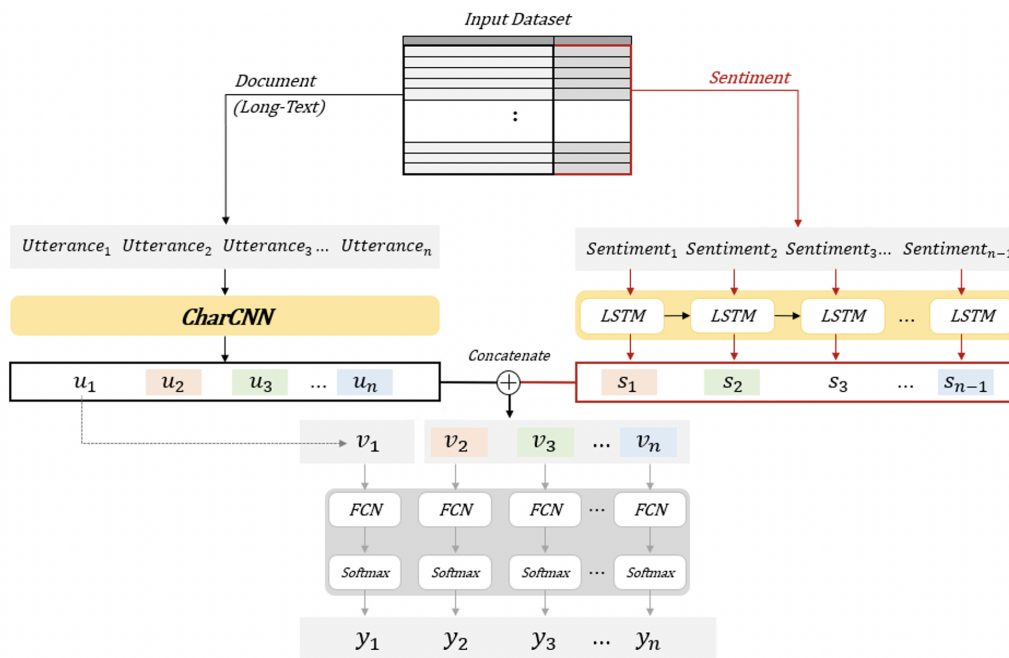Fig. 2. Exemplar representation of utterances

Fig. 1. Model of CharCNN and LSTM

The process involves adding a unique token [BOS] (Beginning of Sentence) at the beginning of the first utterance, inserting the [SEP] token(Separator) at the end of each utterance, and concatenating the utterances that belong to the same statement into a single sentence.

3.2.2 Language model decoder

Fig. 3 shows the method utilizing the transformer[32]-based pre-trained language model KoGPT2. This approach effectively addresses long-term dependency issues and enables contextual information extraction from time-series dialogues.

Language models offer high efficiency and performance as they can leverage pre-trained parameters for various tasks. The proposed model undergoes fine-tuning at all levels to enhance its adaptability for language modeling and emotion classification tasks.

In the proposed model, the tokenized input is ($E_{1,1}$, $E_{1,2}$, $\cdots$, E_{1,i_1}). The contextual token values are represented as ($T_{1,1}$, $T_{1,2}$, $\cdots$, T_{1,i_1}). Since the number of tokens may differ for each dialogue, applying the

maximum pooling technique can be beneficial in preserving important information at each level.

At the next step, the learning models calculate the final loss by incorporating contextual features. This is achieved by summing the cross-entropy loss values of the current sentence and the previous sentences. The final loss is computed using the following equation:

$$FL = \sum_{i=1}^{n} CE(pred_i, true_i) \cdot w^{n-i} \qquad (1)$$

$$CE(pred_i, true_i) = -\sum_{i=1}^{n} true_i \cdot \log(pred_i) \qquad (2)$$

In Eq. (1), $pred_i$ represents the predicted sentiment of the i-th sentence in the dialogue,

which includes a total of n sentences learned thus far.

On the other hand, $true_i$ represents the labeled sentiment in the dataset for the i-th sentence. Additionally, w denotes the weight applied to reduce the influence of previous sentences. We have raised the weight to the power of $(n-i)$ so that sentences uttered earlier than the current sentence receive a lower weight.
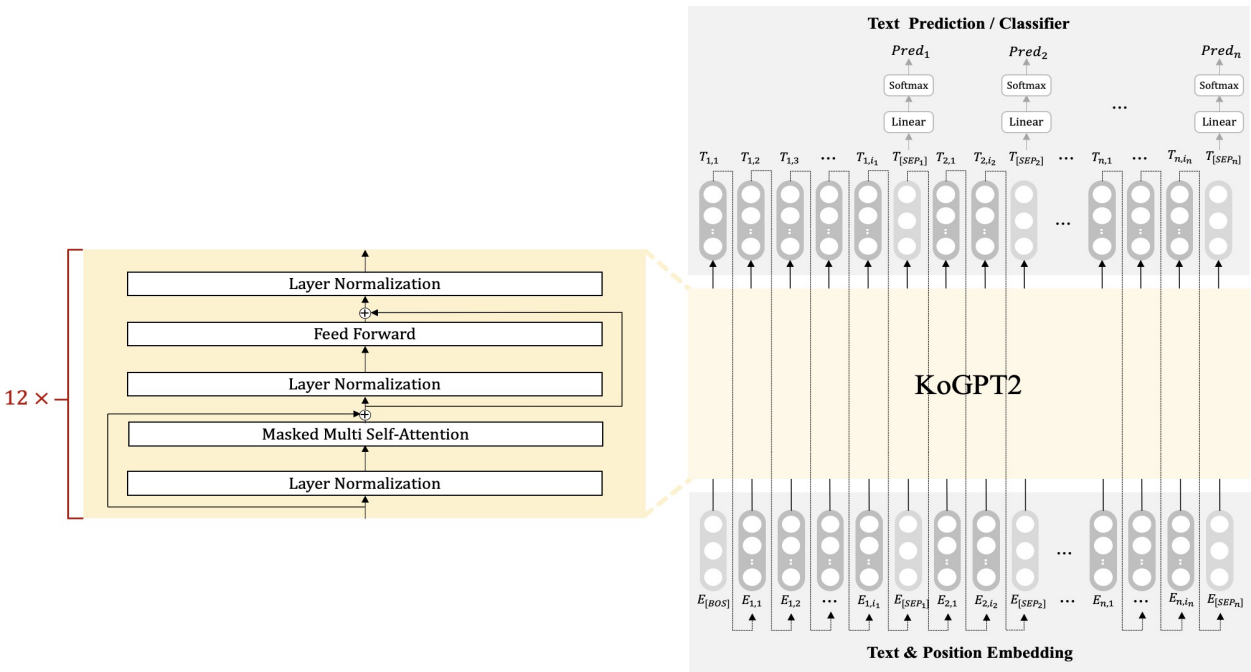

Fig. 3. Model architecture of Context-KoGPT2

In Eq. (2), if the actual sentiment($true_i$) and the predicted sentiment($pred_i$) are the same, the value converges to 0, but, if there is a difference between the two, the value increases. Therefore, this equation represents the cross-entropy loss, the difference between the actual and predicted sentiments.

IV. Experiments

4.1 Empirical setting

4.1.1 Dataset

This study conducted an experiment on conversational data consisting of six sentences per conversation between two speakers. This open-source dataset was built as a data construction project for AI learning, hosted by the Ministry of Science and ICT and supported by the National Information Society Agency.

We utilized a portion of this dataset comprising 12,000 conversations and 72,000 utterances as our experimental data.

To label the collected data, we used a cross-validation approach, where three individuals directly tagged the multi-category emotions while considering the short sentence and its context. The emotions used in the experiment were classified into six categories(surprise, fear, sadness, anger, joy, and disgust), as aligned with existing literature.

Additionally, we included a neutral category, representing a state of no discernible sentiment, resulting in seven emotion classifications[33]. The distribution of utterances per emotion in the collected data is presented in Table 2.

Table 2. Distribution of utterances for each emotion

Emotion	Size of utterances
Surprise	5,843
Fear	6,720
Sadness	12,568
Anger	12,495
Joy	11,440
Disgust	4,694
Neutral	18,150
Total	72,000

Furthermore, we employed K-fold cross- validation to mitigate model evaluation bias and enhance the reliability of performance metrics. This approach guarantees that all data points are used as a part of the test set at least once, aiding in the selection of a relatively better-performing model through the analysis of unseen data.

Table 3. Data distributions between folds

Fold	Data	Sentiment							Total
		Surprise	Fear	Sadness	Anger	Joy	Disgust	Neutral	
1	Train	5116	6033	11632	11518	10290	4227	15984	64800
	Test	727	687	1026	977	1150	467	2166	7200
2	Train	5094	6119	11767	11590	10382	4247	15601	64800
	Test	749	601	891	905	1058	447	2549	7200
3	Train	5173	6089	11720	11537	10299	4346	15636	64800
	Test	670	631	938	958	1141	348	2514	7200
4	Train	5082	5913	11588	11421	10382	4142	16272	64800
	Test	761	807	1070	1074	1058	552	1878	7200
5	Train	5329	5887	11169	11156	10020	4129	17110	64800
	Test	514	833	1489	1339	1420	565	1040	7200
6	Train	5348	6161	11158	10986	10338	4282	16527	64800
	Test	495	559	1500	1509	1102	412	1623	7200
7	Train	5330	6280	11099	11179	10549	4255	16108	64800
	Test	513	440	1559	1316	891	439	2042	7200
8	Train	5347	5580	11166	11161	9953	4061	17532	64800
	Test	496	1140	1492	1334	1487	633	618	7200
9	Train	5401	6071	11346	11001	10272	4194	16515	64800
	Test	442	649	1312	1494	1168	500	1635	7200
10	Train	5367	6347	11277	10906	10475	4363	16065	64800
	Test	476	373	1381	1589	965	331	2085	7200

For our experiment, we divided the training and test datasets into nine folds and one fold, respectively, resulting in 10 training and evaluation iterations(K=10) using different fold combinations. The distribution of data between folds is presented in Table 3.

4.1.2 Simulation parameters

Table 4 displays the weight values assigned based on the position of the previous sentence relative to the current point. These weights capture the influence of previous utterances on the current utterance within the dialogue. As the conversation unfolds, the impact of utterances closer to the present point becomes more significant for future utterances compared to those farther away.

In this study, we defined weight values (w) of 1, 0.9, 0.8, and 0.7. A weight of 1 indicates that the influence of previous conversations is not considered, while lower weights reflect decreasing importance of earlier utterances in relation to the current point.

Table 4. Weights are applied differently for each sentence point

Sentence point	w=1	w=0.9	w=0.8	w=0.7
t	1	1	1	1
t−1	1	0.9	0.8	0.7
t−2	1	0.81	0.64	0.49
t−3	1	0.73	0.51	0.34
t−4	1	0.66	0.41	0.24
t−5	1	0.6	0.33	0.17

4.1.3 Performance measurement

For the experimental evaluation, we used the dialogue data between two speakers as the dataset. We trained, evaluated, and compared four short sentence models and two context models proposed in this paper.

To assess the performance of each model, we conducted 10-fold cross-validation, where the emotion prediction for each statement was used as the basis. We compared the models by calculating the average

error value and accuracy.

The error value was measured using the cross-entropy error function(CEE), which is commonly used in classification models. The following baseline models will be compared to the proposed model in this study.

1. Char + Word + CNN + Self-Attention

Previous studies have utilized a CNN-based method for handling short texts. This approach takes into account the characteristics of each syllable and morpheme by incorporating word-level embeddings.

The input embeddings are processed in parallel and passed through convolutional layers to extract feature maps. This allows for self-attention mechanisms to selectively focus on relevant features.

Furthermore, by connecting the features through a max-pooling layer, the model analyzes the sentiment of each sentence while minimizing the impact of overfitting.

2. Char + Word + Bi-LSTM + Attention

The Bi-LSTM model extends the capabilities of the LSTM model by adding an LSTM layer in reverse order. This allows for incorporating both syllable and word features through word embeddings generated from syllable and morpheme analysis. By considering the sequential and inverse features of sentences, the model gains a more comprehensive understanding of the text.

Additionally, an attention mechanism is employed to address the gradient loss problem, a limitation of the traditional RNN model. The attention mechanism helps the model focus on essential parts of the input and improves its overall performance.

3. BERT[14]

BERT is a pre-trained language model developed by Google in 2018. It was trained on large corpora, including BooksCorpus(800M words) and Wikipedia (2,500M words), to learn general language representations. BERT utilizes the WordPiece tokenizer,

which breaks down terms into sub-word and considers longer and more frequent sub-words as individual units.

In addition, there is a multilingual version of BERT called BERT-Multilingual. It shares the same model architecture as the original BERT but was trained on Wikipedia corpus from 104 languages.

4. KoBERT

KoBERT was specifically developed to address the performance limitations of Korean language tasks using the BERT-Multilingual model as a starting point. To train KoBERT, a large-scale dataset of 25 million sentences from Korean Wikipedia and news articles was collected by SK Telecom.

To capture the unique characteristics of Korean, KoBERT applies the BPE tokenizer at the character level, which helps handle irregular language changes and improve tokenization accuracy for Korean text.

5. KoGPT2

KoGPT2 is the Korean version of the OpenAI GPT2 model[27] specifically designed to improve the performance of Korean language processing tasks. It was trained using a large dataset consisting of Korean Wikipedia, news articles, and a text corpus containing 152 million sentences.

SK Telecom was responsible for collecting and curating this dataset. Similar to KoBERT, KoGPT2 also employs the BPE tokenizer at the character level

to capture the unique characteristics and handle exceptional changes in the Korean language.

4.2 Experimental results

Table 5 demonstrates the results of the sentiment classifiers using the conversational dataset, which comprised 72,000 sentences. The experimental results present the performance of the proposed and comparison models in this study.

Models 1 through 5 correspond to the baseline model for the short sentences given in the previous section, and Models 6 and 7 refer to the models for analyzing dialogues with context. Model 6 is CharCNN+LSTM shown in Figure 1, and Model 7 is the proposed model of this study with different weights of 1, 0.9, 0.8, and 0.7, respectively.

The accuracy(%) was measured through 10-fold cross-validations. Notably, Context-KoGPT2, one of the context models, exhibited superior performance compared to the short sentence models. It achieved an average accuracy that had a 12.56% gap between the minimum values.

While KoBERT and KoGPT2, the transfer learning models for short sentences, displayed higher performance than CharCNN+LSTM for dialogue, they were surpassed by Context-KoGPT2 when applied to dialogue. The highest accuracy was observed with a weight of 0.9(w=0.9).

Table 5. Accuracy in 10-fold cross-validation for each model based on short sentences or dialogue

Type	Model	Fold number									
		1	2	3	4	5	6	7	8	9	10
Short sentence	1	58.5	62.7	60.6	64.8	60.0	65.7	65.3	57.0	65.6	66.1
	2	63.0	66.6	64.8	65.3	63.3	70.0	70.9	62.1	69.3	73.4
	3	64.5	67.7	65.2	67.3	66.3	68.7	70.1	64.2	70.6	70.3
	4	66.3	68.4	68.6	69.8	68.3	69.5	70.2	66.2	71.5	69.8
	5	69.1	70.7	69.5	71.9	70.8	75.6	75.6	68.0	74.6	72.9
Dialogue with context	6	63.7	69.4	68.1	67.2	66.5	70.4	71.9	62.1	71.6	72.2
	7	72.0	72.3	70.6	73.8	72.8	77.2	78.8	71.2	77.0	77.8
		73.1	72.7	71.7	73.9	75.2	77.3	78.8	71.8	77.8	79.6
		71.8	73.4	72.0	74.9	76.0	76.1	77.5	71.0	76.6	79.3
		72.5	71.6	71.8	73.2	73.2	76.7	78.5	69.6	77.0	78.0

Note: The best value per fold is shown in bold.

As a results in Table 6, the proposed model, Context-KoGPT2, consistently achieved the lowest average CEE values in all cases, indicating better performance in sentiment classification. Specifically, the Context-KoGPT2 model with a weight of 0.9 achieved the best performance.

These results further support the effectiveness of incorporating contextual information and utilizing the entire dialogue than individual short sentences, especially when using transfer learning models.

The results presented in Table 7 indicate that the pre-trained KoGPT2 model outperformed other models in both short-text and dialogue tasks. This can be attributed to the fact that KoGPT2 was specifically designed to understand and capture the natural language's features, including the Korean language's unique characteristics.

The lower performance of the CharCNN+LSTM model compared to KoBERT and KoGPT2 in the short-text task suggests that it was confirmed that it is an essential factor to reflect the characteristics of the Korean language as well as the context in the efficiency of the model.

As Korean has a high ratio of homonyms, it is advantageous to provide different vector expressions depend- ing on the context, even for the same word, through a transfer learning model that pre-trains a large amount of Korean corpus in advance[34][35].

Comparing models not fully specialized in Korean, BERT-multilingual showed lower performance than CharCNN+LSTM, despite incorporating some Korean corpus in its pre-training. This highlights the importance of considering contextual features in conversation rather than relying solely on a relatively small Korean corpus.

Table 6. CEE in 10-fold cross-validation for each model based on short sentences or dialogue

Type	Model	Fold number									
		1	2	3	4	5	6	7	8	9	10
Short sentence	1	1.47	1.31	1.42	1.24	1.15	1.18	1.11	1.71	1.31	1.17
	2	1.01	0.94	0.98	0.97	1.03	0.83	0.79	1.11	0.87	0.74
	3	1.04	0.94	0.98	0.97	0.97	0.84	0.83	1.06	0.83	0.83
	4	0.96	0.91	0.89	0.86	0.94	0.88	0.87	0.95	0.89	0.87
	5	0.86	0.83	0.85	0.82	0.88	0.69	0.65	0.99	0.71	0.74
Dialogue with context	6	1.03	0.88	0.87	0.95	0.97	0.84	0.79	1.12	0.84	0.78
	7	0.83	0.81	0.83	0.82	0.82	0.67	0.62	0.90	0.71	0.63
		0.79	0.80	0.85	0.81	0.73	0.70	0.61	0.89	0.71	0.60
		0.85	0.79	0.83	0.76	0.74	0.69	0.65	0.98	0.70	0.60
		0.83	0.83	0.83	0.81	0.81	0.74	0.62	1.07	0.75	0.65

Note: The best value per fold is shown in bold.

Table 7. Results of the accuracy and CEE for each model and text type

Type	Model	Accuracy			CEE		
		Min	Max	Mean	Min	Max	Mean
Short sentence	1	60.07	66.11	62.67	1.118	1.712	1.318
	2	62.15	73.4	66.9	0.744	1.110	0.930
	3	64.25	70.61	67.54	0.829	1.064	0.932
	4	66.35	71.54	68.9	0.873	0.957	0.907
	5	68.02	75.68	71.91	0.656	0.999	0.808
Dialogue with context	6	62.12	72.28	68.36	0.788	1.122	0.916
	7	70.62	78.88	74.40	0.629	0.901	0.769
		71.74	79.65	75.23	0.605	0.899	0.754
		71.02	79.38	74.91	0.609	0.987	0.764
		69.66	78.53	74.25	0.621	1.071	0.797

Note: The best performance values are bolded.

On the other hand, KoGPT2, which specialized in Korean and trained on a sizeable Korean corpus, demonstrated excellent performance in sentiment analysis tasks targeting conversations. These findings emphasize the significance of considering the entire conversation and the need for sentiment analysis models to effectively capture contextual information.

Ⅴ. Conclusion

This paper introduces a context-based modeling approach using KoGPT2 for multi-category sentiment analysis. The study examines 12,000 dialogues, comprising 72,000 Korean utterances labeled with seven emotions.

Four short text and two context models are employed for natural language processing and emotion classification, evaluated based on accuracy and cross-entropy error metrics.

The experimental results demonstrate the model's effectiveness in capturing contextual information from previous utterances and achieving superior performance.

The results of the multi-category sentiment analysis in dialogue confirm the effective capture of contextual information by the proposed model from previous utterances. This study contributes by analyzing dialogue emotions using KoGPT2, a Korean transfer learning model.

The model comprehends the prior conversation's context to predict the emotion of the current statement, demonstrating that superior analysis is achievable through contextual information in dialogue rather than examining short sentences separately.

Consequently, this model can be practically applied to enhance the performance of real-time chatbot systems, where understanding previous contexts is crucial.

Furthermore, the study highlights the importance of considering the preceding sentence in real-time chatbot interactions for analyzing the emotions of the current sentence. Real-time sentiment analysis should incorporate the previous context, as the current speech is influenced by prior speech and contextual information.

The proposed model can be employed in various real-time systems, including counseling systems and AI speakers, which utilize interactive agent technology.

Nevertheless, there are certain challenges that warrant further investigation. The proposed model faces difficulties in fully capturing the characteristics of each short sentence, especially when compared to models that employ bidirectional training methods.

This limitation stems from the sequential training approach of KoGPT2, which operates in a single direction, unlike models like KoBERT. Future research could explore techniques such as bidirectional training to enhance the model's understanding of short sentences.

Another aspect for future work is analyzing conversations involving more than three speakers. Developing methods to recognize and consider the context of multiple speakers would significantly improve the model's capability to determine emotions in complex dialogue scenarios.

Expanding the dataset would enhance the model's generalizability and performance. The model can better evaluate and improve its performance by utilizing larger and more diverse datasets.

Additionally, exploring real-time sentiment analysis with temporal dynamics holds promise. Incorporating temporal information and tracking changes in sentiment over time would enable more dynamic and nuanced emotion analysis in ongoing conversations.

In conclusion, this study presents a comprehensive analysis of sentiment analysis in dialogue using a context-based modeling approach with KoGPT2. The model's ability to capture contextual information and its performance in emotion classification demonstrate its potential for enhancing real-time chatbot systems.

References

[1] R. Feldman, "Techniques and applications for sentiment analysis", Communications of the ACM, Vol. 56, No. 4, pp. 82-89, Apr. 2013. https://doi.org/10.1145/2436256.2436274.

[2] B. Liu, "Sentiment Analysis and Opinion Mining", Synthesis Lectures on Human Language Technologies, Vol. 5, No. 1, pp. 1-167, Apr. 2012.

[3] A. Bakliwal, J. Foster, J. Puil, R. O'Brien, L. Tounsi, and M. Hughes, "Sentiment analysis of political tweets: Towards an accurate classifier", Proc. of Computational Linguistics, Atlanta, Georgia, pp. 49-58, Jun. 2013.

[4] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep Contextualized Word Representations", Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, Vol. 1, pp. 2227-2237, Jun. 2018.

[5] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, "A survey on empathetic dialogue systems", Information Fusion, Vol. 64, pp. 50-70, Dec. 2020. https://doi.org/10.1016/j.inffus.2020.06.011.

[6] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges", Artificial Intelligence Review, Vol. 55, No. 7, pp. 5731-5780, Feb. 2022. https://doi.org/10.1007/s10462-022-10144-1.

[7] S. Ghosh, O. Vinyals, B. Strope, S. Roy, T. Dean, and L. Heck, "Contextual lstm (clstm) models for large scale nlp tasks", arXiv preprint, arXiv:1602.06291, Feb. 2016. https://doi.org/10.48550/arXiv.1602.06291.

[8] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations", Proc. of the AAAI Conference on Artificial Intelligence, Honolulu Hawaii USA, Vol.

33, No. 1, pp. 6818-6825, Jan. 2019. https://doi.org/10.1609/aaai.v 33i01.33016818.

[9] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation", arXiv preprint, arXiv:1908.11540, Aug. 2019. https://doi.org/10.48550/arXiv.1908.11540.

[10] D. W. Shin, Y. S. Lee, J. S. Jang, and H. C. Lim, "Emotion Classification in Dialogues Using Embedding Features", Human and Language Technology, pp. 109-114, Oct. 2015.

[11] M. K. Kim and H. S. Kim, "Integrated Dialogue Analysis using Long Short-Term Memory", Human and Language Technology, pp. 119-121, Oct. 2016.

[12] M. K. Kim and H. S. Kim, "Utterance Intention Analysis Using CNN-LSTM Neural Network", Korean Language Information Science Society, pp. 122-124, Oct. 2017.

[13] B. M. Kim and S. B. Park, "Model of Chatbot Answers by Language Style and Sentiment", Journal of KIISE, pp. 1480-1482, Jul. 2020.

[14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint, arXiv:1810.04805, Oct. 2018. https://doi.org/10.48550/arXiv.1810.04805.

[15] S. J. Lim and H. K. Kim, "Current Status of Deep Learning Pre-training for Natural Language Processing and Application to Korean", Journal of D-Culture Archives, Vol. 2, No. 2, pp. 111-118, Oct. 2019.

[16] C. H. Lee, D. Y. Lee, Y. A. Hur, K. S. Yang, and H. S. Lim, "Comparing Byte Pair Encoding Methods for Korean", Annual Conference on Human and Language Technology, pp. 291-295, Oct. 2018.

[17] A. Yessenalina, Y. Yue, and C. Cardie. "Multi-level structured models for document-level sentiment classification", In Proc. of the 2010

conference on empirical methods in natural language processing, Cambridge, MA, pp. 1046-1056, Oct. 2010.

[18] N. Farra, E. Challita, R. A. Assi, and H. Hajj, "Sentence-level and document-level sentiment mining for arabic texts", In 2010 IEEE international conference on data mining workshops, Sydney, NSW, Australia, pp. 1114-1119, Dec. 2010. https://doi.org/10.1109/ICDMW.2010.95.

[19] N. Engonopoulos, A. Lazaridou, G. Paliouras, and K. Chandrinos, "ELS: a word-level method for entity-level sentiment analysis", Proc. of the International Conference on Web Intelligence, Mining and Semantics WIMS '11, New York, United States, No. 12, pp. 1-9, May 2011. https://doi.org/10.1145/1988688.1988703.

[20] T. Brown, "What Consumers Think about brands on social media, and what businesses need to do about it", Report, Keep Social Honest, 2013.

[21] H. Zhou and F. Song, "Aspect-level sentiment analysis based on a generalized probabilistic topic and syntax model", In The Twenty-Eighth International Flairs Conference, Hollywood, Florida, USA, Apr. 2015.

[22] D. M. E. D. M. Hussein, "A survey on sentiment analysis challenges", Journal of King Saud University-Engineering Sciences, Vol. 30, No. 4, pp. 330-338, Oct. 2018.

[23] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach", In ICML, Jan. 2011

[24] W. Liu, W. L. Zheng, and B. L. Lu, "Emotion recognition using multimodal deep learning", In International conference on neural information processing, Springer, Cham. pp. 521-529, Oct. 2016.

[25] J. Xu, D. Chen, X. Qiu, and X. Huang, "Cached long short-term memory neural networks for document-level sentiment classification", arXiv preprint, arXiv:1610.04989, Oct. 2016. https://doi.org/10.48550/arXiv.1610.04989.

[26] H. T. Nguyen and M. L. Nguyen, "An ensemble method with sentiment features and clustering support", Neurocomputing, Vol. 370, pp. 155-165, Dec. 2019. https://doi.org/10.1016/j.neucom.2019.08.071.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners", OpenAI blog, Vol. 1, No. 8, Aug. 2019.

[28] C. Zhou, C. Sun, Z. Liu, and F. Lau, "A C-LSTM neural network for text classification", arXiv preprint, arXiv:1511.08630, Nov. 2015. https://doi.org/10.48550/arXiv.1511.08630.

[29] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification", In Proc. of the 2016 conference on empirical methods in natural language processing, Austin, Texas, pp. 606-615, Nov. 2016.

[30] J. M. Yoon and Y. J. Ko, "Speech-Act Analysis System Based on Dialogue Level RNN-CNN Effective on the Exposure Bias Problem", Journal of KIISE, pp. 911-917, Sep. 2018. https://doi.org/10.5626/JOK.2018.45.9.911.

[31] M. Y. Seo, T. S. Hong, J. A. Kim, Y. J. Ko, and J. Y. Seo, "Hierarchical attention-based CNN-RNN networks for The Korean Speech-Act Analysis", Human and Language Technology, pp. 243-246, Oct. 2018.

[32] A. Vaswani, et al., "Attention is all you need", In Advances in neural information processing systems, Dec. 2017.

[33] P. Ekman, et al., "Universals and cultural differences in the judgments of facial expressions of emotion", Journal of personality and social psychology, Vol. 53, No. 4, pp. 712-717, Oct. 1987. https://psycnet.apa.org/doi/10.1037/0022-3514.53.4.712.

[34] H. J. Park and K. S. Shin, "Aspect-Based Sentiment Analysis Using BERT: Developing Aspect Category Sentiment Classification Models", Journal of Intelligence and Information Systems,

Vol. 26, No. 4, pp. 1-25, Dec. 2020.

[35] S. Kim, Y. Song, C. Song, and J. Han, "A study on semantic ambiguity in the korean named entity recognition", In Annual Conference on Human and Language Technology, pp. 203-208, Oct. 2021.

Authors

Seon-Haeng Kim

2012 ~ 2019 : B.S. degree in Dept. of Business Administration, Hannam University
2019 ~ 2021 : M.S. degree in Dept. of Business Informatics, Hanyang University
2022 ~ present : Associate, Production AIX Task Team, LG Electronics
Research Interest : Natural Language Processing, Machine Learning, Noise/Vibration Anomaly Detection, and Artificial Intelligence

So-Young Jun

2016 ~ 2020 : B.S. degree in School of Business, Hanyang University
2020 ~ 2023 : M.S. degree in Dept. of Business Informatics, Hanyang University
2023 ~ present : Intern, Data Team, Storelink
Research Interest : Business Analytics, Natural Language Processing, Financial Engineering, and Artificial Intelligence for Business Applications

Jong-Woo Kim

1985 ~ 1989 : B.S. degree in Mathematics, Seoul National University
1989 ~ 1991 : M.S. degree in Dept. of Management Science, KAIST
1990 ~ 1995 : Ph.D. degree in Dept. of Industrial Management, KAIST
1995 ~ 1996: Post Doc., KAIST
1996 ~ 2003 : Associate Professor, Dept. of Information and Statistics, Chungnam National University
2003 ~ present : Professor, School of Business, Hanyang University
Research Interest : Intelligent Information Systems, Data Mining, Social Network Analysis, Text Mining, Business Machine Learning Applications, Collaborative Systems, and E-commerce Recommendation Systems