DOI: 10.6688/JISE.20220X_38(X).00XX

Designing the KoBERT-based Sentiment Classifier and Selective Llama2 model to improve the Performance of a Sentiment-based Chatbot System

HYEONJI KIM[1] AND YOOSOO OH[+2]

[1]Department of Information and Communication Engineering,
Daegu University, Gyeongsan-si, Republic of Korea
[2]School of AI, Daegu University, Gyeongsan-si, Republic of Korea

E-mail: hyunji.k.0410@gmail.com, yoosoo.oh@daegu.ac.kr

Chatbots with emotional elements analyze users' sentiments in daily life and derive appropriate emotional responses. This paper proposes a design of a KoBERT-based emotion classifier and selective Llama2 model to improve the performance of a sentiment-based chatbot system. The proposed system comprises a Sentiment Analysis Module, a Sentiment Category Selector, and a Selective Llama2 Chatbot Module. The KoBERT-based sentiment analysis module trains and fine-tunes a KoBERT pre-training model for analyzing sentiment from user's conversation sentences. The sentiment analysis category selector is a module that obtains the number of cases according to the sentiment analysis results and selects the sentiment category. The Llama2 chatbot module consists of a model trained by fine-tuning Llama2, a pre-trained model, according to the category results of the emotion category selector. The proposed system generates chatbot-like answers with emotional elements. This paper compares the performance of the proposed system with that of an existing chatbot that learns all the data to form a chatbot. We measure the performance using cosine similarity, which quantifies the similarity between the responses of the two systems, and BLEU accuracy, which measures the quality of the responses in terms of their similarity to human responses. Our system outperforms the existing chatbot system with a cosine similarity of 0.86 and a BLEU accuracy of 0.906.

Keywords: Deep Learning, Sentiment Analysis, Chatbot, KoBERT, Llama2

1. INTRODUCTION

Usually, human conversations include contextualized emotions. Conversely, chatbots need help keeping up with the flow of a conversation or generating consistent responses [1]. Verbal, emotional conversations are contextualized, which means that responses vary depending on the sentence. For example, a sentence like "I have had a busy day today" can be either a strenuous sentiment because you are too busy [strenuous] or a joyful sentiment because you have much work coming in [joyful], depending on the context. In a sentence like "I did not get the job interview today," you might feel sad because you did not get the job interview or angry because you did not do well in the interview. Depending on the situation, the same sentence can express different emotions in these examples. Since existing chatbots learn the entire sentence data related to emotions, there are cases where chatbot answers are derived without considering the sentiments of the situation.

Therefore, we propose a design of a KoBERT-based sentiment classifier and selective

Received; revised; accepted.
Communicated by Editor.

[+] Corresponding author: Yoosoo Oh

Llama2 model to improve the performance of a sentiment-based chatbot system. The proposed system analyzes the sentiment of user sentences by fine-tuning the KoBERT pretraining model and derives the probability of each sentiment. KoBERT is a model that improves the Korean performance of BERT, which SKT Brain developed [2]. The proposed system obtains the number of cases of emotion categories for contextual emotion analysis and builds a chatbot model according to the number of cases. The chatbot model is based on the number of emotion cases learned by fine-tuning Llama2. Llama2 is a large-scale AI language model released by Meta [3]. We selected a Llama2 chatbot model designed based on the results of emotion analysis because of improving chatbot answers. We also verified the performance of the proposed system by analyzing the accuracy of existing emotion chatbots and the proposed system. As a result, our system outperforms the existing chatbot in eliciting emotional responses based on context.

2. RELATED RESEARCH

The BERT Model is a bidirectional encoder language model that applies transformers [2] [4]. The BERT Model is trained by masking a part of the randomly input tokens and then predicting the correct answer of the masked tokens [4]. The BERT Model can also be trained using Next Sentence Prediction (NSP) to understand the context [2]. The KoBERT is a model trained on large-scale corpora collected through Wikipedia and news and applying data-based tokenization techniques [2].

Wotaifi TA et al. proposed a hybrid model for fake news prediction that fuses CNN for identifying text characteristics, BiGRU for identifying text order, and machine learning based on dictionary learning models (Glove, FastText, BERT) to reflect word representation [5]. They assigned embedding values to words using Glove, FastText, and BERT to represent the semantic meaning of words. They improved the random forest with a feature selection method based on the assigned embedding values. They used three test data sets (Fake-or-Real, AraNews, and Sentimental LAIR datasets), yielding accuracies as high as 0.9935, 0.9473, and 0.7481, respectively [5].

The Llama2 is a large-scale AI language model published by Meta, trained on models with up to 65 billion parameters to improve the transformer architecture [3]. The Llama2 is trained with 2 trillion tokens, 40% more than the original Llama and 2x the context length. The Llama2 utilizes human feedback-based reinforcement learning (RLHF) and ghost attentions for fine-tuning, allowing it to maintain the overall flow of the conversation even as it moves on.

SoYeop Yoo et al. proposed PolarisX-bot, a chatbot system that builds a self-expanding knowledge graph after collecting real-time data and utilizes the collected data [6]. The PolarisX-bot builds a self-expanding knowledge graph using a BERT model that can identify relationships between words to extract new words [6]. The PolarisX-bot is a chatbot for the self-expanding knowledge graph using Google's DialogFlow service. The PolarisX-bot presents multiple answers to the user based on the self-expanding knowledge graph built by the user in a graph-like visualization, making the results intuitive [6]. They showed that the accuracy of the BERT-based self-expanding knowledge graph was 0.7528 for train accuracy and 0.7885 for test accuracy [6].

Wonmin Lee et al. proposed a sentence generation model that analyzes emotions through a fine-tuned BERT and trains a fine-tuned GPT model to reflect emotional information by pairing emotion-label values in the dialog. They showed that the BERT and GPT pipeline generation models can generate sentences by predicting the next word based on the emotion labels [7]. Their proposed system achieved a 12% accuracy improvement over the baseline method [7].

We found that sentiment analysis chatbots are trained using only all emotional data by analyzing existing works. They are trained to pair the sentiment analysis results with previous sentences to preserve the emotional information. To improve the analysis of sentiment aspects for chatbot systems, we build an integrated model by combining Llama2, which shows high performance in identifying the context of conversations, and the KoBERT model trained with Korean data.

3. PROPOSED CHATBOT

In this paper, we design a KoBERT-based sentiment classifier and a selective Llama2 model to improve the performance of sentiment analysis chatbot systems. Figure 1 shows the proposed model's overall behavioral diagram.

The proposed system comprises a sentiment analysis module, a sentiment category selector, and a selective Llama2 chatbot module. The emotion analysis module analyzes the user input dialog sentences for three emotions (sadness, joy, and anger) and derives probability values. The emotion category selector is responsible for identifying the emotions in the sentence and selecting the final emotion based on the probability values derived from the emotion module. The selective Llama2 chatbot module selects a Llama2 chatbot model trained for each emotion based on the final emotion derived from the emotion category selector. It derives chatbot-type answers from the model chosen.

The KoBERT-based sentiment analysis module analyzes the emotions in users' verbal sentences. The proposed system is trained using human sentence-1, emotion_classification, from emotional conversation corpus data provided by AI_Hub [8]. The composition of the data used is shown in Table 1. In this paper, we used six emotion categories (sadness, joy, anger, hurt, sadness, anxiety, embarrassment) and trained using 7000 data points of the three most representative emotions (sadness, joy, and anger). We trained our proposed model by fine-tuning the KoBERT pre-trained model, and we established the KoBERT-based sentiment analysis module to analyze the emotions among users' conversations when the users input the conversation contents to the chatbot. Our system derives predicted probability values for the three emotions used for training: sadness, joy, and anger. The predicted emotion probability values of the emotion analysis module are shown in Table 2.

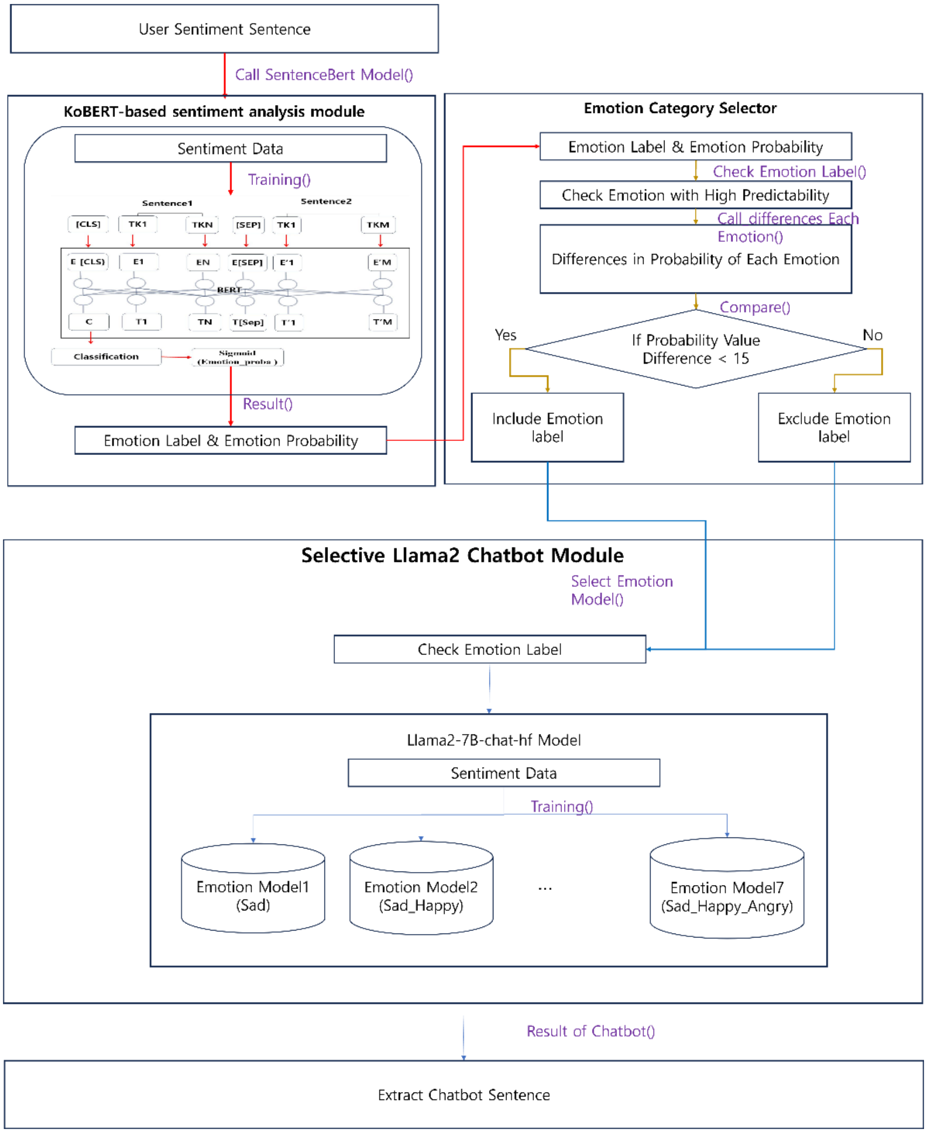

Fig. 1. Diagram of the KoBERT-based sentiment classifier and selective Llama2 model for improving the performance of sentiment analysis chatbot systems.

Table 1. *Organize the data you use*

	Human Sentence-1	System Response1	Sentiment_Category
1	I think it's because of something I said that the kids at school avoid me.	You are experiencing bullying at school.	Sadness
2	You want to throw a surprise party for your best friend's birthday this week.	You want to throw a big surprise party for your best friend's birthday.	Joy
3	I felt insulted by a friend who made a comment about my future plans.	You must be very upset that this happened.	Anger
4	I'm working on a difficult project and I feel so sad that no one supports me.	You're sad because no one supports you.	Sadness
5	I'm angry that people in the office seem to be ignoring me on purpose.	The people in the office must be making you feel bad	Anger

Table 2. *Sentiment Analysis Module Results*

	Human Sentence 1	Predicted Sentiment Probability [sadness, joy, anger]	Predictive Sentiment
1	My heart sank when I heard that my friend was assaulted for standing up for me!	[10.12551, -4.98441, -4.325249]	Sadness_Angry
2	Today my teacher apologized to me for not knowing I was being bullied.	[8.780641,-1.4066051, -6.2898874]	Joy_Sadness

3	A few days ago, I was being beaten up by a kid from another class and my class president came and stopped it.	[-5.4003463,10.35224, -3.6633172]	Joy_Angry
4	I had a deep talk with a friend today.	[-4.345507,10.8411875, -4.700371]	Joy
5	I feel sad because I feel stupid for letting it go even when I feel bad.	[10.155785,-5.164654, -4.1531515]	Sadness_Angry

The sentiment category selector generates the number of cases, excluding non-response cases, based on the sentiment category used. The number of cases generated is 7: sad, joy, anger, sad_happy, sad_angry, joy_angry, sad_happy_angry, sad_happy_angry. In this paper, we define the seven emotions generated by the number of cases as emotion categories. The emotion category selector uses the predicted emotion probability values derived from the KoBERT-based emotion analysis module to find the difference between each predicted emotion probability value. The proposed system finds the difference between each probability value based on the emotion category with the most significant predicted probability value. For example, if sadness has the most significant predicted probability, the proposed system finds the difference between the predicted probability values of happy_sad and sad_angry.

To set the sentiment prediction probability threshold, we calculated the median value of each sentiment probability in the test data. As a result, we calculate the final sentiment based on 15, which is the median value of the sentiment prediction probability of the test data. Suppose the difference between the prediction probabilities of happy_sad and sad_angry is 15 or less, and the difference between the prediction probabilities of sad_sad and sad_angry is 15 or more. In that case, the system recognizes that the emotion of the predicted sentence is not only sad but also joy and derives happy_sad, excluding the emotion of anger. Table 3 shows the algorithmic pseudocode of the emotion category selector.

Table 3. Sentiment Category Selector Algorithm Pseudocode

Sentiment Category Selector Algorithm Pseudocode

```
Declare an emotion_result list to store the results.
Set the emotion threshold to 15.
Call the Sentiment Category Selector function.
If the maximum probability value corresponds to the "sadness" class:
   If sadness_anger is below the emotion threshold or joy_sadness is below the emotion threshold:
  If both sadness_anger and joy_sadness are below the emotion threshold:
    Then add the "all" emotion class to the emotion_result list.
      Else if sadness_anger is below the emotion threshold and joy_sadness is above the emotion threshold:
    Then add the "sadness_anger" emotion class to the emotion_result list.
      Else if sadness_anger is above the emotion threshold and joy_sadness is below the emotion threshold:
    Then add the "joy_sadness" emotion class to the emotion_result list.
  Else:
    Then add the "sadness" emotion class to the emotion_result list.
If the maximum probability value corresponds to the "joy" class:
   If joy_anger is below the emotion threshold or joy_sadness is below the emotion threshold:
  If both joy_anger and joy_sadness are below the emotion threshold:
    Then add the "all" emotion class to the emotion_result list.
      Else if joy_anger is below the emotion threshold and joy_sadness is above the emotion threshold:
    Then add the "joy_anger" emotion class to the emotion_result list.
      Else if joy_anger is above the emotion threshold and joy_sadness is below the emotion threshold:
    Then add the "joy_sadness" emotion class to the emotion_result list.
Else:
  Then add the "joy" emotion class to the emotion_result list.
If the maximum probability value corresponds to the "anger" class:
   If joy_anger is below the emotion threshold or sadness_anger is below the emotion threshold:
  If both joy_anger and sadness_anger are below the emotion threshold:
    Then add the "all" emotion class to the emotion_result list.
      Else if joy_anger is below the emotion threshold and sadness_anger is above the emotion threshold:
    Then add the "joy_anger" emotion class to the emotion_result list.
      Else if joy_anger is above the emotion threshold and sadness_anger is below the emotion threshold:
    Then add the "sadness_anger" emotion class to the emotion_result list.
Else:
  Then add the "anger" emotion class to the emotion_result list.
```

The predictive sentiment analysis results of the proposed system are shown in Table 2. The selective Llama2 chatbot module generates emotional responses appropriate to the user's conversation. The selective Llama2 chatbot module uses human sentence-1 and system response-1 from the emotional conversation corpus data collected, as shown in Table 1. The selective Llama2 chatbot module organizes the emotion data according to the emotion category selector's built-in emotion category and trains the Llama2-chat-7B model with fine-tuning and 4-bit quantization for model lightweight. The Llama2 is a large-scale language model that improves GPT's transformer model [3]. The Llama2 has models according to model size (7B, 13B, 70B), and the Llama2-chat-7B shows high accuracy for models that grasp the flow of conversations, such as chatbots through learning as human feedback-based reinforcement learning [3]. The proposed system selects each model corresponding to the emotion derived from the emotion category selector among each chatbot model trained with Llama2-chat-7B per emotion category. The proposed system uses the selected model to generate emotional chatbot responses.

4. EXPERIMENTS

This paper evaluates the proposed system's performance improvement using learning accuracy, prediction accuracy, and loss as the main metrics. Additionally, the performance of the sentiment analysis module and the selective Lama2 chatbot model is assessed using learning accuracy, recall, and loss. We used the accuracy formula in Equation 1 to check its performance. In Equation 1, y is the correct answer value, y' is the predicted value, and N is the total data [9].

$$Accuracy = \frac{1}{N}\sum_{i=0}^{n-1} 1(y_i^{'} = y_i) \tag{1}$$

The experimental results presented that the proposed KoBERT-based sentiment analysis module had a learning accuracy of 1.0, a loss value of about 0.001, and a prediction accuracy of about 84%. Figure 2 shows a graph of the learning/prediction accuracy on the left and the loss value for learning on the right.

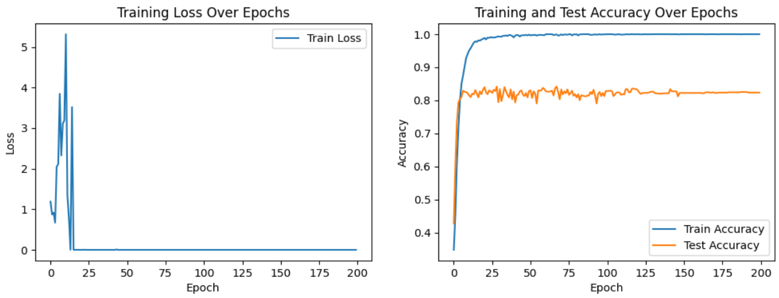

Fig. 2. Precision graph for the KoBERT-based sentiment analysis module.

We measured and compared the chatbot system accuracy of existing chatbot systems (which learn the data as a whole), the KoGPT model widely used for sentence generation, and the proposed system (a selective Llama2 chatbot module).

Table 4 shows the results of chatbot answers using the proposed selective Llama2 chatbot model. This paper compares the correct answers in Table 4 with the answers generated by the proposed chatbot system. As a result, we can see that the correct answer of No. 1 is ['Surprised'], and the chatbot answer of the proposed system is ['I thought my heart was stopping']. The correct answer to No. 2 is ['You apologized'], and the chatbot answer of the proposed system is ['You apologized']

Table 4. Chatbot answer results of the proposed system

	Human Sentence 1	Correct answer	Chatbots in your proposed system Answer
1	My friend created your heart sinks when you hear that you were assaulted while out and about.	You're very upset to hear that your friend was assaulted.	Your heart stopped when you heard that your friend had been assaulted because you stood up for him.
2	Today my teacher told me that I was I didn't realize I was being bullied and apologized.	You apologized and said you didn't know you were being harassed.	You've apologized for not realizing you were being harassed.
3	My friend is a kid I'm feeling pretty grim after getting beaten up.	I'm sure you don't feel comfortable that your close friend got hit.	Your friend just got beat up by some kid, and you're feeling pretty down.

4	I had a deep conversation with a friend today.	Friends and heartfelt We can talk. That must have been nice.	Friends with Benefits Share your chi today anything good happening?
5	If you feel bad, just use the I'm an idiot for moving on Sad because it's the same.	The fact that it was overlooked Sad because I feel stupid, You're up.	You feel bad, you're sad

Therefore, Table 4 shows that the chatbot answers generated by the proposed model are similar to the correct answers. This paper numerically verifies the similarity between the chatbot answer generated by the proposed model and the correct answer. In addition, this paper verifies the performance of the proposed system chatbot compared to existing chatbot systems [KoGPT, Llama2]. This paper uses cosine similarity and BLEU accuracy formulas to measure similarity [10]. The closer the cosine similarity approaches 1, the higher the similarity [10] [12]. The cosine similarity formula used is shown in Equation 2. In Equation 2, A and B are the values of the vectorized words.

$$Cosine\ Similarity = \frac{A*B}{||A||\,||B||} \tag{2}$$

In this paper, the accuracy is derived from the sum of similarities between the words used in a sentence, and the words in the sentence are vectorized through Countervector embedding, which is one of the embedding models [13]. Table 5 shows the cosine similarity of each model.

Table 5. Cosine similarity values for chatbot systems

Chatbot System Name	Cosine Similarity Value
KoGPT(Based_System)	0.73
Llama2(Based_System)	0.81
Selective Llama2 chatbot module (Proposed_System)	0.86

Figure 3 is a graphical representation of the cosine similarity results of the proposed and existing chatbot models [KoGPT, Llama2]. Figure 3 represents the cosine similarity results of KoGPT, Llama2, and the Proposed System (the selective Llama2 chatbot module) from left to right.

In the experimental results, the proposed system achieved the highest cosine similarity of 0.86 compared to the existing chatbot systems. We can also see from the cosine similarity graph that the proposed chatbot system produces a more distinct graph.

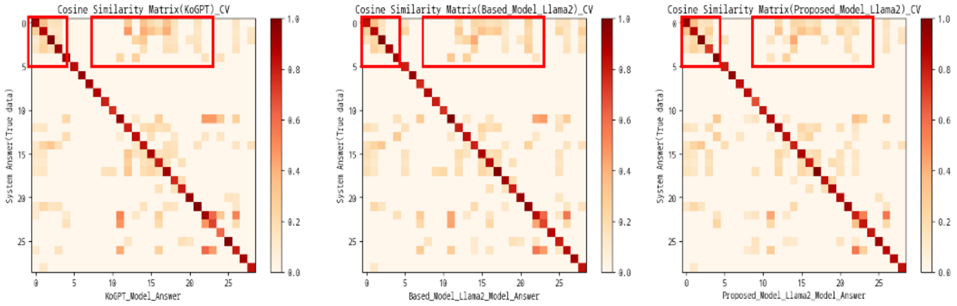

Fig. 3. Cosine similarity result graph (From the left: KoGPT, Llama2, Proposed System)

We compared the accuracy using the BLEU value, which is widely used to measure chatbot accuracy. The BLEU is measured using the n-gram method, such as how many words of the correct sentence are included in the words of the predicted sentence derived through the chatbot system [11]. The formula for BLEU used is shown in Equation 3. Pn is the calibrated precision of the gram, N is the maximum number of n in the n-gram, Wn is the weight, and BP is the penalty for sentence length [11]. The BLEU used is shown in Equation 3. Table 6 shows the BLEU results between chatbot systems. Table 6 shows the BLEU values as a percentage.

$$BLEU = BP * \exp(\sum_{n=1}^{N} W_n log P_n) \quad (3)$$

Table 6. BLEU values for chatbot systems

Chatbot System name	BLEU Value
KoGPT(Based_System)	0.887
Llama2(Based_System)	0.899
Selective Llama2 chatbot module (Proposed_System)	0.906

Figure 4 shows graphs of the BLEU results for the 30-test data, from left to right, for KoGPT, Llama2, and the Proposed System (the selective Llama2 chatbot module).

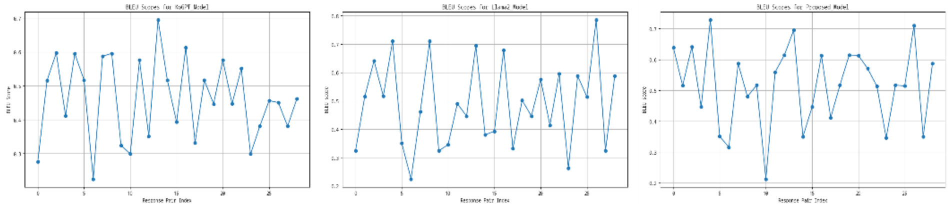

Fig. 4. BLEU result graph (From the left: KoGPT, Llama2, Proposed System)

The experimental results showed that the BLEU value of the proposed system was 0.906 compared to the existing chatbot, resulting in high accuracy.

5. CONCLUSIONS

In conclusion, we proposed a KoBERT-based sentiment classifier and selective Llama2 model design to improve the sentiment analysis of the chatbot system's performance. The proposed system achieves high performance compared to existing chatbot systems. We achieved higher performance by generating chatbot models for each emotion category. However, we needed a lot of computer memory due to model segmentation. Large memory consumption can reduce the performance of the computer. To overcome this drawback, we will build a model segmentation with less memory in future research. Due to the limitation of computer resources, only three categories of emotion_category, sadness, happiness, and anger, have been trained. In future research, we will use all six emotions in the emotion_big category to learn and build an optimized model. In addition, we analyzed emotions by considering only one emotion per sentence without considering the order of emotions. We will also consider the order of emotions when multiple sentences are input. Our emotional chatbot system can be used in a care system because it can respond to emotional factors. Therefore, we can apply our chatbot system to IoT or embedded care systems.

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2022S1A5C2A07091326)

REFERENCES

1. Seunguook Lim, & Jihie Kim (2021). RNN model for Emotion Recognition in Dialogue by incorporating the Attention on the Other's State. Journal of KIISE, 48(7), 802-808, 10.5626/JOK.2021.48.7.802
2. Lee, Sangah, et al. "Kr-bert: A small-scale korean-specific language model." arXiv preprint arXiv:2008.03979 (2020).

3. Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).
4. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
5. Wotaifi TA, Dhannoon BN (2023) Developed Models Based on Transfer Learning for Improving Fake News Predictions. JUCS - Journal of Universal Computer Science 29(5): 491-507.
6. Yoo, SoYeop, and OkRan Jeong. "Auto-growing knowledge graph-based intelligent chatbot using BERT." ICIC Express Lett
7. Lee, Won-Min, and Byung-Won On. "Generating Emotional Sentences Through Sentiment and Emotion Word Masking-based BERT and GPT Pipeline Method." The Journal of Korean Institute of Information Technology 19.9 (2021). pp: 29-40.
8. AI_Hub, https://www.aihub.or.kr/
9. Xu, Canwen, et al. "Beyond preserved accuracy: Evaluating loyalty and robustness of BERT compression." arXiv preprint arXiv:2109.03228 (2021).
10. Xia, Peipei, Li Zhang, and Fanzhang Li. "Learning similarity with cosine similarity ensemble." Information sciences 307 (2015): 39-52.
11. Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
12. Hammad MM, Al-Smadi M, Baker QB, D MA-a, Al-Khdour N, Younes MB, Khwaileh E (2020) Question to Question Similarity Analysis Using Morphological, Syntactic, Semantic, and Lexical Features. JUCS - Journal of Universal Computer Science 26(6): 671-697. https://doi.org/10.3897/jucs.2020.036
13. Vu, Dieu, et al. "Revisiting Supervised Word Embeddings." *J. Inf. Sci. Eng.* 38.2 (2022): 413-427.

Hyeonji Kim received her Bachelor's degree in Department of AI Software, School of AI from Daegu University in 2022. Currently studying master course at the Department of Information and Communication Engineering at Daegu University

Yoosoo Oh (范晉維) received his Bachelor's degree in the Department of Electronics and Engineering from Kyungpook National University in 2002. He obtained his Master's degree in the Department of Information and Communications from Gwangju Institute of Science and Technology (GIST) in 2003. In 2010, he received his Ph.D. degree in the School of Information and Mechatronics from GIST. In September 2012, he joined Daegu University, where he is currently an associate professor in the School of AI, Daegu University. From 2017~2019, he worked as a center director of the Mixed Reality Convergence Research Center at Daegu University. From 2015-2017, He worked as a director in the Enterprise Supporting Office of LINC Project Group, Daegu University. Currently working as a center director of DU Smart Drone Center, AZIT MakerSpace Center, and Gyeongbuk Technopark Daegu University Center. His research interest is Machine Learning and Deep Learning for Edutech and Industry applications, Intelligent Middleware.