check for updates

ARTICLE

A Study of BERT-Based Classification Performance of Text-Based Health Counseling Data

Yeol Woo Sung, Dae Seung Park and Cheong Ghil Kim[*]

Namseoul University, Cheonan, Chungcheongnam-do, Korea
*Corresponding Author: Cheong Ghil Kim. Email: cgkim@nsu.ac.kr

ABSTRACT

The entry into a hyper-connected society increases the generalization of communication using SNS. Therefore, research to analyze big data accumulated in SNS and extract meaningful information is being conducted in various fields. In particular, with the recent development of Deep Learning, the performance is rapidly improving by applying it to the field of Natural Language Processing, which is a language understanding technology to obtain accurate contextual information. In this paper, when a chatbot system is applied to the healthcare domain for counseling about diseases, the performance of NLP integrated with machine learning for the accurate classification of medical subjects from text-based health counseling data becomes important. Among the various algorithms, the performance of Bidirectional Encoder Representations from Transformers was compared with other algorithms of CNN, RNN, LSTM, and GRU. For this purpose, the health counseling data of Naver Q&A service were crawled as a dataset. KoBERT was used to classify medical subjects according to symptoms and the accuracy of classification results was measured. The simulation results show that KoBERT model performed high performance by more than 5% and close to 18% as large as the smallest.

KEYWORDS

BERT; NLP; deep learning; healthcare; machine learning

1 Introduction

With the development of high-speed wireless communication and the spread of various mobile devices, the Internet is overflowing with various opinions and information of individuals especially using SNS [1,2]. As a result, online data has rapidly increased and is being used in various fields to extract meaningful information from such accumulated unstructured data, in which SA (Sentiment Analysis) and Chatbot services using NLP (Natural Language Processing) are representative. In addition, people analyze product sale, service strategies, and lyrical trends by exploring subjective sentiment information in articles and reviews [3–6]. Shanmuganathan et al. [3] proposed a machine learning methodology to detect the flu (influenza) virus spreading among people mainly across Asia. Zeng et al. [4] proposed a mixed-model account of CNN-BiLSTM-TE (Convolutional Neural Network, Bidirectional Long-Short-Term Memory, and Subject Extraction) model to solve the problems of low precision, insufficient feature extraction, and poor contextual ability of existing

text sentiment analysis methods. Heo et al. [5], presented an approach for detection of adverse drug reactions from drug reviews to compensate limitations of the spontaneous adverse drug reactions reporting system. Salminen et al. [6] introduced a cross-platform online hate classifier.

The situation where face-to-face contact is restricted due to the prolonged COVID-19 pandemic is also being applied to healthcare fields. Under this circumstance, an area attracting attention is non-face-to-face healthcare services such as teleconsultation, telemedicine, and remote monitoring [7,8]. Nasr et al. [7] mentioned the need for innovative models to replace the traditional health care system as the number of individuals with chronic diseases increases significantly, and the models further evolved into the smart health care system of the future including hospitals, nursing homes and long-term care centers. It addressed the need to provide more personalized health care services and less reliance on traditional offline health care institutions such as hospitals, nursing homes, and long-term healthcare centers. In particular, Rahaman et al. [8] introduced IoT-based health monitoring systems as the most important healthcare application field as IoT is an important factor that changes technological infrastructure. Here, it is possible to reduce contact for health promotion and disease prevention such as counseling and management in the pre-hospital stage and healthcare in situations where face-to-face treatment is difficult [9,10]. Miner et al. [9] introduced the several useful aspects of Chatbots in the fight against the COVID-19 pandemic together with challenges in information dissemination. Jovanović et al. [10] tried to provide the goals of healthcare chatbots by characterizing in service provisioning and highlighting design aspects that require the community's attention, emphasizing the human-AI interaction aspects and the transparency in AI automation and decision making. As for those services AI is applied in the form of digital healthcare including chatbots [11–13]. Wu et al. [11] summarized the latest developments related to telemedicine and discussed the obstacles and challenges to its wide adoption with a focus on the impact of COVID-19. Kandpal et al. [12] introduced contextual based chatbots using Machine Learning and Artificial Intelligence techniques to store and process the training models which help the chatbot to give a better and appropriate response when the user asks domain specific questions to the bot. Softić et al. [13] presented the health chatbot application created on the Chatfuel platform which can identify users' symptoms through a series of queries and guides them to decide whether to go to doctor. Digital healthcare enables prevention, diagnosis, treatment, and follow-up management anytime, anywhere, which is usually a combination of digital technology, smart technology, and health management technology. Table 1 shows the use of AI in the healthcare industry divided by Clinical and Non-clinical area [14]. Chebrolu et al. [14] introduced how health care organizations can scale up their AI investments by pairing with a robust security and data governance strategy. AI is being used in symptom analysis, scientific discovery, and risk management in the clinical field. In the non-clinical field, it is being used in management tasks automation, fraud and misuse detection and prevention, and artificial intelligence-based counselling.

As for chatbot, it is a freely communicating software with humans using NLP that provides an appropriate answer to a user's question or various related information through voice or text conversation, and generally uses a chat interface on the web or messenger [15]. Kim et al. [15] examined existing attempts to utilize Information Technology (IT) in the field of counseling and psychotherapy, and current trends in recent overseas cases of the application of Artificial Intelligence (AI) technology in the field of chatbots. Even though chatbots can perform many tasks, the main function is to understand human speech and respond appropriately. The deep learning-based NLP AI engine, which has been continuously developing in recent years [12], is being applied in a way that enhances this function. In the healthcare field, In the healthcare field, AI chatbot collects user data through symptom related conversations in advance, so that it can be used as basic data for the user during face-to-face

consultation or treatment with a doctor. In particular, the deep learning-based dialogue engine helps to accurately recognize the patient's conversations with various expressions according to the context.

Table 1: Applications of AI in healthcare

	Clinical		Non-clinical
Efficiency	Symptom analysis for COVID-19	Efficiency	Automating administrative tasks
	Accelerating scientific discovery processes		Accelerating and amplifying insights generation
	Smart workforce management – Planning and scheduling – Talent crowdsourcing		Fraud, waste, and abuse (FWA) detection and prevention
Risk management	Medication-related error detection	Customer experience	Conversational AI

This paper verifies the performance of NLP algorithms based on deep learning, which is necessary to accurately determine the treatment subject when consulting a user for a disease when operating a consultation healthcare system using a chatbot. Among the performance of Bidirectional Encoder Representations from Transformers (BERT) was compared with other algorithms of CNN, RNN, LSTM, and GRU. For this purpose, we crawled the health counseling data of Naver Q&A service as dataset. A Korean BERT model, KoBERT, was used to classify medical subjects according to symptoms and the accuracy of classification results was measured.

The rest of this paper is structured as follows. Section 2 reviews the basic of CNN, RNN, LSTM and GRU; Section 3 overviews BERT; Section 4 introduces dataset and implementation; Section 5 includes simulation results. Finally, Section 6 concludes this paper.

2 Background

There are two approaches to developing a chatbot depending on the algorithms and the techniques adopted: pattern matching and machine learning approaches. This section reviews four machine learning algorithms for dialog modeling. They are all representative deep learning algorithms for NLP with a well-known reputation in time series data analysis and context identification [16].

2.1 CNN

CNN (Convolutional Neural Network) is a model that extracts various features of data using filters. It is mainly used to find patterns for image recognition. As there are filters as many as the number of channels of data, image features are extracted while moving the filters. After applying the image to the filter, we use pooling to resize it to emphasize the features. Recently, research using CNN has been conducted in the field of NLP, and it is showing effective results. Fig. 1 shows the architecture of the CNN model [17].
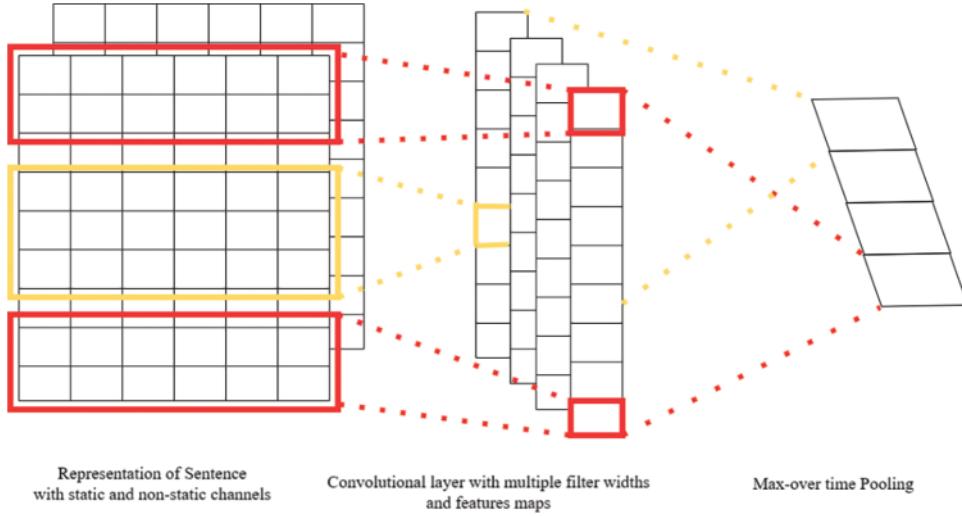
Representation of Sentence
with static and non-static channels

Convolutional layer with multiple filter widths
and features maps

Max-over time Pooling

Figure 1: CNN architecture

2.2 RNN

Recurrent Neural Network (RNN) classifies text by finding patterns in ordered data. As the previous input information is accumulated, the current information is expressed. Fig. 2 [18] depicts the structure of RNN with its own recurrent weight (W), which reflects the time series characteristics of predicting future information based on present information through past information through W while recognizing the pattern [19].
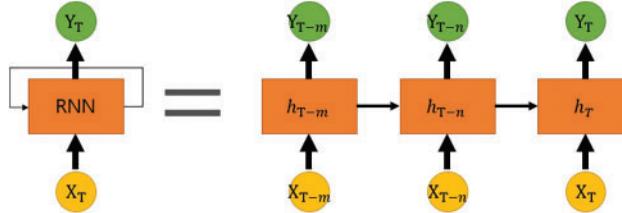

Figure 2: RNN architecture

The equations required for processing RNN are shown as (1) and (2).

$$h_T = tanh\,(W_{hh}h_{T-1} + W_{xh}x_T) \tag{1}$$

$$Y_T = W_{hY}h_T \tag{2}$$

In the above equations, the input value X_T affects the next output value h_{T+1} while outputting the result value h_T. Eq. (2) expresses the exaggeration that the previous data affects the current data as an equation. In RNN, the h_T value, called the state value, is a state value representing the present time based on time. h_{T-1} represents the previous ecological value. This model always refers to the value of h_{T-1} to calculate the value of h_T. In other words, Y_T is a value obtained by multiplying the value of h_T by the weight W_{hY}.

2.3 LSTM

LSTM (Long Short Term Memory) is a structure in which cell-state is added to RNN. In RNN, when the time interval between two points is large, the learning ability is greatly reduced. This limitation is because Back Propagation process does not convey enough information. The reason is that the weight is lost as it passes through numerous layers. LSTM was introduced to solve this problem. Using LSTM, information can be effectively transmitted without losing even at many intervals as long as certain information does not adversely affect the gradient [20].

Fig. 3 shows the structure of the LSTM which introduces Cell State and serves to deliver previously inputted information. It consists of Forget Gate which determines whether to reflect cell state among past information, Input Gate for how much input information will be reflected, and Output Gate which determines how much Cell State will be reflected in the State value to be transmitted to the next cell [20].
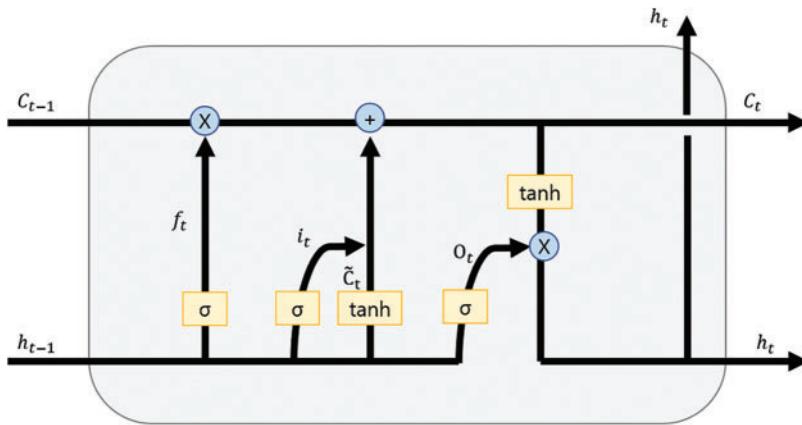

Figure 3: LSTM architecture

Forget Gate plays a role in deleting data deemed unnecessary among the transferred information. The equation of Forget Gate is as (3) [20].

$$f_t = \sigma \left(W_f \cdot [h_{t-1}, x_t] + b_f \right) \tag{3}$$

Input Gate determines how much of the current input value is reflected in Cell State. A lot of high-value data is reflected in cell state, and if not, the reflection is minimized. The equations are (4)–(6) [20].

$$i_t = \sigma \left(W_i \cdot [h_{t-1}, \ x_t] + b_i \right) \tag{4}$$

$$\tilde{C}_t = tanh \left(W_c \cdot [h_{t-1}, \ x_t] + b_C \right) \tag{5}$$

$$C_t = f_t * \ C_{t-1} + i_t * \ \tilde{C}_t \tag{6}$$

Output Gate is the gate that determines whether to transfer the final Cell State value to Hidden State. Finally, the h_t value is obtained and transferred to the next cell. The equation of the output gate is (7), (8) [20].

$$o_t = \sigma \left(W_o \cdot [h_{t-1}, \ x_t] + b_o \right) \tag{7}$$

$$h_t = o_t * \ tanh \left(C_t \right) \tag{8}$$

2.4 GRU

Gated Recurrent Units (GRUs) are inspired by LSTM. It is a structure that efficiently processes the existing gate by reducing it to two. Although it is used as a hidden state method, it effectively solves the Long-Term Dependency problem [21].

Fig. 4 shows the structure of the GRU. It consists of Reset Gate and Update Gate. The Reset Gate determines how the new input is merged with the old memory. Update Gate determines how much of the previous memory to remember. The equations are (9)–(12) [21].

$$r_t = \sigma \left(W_r \cdot [h_{t-1}, \ x_t] \right) \tag{9}$$

$$z_t = \sigma \left(W_z \cdot [h_{t-1}, \ x_t] \right) \tag{10}$$

$$\tilde{h}_t = \tanh \left(W \cdot [r_t * h_{t-1}, \ x_t] \right) \tag{11}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{12}$$

z_t receives the previous hidden state and input and determines how much value to reflect in the previous hidden state. After receiving the previous hidden state and input, r_t processes it as a sigmoid function. $\tilde{h}_t$ represents the current cell state, and h_t determines how much to delete the previous hidden state and reflect the current state and becomes the hidden state of the next step [21].
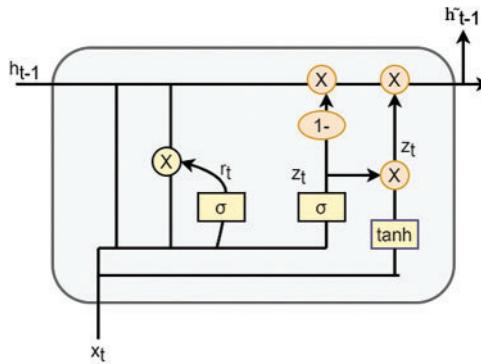

Figure 4: GRU architecture

3 BERT

The BERT (Bidirectional Encoder Representations from Transformers) model [22] shown in Fig. 5, one of the deep learning models, encodes the context of the input data in both directions based on the Transformer module [23], unlike the existing language representation models reviewed in Section 2. In this way, the BERT model provides a universal numerical representation of natural language that can be applied to many tasks related to natural language processing.

As is well known, deep learning models have achieved great results in a short time in the computer vision field for object detection and segmentation. This is because the model pre-trained with ImageNet extracted general-purpose feature vectors for images. If fine-tuning is performed with data on a given task based on this pre-trained model, good results can be obtained with a relatively small amount of data. Since the BERT model learns a feature vector that reflects the surrounding semantic information from a large amount of data, it can be applied to various natural language processing tasks using this feature vector.
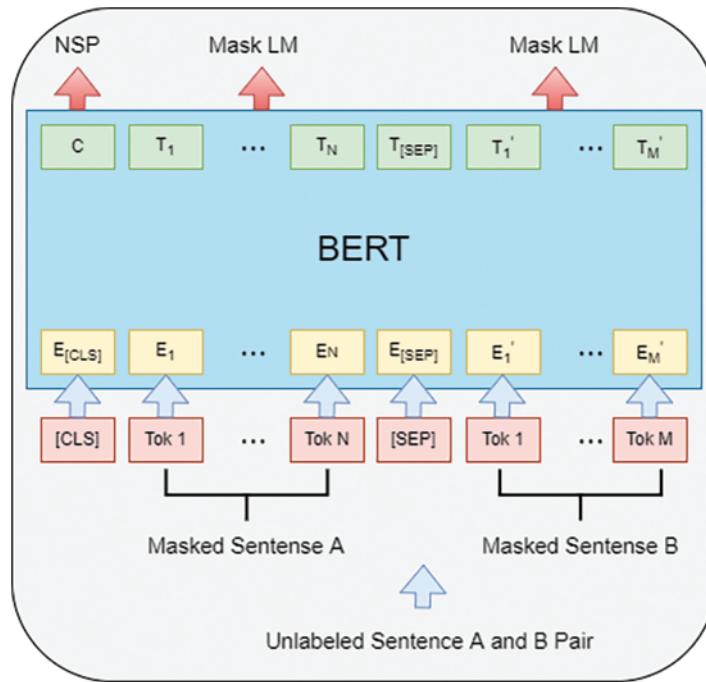
Figure 5: BERT pre-training

As for the pre-training, there are two methods. One is the method of predicting a masked word by masking some of the words in a sentence, the other is the method of predicting whether two sentences are a continuous sentence when given. Masked LM randomly masks the input value and predicts the masked token through the context of the surrounding words. At this time, randomly change to [Mask] token at a rate of 15% in each sentence. 80% of these are converted to [Mask] tokens. 10% change the token to a random word, and 10% leave the token as the original word. [Mask] token is used only for pre-training, not for fine-tuning [22]. In NSP (Next Sentence Prediction), QA (Question Answering) or NLI (Natural Language Inference), it is important to understand the relationship between two sentences. BERT predicts whether two sentences are continuous or not [22].

Pre-training using a large-scale corpus requires a lot of resources and time for learning, so people use the pre-trained model published by Google and fine-tune it. The BERT model that has been trained in advance using a Korean dataset of SK T-Brain's KoBERT [24] is emerging.

4 Dataset and Implementation

In this section, we introduce a proposed research framework to compare the performance of NLP integrated with machine learning for the accurate classification of medical subjects on text-based health consultation data. Performance measurement is performed by comparing the accuracy of classification. The first step of the proposed model is data collection for training. The next step is data preprocessing. After that, model training is performed. Finally, a validation step is run with the test data as shown in Fig. 6.

A crawling software dedicated to data collection was implemented with C# Winform Application and .Net Framework 4.5. The access identifier Selector Path was used to access the page and to identify specific elements for collection and organization. However, since random tag IDs are assigned to

medical answers, it is difficult to access with a simple Selector Path, so Full Xpath was used. In addition, the answers to questions from the collected data were filtered with only the answers of experts with a medical license.
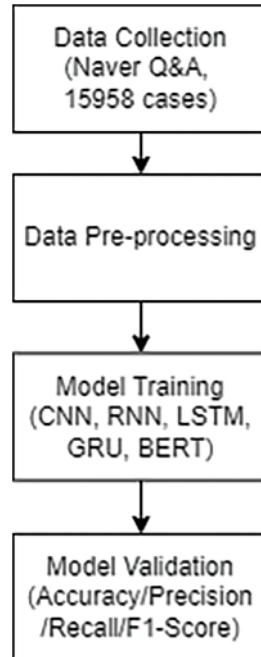

Figure 6: Research framework

Text data containing symptom information is necessary for executing various algorithms through minimal pre-processing after crawling the contents of the Naver Q&A Service. The pre-processing process is described in more detail later. The total number of data collected in this way is 15,958. 20% of them divided using the scikit-learn library is used to evaluate the performance of learning models mentioned above. As a result, the total number of data used for training is 12,766, and the number of data used for testing is 3192. As for the labeling of the data set, the counseling subject information in the crawling data was classified into nine categories of medical subjects such as Orthopedics, Dermatology, Otorhinolaryngology, Surgery, Neurosurgery, Thoracic Surgery, Colorectal/Anal Surgery, Neurology, and Internal Medicine, and they are assigned from 0 to 8 in order, respectively. The period of data collection was from December 03, 2021, to January 15, 2022. Table 2 shows sample data examples with English translation from the datasets crawled.

Table 2: Data sample

Question	Label
Is it tinnitus?!~Is it an ear disease?	2
I'm trying to get treatment for cold hands and feet~Life is so hard.	5
Will my finger tendon inflammation get better?~I don't think it's getting better.	0
The left floating chamber hurts so much~I am 29 years old.	4

(Continued)

Table 2 (continued)

Question	Label
I have a headache and I think I'm going to vomit~Why is this?	3
How to get rid of tiny thorns on your hands~How do I remove this thorn?	1
Around the anus when sitting~Why is this?	6
If the liver level is high, endoscopy~Or should I go to the hospital?	8
Do I have a migraine?~What should I do?	7

Fig. 7 shows the amount of data for each medical subject. It shows that the data except for the thoracic surgery data are evenly distributed.
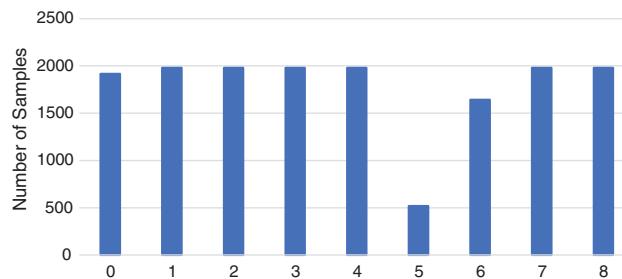
Figure 7: Dataset distribution

As for pre-processing, it is carried out in the same order as shown in Fig. 8. First, special characters and space characters are removed in order to leave only Korean in the crawled data. In addition, stopwords are removed from sentences in which special characters and whitespace characters are removed. For the morpheme analyzer, the open-source Korean morpheme analyzer Okt (Open Korean text) [25] made by twitter is used. This analyzer supports four functions: normalization, tokenization, rooting, and phrase extraction. After morphological analysis using Okt, rare words are removed based on the number of appearances. The threshold value is set to 2 times. Consequently, 9,869 rare words are removed out of a total of 18,809 words, so that the ratio of words that appeared twice or less is about 52.5%.

For BERT, the KoBERT model developed by SK T-Brain was used. To measure the accuracy according to the models, CNN, RNN, LSTM, GRU, and BERT are implemented. The pre-processed data is embedded through an embedding layer. Finally, using the dense layer and the activation function softmax function, it is outputted.

Fig. 9 shows the structure of the classification model using BERT. The text pre-processed through BERT Tokenizer goes through the BERT model and the Dense Layer for final classification. Through this, the label is finally predicted according to the input value.

Implementation is carried out in Google's Colab [26]. The questions are labeled as medical subjects. To compare model performance, CNN, RNN, LSTM, GRU, and BERT models are implemented using Tensorflow and Pytorch libraries.
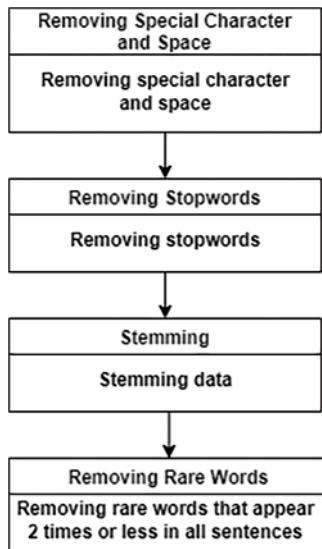
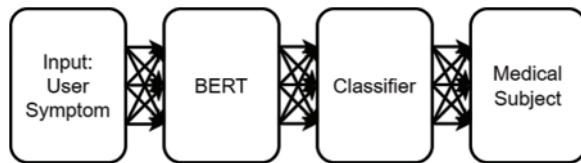**Figure 8:** Data pre-processing stages

Figure 9: Overall structure of BERT classification model

5 Simulation Result

This study was conducted in Colab, a machine learning environment provided by Google, to classify Korean health counseling data. The specifications of the Colab environment are shown in Table 3.

Table 3: Colab specifications

OS	Linux Ubuntu 18.04
Program language	Python 3.7.12
CPU	Intel(R) Xeon(R) CPU @ 2.20 GHz
GPU	Tesla T4

This paper verifies the performance of NLP algorithms based on deep learning by comparing the accuracy of classifying the treatment subject from text-based health counselling data. In general, the results of the classification model come out as two values, positive and negative [18]. At this time, four values of the confusion matrix were used to check whether this value was correctly classified. The four values can be divided into True Positive (TP), False Positive (FP), or True Negative (TN) and False Negative (FN). These evaluation methods are Precision, Recall, and Accuracy, which are often used as

evaluation indicators for machine learning classification models, and the equations for each evaluation indicator are as follows [27]:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Accuracy = \frac{TP}{TP + FP + FN + TN} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precison + Recall} \tag{16}$$

Precision refers to a value that is actually True positive among all data. Recall refers to a value in which the value determined to be actually True by the recall is actually True. Finally, accuracy refers to the percentage of all data that is correctly judged as True. The reason why multiple evaluation indicators are used is that each evaluation indicator cannot be an absolute evaluation indicator. In this paper, the model was evaluated through the above four evaluation indicators.

The difference in the results according to algorithms was clearly shown.

Table 4 and Fig. 10 show the results for each model. Recall result of each algorithm showed 73.7% of CNN, 53.3% of RNN, 66.4% of LSTM, 70.5% of GRU, and about 75.8% of BERT. Next, the result values for Precision showed performance of CNN 73.7%, RNN 51.9%, LSTM 69.8, GRU 69.1, and BERT 74.6%. And the F1-score test result showed the performance of CNN 73.6%, RNN 52.4%, LSTM 69.8%, GRU 69.1%, and BERT 75.1%. Finally, the Accuracy measurement result showed the performance of CNN 74.2%, RNN 58%, LSTM 71.7%, GRU 70.1%, and BERT 76.3%. In the case of BERT, it not only shows the best performance in accuracy, but also shows excellent performance in all other conditions. The accuracy appeared in the order of BERT, CNN, LSTM, GRU, and RNN. It was confirmed that BERT was about 2% different than the second highest CNN.

Table 4: Simulation results

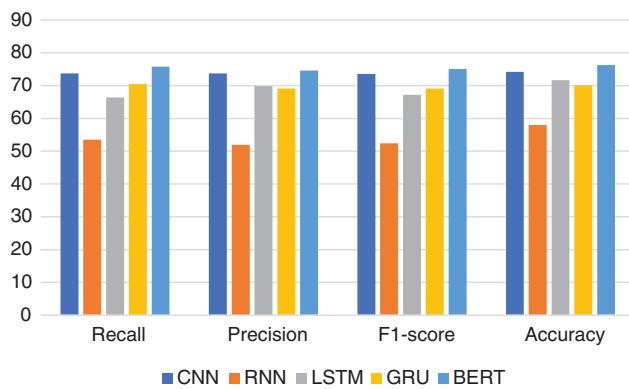| Model	Recall	Precision	F1-score	Accuracy
CNN	73.7	73.7	73.6	74.2
RNN	53.5	51.9	52.4	58.0
LSTM	66.4	69.8	67.2	71.7
GRU	70.5	69.1	69.1	70.1
BERT	75.8	74.6	75.1	76.3

Figure 10: Comparison results

6 Conclusions and Future Work

In this paper, a medical subject classification model according to symptom data for a healthcare chatbot was implemented, and the performance of each algorithm was analyzed. The dataset was collected by crawling the data of the Naver Q&A service. As a result of comparing the performance of CNN, RNN, LSTM, GRU, and BERT, the BERT model showed the best performance in all four evaluation indicators.

It is expected that this system will allow the user to assign appropriate medical subjects according to the symptoms of the user. In this study, 9 medical subjects were classified, but performance verification is required after securing additional data through the expansion of medical subjects. Next, performance comparison was made through one BERT model, but in the situation where various domestic BERT models and various natural language processing algorithms are emerging, in order to improve the performance of this model, data set processing and expansion, additional algorithm performance analysis, and other additional studies such as performance verification of the Korean BERT model are needed. Furthermore, further studies on the medical chatbot system and the intervention module of the multi-chatbot system based on this model are planned.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Gwon, H. M., Seo, Y. S. (2021). Towards a redundant response avoidance for intelligent chatbot. *Journal of Information Processing Systems, 17(2),* 318–333.
2. Cha, G. H. (2021). The kernel trick for content-based media retrieval in online social networks. *Journal of Information Processing Systems, 17(5),* 1020–1033.

3. Shanmuganathan, V., Yesudhas, H. R., Madasamy, K., Alaboudi, A. A., Luhach, A. K. et al. (2021). AI based forecasting of influenza patterns from twitter information using random forest algorithm. *Human-Centric Computing and Information Sciences, 11(3),* 1–14.

4. Zeng, Y., Zhang, R., Yang, L., Song, S. (2021). Cross-domain text sentiment classification method based on the CNN-BiLSTM-TE model. *Journal of Information Processing Systems, 17(4),* 818–833.

5. Heo, E. Y., Jeong, J. H., Kim, H. H. (2021). Detection of adverse drug reactions using drug reviews with BERT+ algorithm. *KIPS Transactions on Software and Data Engineering, 10(11),* 465–472.

6. Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. G., Almerekhi, H. et al. (2020). Developing an online hate classifier for multiple social media platforms. *Human-Centric Computing and Information Sciences, 10(1),* 1–34. DOI 10.1186/s13673-019-0205-6.

7. Nasr, M. M., Islam, M. M., Shehata, S., Karray, F., Quintana, Y. (2021). Smart healthcare in the Age of AI: Recent advances, challenges, and future prospects. *IEEE Access, 9,* 145248–145270. DOI 10.1109/ACCESS.2021.3118960.

8. Rahaman, A., Islam, M. M., Islam, M. R., Sadi, M. S., Nooruddin, S. (2019). Developing IoT based smart health monitoring systems: A review. *Revue D'Intelligence Artificielle, 33(6),* 435–440. DOI 10.18280/ria.

9. Miner, A. S., Laranjo, L., Kocaballi, A. B. (2020). Chatbots in the fight against the COVID-19 pandemic. *NPJ Digital Medicine, 3(1),* 1–4. DOI 10.1038/s41746-020-0280-0.

10. Jovanović, M., Baez, M., Casati, F. (2020). Chatbots as conversational healthcare services. *IEEE Internet Computing, 25(3),* 44–51. DOI 10.1109/MIC.2020.3037151.

11. Wu, Z., Chitkushev, L., Zhang, G. (2020). A review of telemedicine in time of COVID-19. *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3005–3007. Seoul, Korea.

12. Kandpal, P., Jasnani, K., Raut, R., Bhorge, S. (2020). Contextual chatbot for healthcare purposes (using deep learning). *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pp. 625–634. London, UK.

13. Softić, A., Husić, J. B., Softić, A., Baraković, S. (2021). Health chatbot: Design, implementation, acceptance and usage motivation. *2021 20th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pp. 1–6. East Sarajevo, Bosnia and Herzegovina.

14. Chebrolu, K., Ressler, D., Varia, H. (2020). Smart use of artificial intelligence in health care. https://www2.deloitte.com/us/en/insights/industry/health-care/artificial-intelligence-in-health-care.html.

15. Kim, D. Y., Cho, M. K., Shin, H. C. (2020). The application of artificial intelligence technology in counseling and psychotherapy: Recent foreign cases. *The Korean Journal of Counseling and Psychotherapy, 32(2),* 821–847. DOI 10.23844/kjcp.2020.05.32.2.821.

16. Park, S. U. (2021). Analysis of the status of natural language processing technology based on deep learning. *The Korea Journal of BigData, 6(1),* 63–81. DOI 10.36498/KBIGDT.2021.6.1.63.

17. Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751. Doha, Qatar.

18. Zaremba, W., Sutskever, I., Vinyals, O. (2014). Recurrent neural network regularization. arXiv preprint arXiv:1409.2329.

19. Lee, E. J. (2017). Basic and applied research of CNN and RNN. *Broadcasting and Media Magazine, 22(1),* 87–95.

20. Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9(8),* 1735–1780. DOI 10.1162/neco.1997.9.8.1735.

21. Chung, J., Gulcehre, C., Cho, K., Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

22. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. et al. (2017). Attention is all you need. In: *Advances in neural information processing systems*, vol. 30, pp. 6000–6010.

24. KoBERT. https://github.com/SKTBrain/KoBERT.

25. Konlpy, Okt. https://konlpy.org/en/latest/api/konlpy.tag/.

26. Google Colab. https://colab.research.google.com/.

27. Muhammad, A. F., Susanto, D., Alimudin, A., Adila, F., Assidiqi, M. H. et al. (2020). Developing English conversation chatbot using dialogflow. *2020 International Electronics Symposium (IES)*, pp. 468–475. Surabaya, Indonesia.