ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20220406064517
Original scientific paper

A Spoken Dialogue Analysis Platform for Effective Counselling

Seok Kee LEE*, Sung-Dong KIM

Abstract: This paper proposes a spoken dialogue analysis platform (SDAP) that could assist counsellors in person-to-person counselling by analysing counselling conversations and providing key information that could enhance the counsellors' understanding of the counselees' conditions and situations. The proposed platform has two main modules: a speech recognition module and a text analysis module that are specifically built for the Korean language. The speech recognition module uses NAVER CLOVA Speech service to convert voice recordings of counselling dialogues into text. The Korean text analysis environment of the text analysis module was built using NLTK, KoNLPy and scikit-learn library, and, for now, the module provides two types of text analysis: keyword analysis and sentiment analysis. The results of the text analyses that provide keywords and analysis of customers' emotional state can help counsellors to provide appropriate feedback to the counselees easily and more quickly, making the counselling fast and effective and reducing the counselees' waiting time. In the experiments, the text analysis module building process is elaborated in detail, and the usefulness of the proposed SDAP is exemplified by case studies on actual counselling conversations at a dental clinic and a fitness centre.

Keywords: dialogue analysis; keyword analysis; sentiment analysis; speech recognition; text analysis

1 INTRODUCTION

Many of the professional or personalized services that are available today require customers or clients who seek to use those services to engage in a preliminary consultation. For instance, when people go to a medical clinic or a fitness centre for the first time, they are usually met by a counsellor (or a customer service representative), who asks a specific set of questions aimed at understanding the counselee's situation and condition. The counsellor then provides necessary information based on the contents of the conversation. For example, in the case of a medical clinic, the counselees get information on the treatment they will be receiving after counselling; at a fitness centre, they will be told about the exercises or training programs that are suitable for them. The heavy burden on counsellors to perform many consultations on any given day could make it difficult for them to provide suitable feedback each time. At times, there may also arise a situation where the counsellor is already in a consultation or away from the office, requiring counselees to wait or to schedule another appointment. In this regard, counselling support systems help counsellors to provide feedback more quickly and effectively by automatically generating the information necessary for appropriate feedback from a recorded counsellor-counselee conversation based on a predetermined inquiry. Moreover, these systems make it possible to hold consultations even when the designated counsellor is absent.

Early versions of counselling support systems merely provided a counselling dialogue recording function. Then the "xtrmSolution" was developed to assist call centre agents by providing agents with important keywords for identifying customers' intentions. Since then, new technological developments have enhanced the functionality of counselling support systems, including customizations for specific industries or purposes. For instance, "Ali Duo" was developed to improve the efficiency in selling insurance and financial products. Today, customer or client consultations are being conducted in a broad range of fields. Companies have set up call centres to handle customer inquiries, some introducing chatbots to support call centre operations and even replace human agents. Still, consultations involving important or complex content cannot be covered by pre-determined scenarios, thus requiring human interaction. Therefore, there is a growing need for counselling support systems that could provide conversation analysis and information generation functions.

The stability and accuracy of speech recognition has been raised and reached to a level that allows its practical application by the recent development of deep learning technology. In fact, big IT companies, such as Google, Microsoft and Naver, offer commercially available speech recognition tools for developers to use. Also, various natural language text analysis methods using machine learning and deep learning have been released for extracting or generating information that could enhance understanding of all kinds of texts. Leveraging these technologies, this paper presents a spoken dialogue analysis platform (SDAP) that analyses recordings of voice conversations held between the counsellor and the counselee and generates the information necessary for providing appropriate feedback. The proposed platform is composed of two modules: a speech recognition module for converting recorded audio files into text, and a text analysis module for extracting key information from the transcribed conversations using keyword analysis and sentiment analysis functions. Through these two modules, the proposed platform can facilitate consultations, especially in terms of lessening the burden of processing the counselees' attitudes and responses, so that the counsellor can quickly provide appropriate feedback to the counselee. Thus, the SDAP can reduce the fatigue felt by the counsellor and shorten the length of consultations, which, in turn, reduces the amount of time counselees must wait to receive consultation. In addition, this platform can be extended to be used in various counselling fields by adding other text analysis functions such as named entity recognition and intention identification.

The proposed SDAP is specifically built for the Korean language and is created as a user-friendly web-based platform so that anyone could use it easily even if they do not have much knowledge of computer-based technologies. Users only need to upload the recording of the conversation onto the platform and click on button for the analysis they wish to perform.

This paper consists of 5 sections. The overall structure of the proposed SDAP is presented in Section 2. Section 3 explains each component of the platform in detail. Section 4

elaborates on how the text analysis module was developed and uses actual examples of counselling at a dental clinic and a fitness centre to demonstrate its usefulness. Section 5 concludes the paper with a summary and suggestions for further research and development.

2 RELATED WORKS

A spoken dialogue system refers to a computer system that can hold voice-based conversations with people and could be defined as "informatics systems that allow humans to interact with such systems using natural language" [1]. Existing spoken dialogue systems generally perform voice recognition and voice synthesis.

A significant amount of research has been conducted on the development of speech language processing systems capable of speech recognition and generation. Traditionally, speech recognition technologies have depended on Hidden Markov models (HMM) [2], but these models were unable to achieve practical levels of performance. Thus, other approaches were explored and adopted. One approach applied neural networks to speech recognition by building a speech recognizer made up of a listener, a speller, an encoder, and a decoder, and applying a recurrent neural networks for the encoder and decoder. The neural network approach was found to achieve better performance for conversational speech recognition [3]. Another approach involved data augmentation using multi-condition training data, which was found to provide more robust speech recognition [4]. For speech synthesis, statistical parametric methods were adopted, of which HMM-based synthesis has shown to be effective [5]. Also, the development of deep learning has improved the quality of synthesized speech performed by neural network-based systems [6].

Cloud service providers, such as AWS (Amazon Web Service), Alibaba Cloud, Naver Cloud, KT Cloud, Microsoft Azure, and Google Cloud, currently provide high-performing speech recognition and synthesis services as application programming interfaces (APIs). Thus, more emphasis is placed nowadays on developing devices and applications that make use of these services rather than conducting fundamental research on speech recognition and synthesis. For instance, smart speakers that are commercially available in the market, such as Naver CLOVA Speaker, Google Nest Audio, Amazon Echo, and Apple HomePod Mini utilize these speech recognition and synthesis APIs to provide smart services.

In [7], the authors proposed a system that recognizes voice data, converts it into text, and analyses and classifies the text data efficiently, which is similar to the SDAP proposed in this paper in that this system combines speech recognition and text analysis. Recently a state-of-the-art speech recognition system achieved a 1.4%/2.6% word error rate [8], establishing the high level of accuracy achieved by speech recognition technology. Another study also proposed a counselling support system that combines speech recognition and text analysis [9], which demonstrated the usefulness of a platform that helps counsellors' understanding of conversations with counselees through speech recognition and text analysis. It is expected that the proposed SDAP integrating speech recognition and text analysis technologies will become an important and necessary tool for counselling in various industries and fields.

3 STRUCTURE OF SPOKEN DIALOGUE ANALYSIS PLATFORM

The overall structure of the proposed SDAP is shown in Fig. 1. SDAP is a web-based system with a server built using Flask on AWS's Amazon Elastic Compute Cloud (EC2), which provides secure and resizable cloud-computing capacity [10], to ensure stable service. Ubuntu 18.04 was installed as the operating system. An Amazon EC2 server can be created simply by accessing AWS's platform and configuring a virtual server, referred to as "instance", on the cloud to fit your needs and environment (e.g., operating system and hardware specifications). Flask, a Python-based micro web framework [11], was chosen as the server programming framework because text analysis and machine learning services are generally built in Python language. Flask requires the pre-installation of Python 3.6 or newer versions, and it is recommended to install Flask on the environment *after* generating the specific virtual environment.

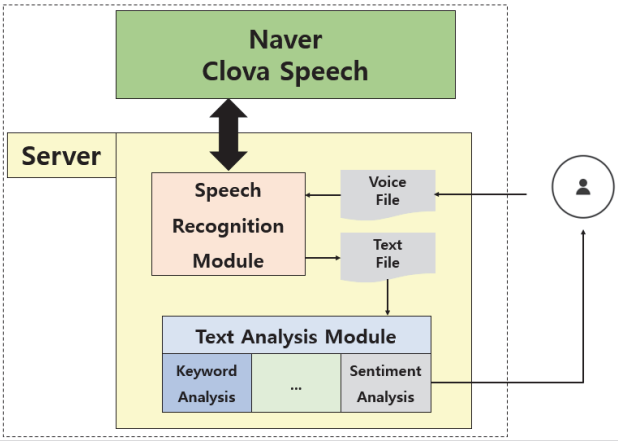

Figure 1 Structure of the spoken dialogue platform

The SDAP's two main modules (speech recognition/text analysis) were loaded on the server. The speech recognition module uses Naver's CLOVA Speech tool [12] to convert voice recording files into text format. CLOVA Speech takes a long audio/video file and processes it using speech recognition technology to provide outputs in text format. There is also CLOVA Speech Recognition (CSR) [13], which is optimized for short speech recognition (i.e., shorter audio/video files). The environment for the SDAP's text analysis module for Korean text was built using the Natural Language Toolkit (NLTK) [14], KoNLPy [15], and Scikit-Learn library [16].

At present, the text analysis module provides two types of analyses (keyword and sentiment), but additional text analysis functions could be added in the future.

4 COMPONENTS AND DEVELOPMENT PROCESS OF SPOKEN DIALOGUE ANALYSIS PLATFORM

This section describes how the speech recognition and text analysis modules, which are the main components of the proposed SDAP, were developed and how they function. The speech recognition module performs speech

recognition using Naver's CLOVA Speech service. The text analysis module applies natural language processing methods to analyse the text provided by the speech recognition module and extract or generate information that could guide counselee feedback.

Notably, the text analysis module could be expanded and upgraded continuously by adding more functions, which can enhance the practicality and efficiency of this platform as a counselling support system. There are currently diverse types and methods of text analysis, including sentiment analysis, intention classification, text classification and categorization, named entity recognition, question-answering, document summary, SNS analysis, and machine translation. Machine learning techniques have been applied to many of these methods, while deep learning-based models have also been developed and applied at the practical level in recent years, furthering the research on text understanding.

4.1 Speech Recognition Module

There are a variety of commercially available voice recognition APIs available as a cloud service. The speech recognition APIs offered by Amazon, Google, Microsoft, and Naver provide deep learning-based voice recognition with practically acceptable performance. Naver's CLOVA Speech service was chosen as the most suitable API for the proposed SDAP's speech recognition module, since Naver specifically targets and develops tools for the Korean language. Although Amazon, Google, and Microsoft also offer natural language processing services for the Korean language, these services are geared towards processing diverse languages, so their accuracy for the Korean language is not as high as Naver's CLOVA Speech.

CLOVA Speech can be used in two ways: API call method and CLOVA Speech builder. The API call method was used for our proposed SDAP [17]. To use the API call method, first, a domain was created, which is the unit that manages the speech recognition target and result files. An API key is assigned to each domain, which serves as a unique identifier for authenticating users' API calls. The SDAP sends the target file for speech recognition to the unique API call URL provided by the domain and receives the result processed by CLOVA Speech's speech recognition engine. Also, a storehouse, called object storage, is created for storing the target and result files.
The speech recognition using CLOVA Speech is processed as follows. When the user uploads the voice recording file to the server through the SDAP's user interface, a speech recognition request object is created on the server and sent to the CLOVA Speech engine. The engine processes the file and sends the output data back to the server in JavaScript Object Notation (JSON) format. Then, the transcribed text is extracted from the JSON- format output data and shown on the user interface to become the input to the text analysis module.

4.2 Text Analysis Module

As previously mentioned, the proposed SDAP specifically targets Korean speech dialogue for analysis. Since text analysis machine learning algorithms are mainly developed in Python language, a Korean text analysis environment was constructed using text analysis tools on the Python development environment. The proposed SDAP will ultimately provide various text analysis functions, but at present, it provides two types of text analysis: keyword analysis and sentiment analysis. The following section describes the environment for Korean text analysis that was built for the SDAP and the text analysis module's keyword analysis and sentiment analysis functions.

4.2.1 Environment for Korean Text Analysis

The project development environment was built on Anaconda [18]. Anaconda is a distribution (prebuilt and preconfigured collection of packages) of Python. It is well known and widely used for data science and data processing. Anaconda distribution also provides the Conda package manager that is a management system for setting up Python environments and installing additional conda packages that are available in the Anaconda distribution. The open-source edition of Anaconda was installed. The Anaconda package includes a compatible Python module, so it is recommended to uninstall any pre-installed Python versions before installing Anaconda to ensure compatibility. Next, a virtual environment was generated, and the necessary tools for Korean text analysis were installed on the virtual environment so that they can be effectively and easily managed. Different types of text analysis require different development tools, and for the two text analysis functions, the following tools were installed: NLTK, KoNLPy, MeCab morpheme analyser, and Scikit-Learn.

NLTK is well known and widely used platform for building Python programs to work with natural language data. It provides easy-to-use interfaces that cover over 50 corpora and lexical resources such as WordNet. It also provides a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. After installing NLTK, NLTK data should be downloaded using NLTK's data downloader. The data is necessary for tasks such as pre-processing for a specific language.

KoNLPy was installed for Korean text analysis. KoNLPy is a Python package for natural language processing of Korean language and provides libraries related to Korean text and language processing such as morpheme analysis, part-of-speech tagging, and syntax analysis. KoNLPy is implemented using Java language, so Java Development Kit (JDK) [19] and JPype should be installed before installing KoNLPy on a Windows operating system. JPype is a Python module that provides full access to Java from within Python. By using JPype, Python programs can make use of only Java libraries, explore and visualize Java structures, develop and test Java libraries, perform scientific computing, and much more [20]. MeCab morpheme analyser [21] is a part-of-speech morphological analyser that is known to perform better than those provided by KoNLPy. It was additionally installed to improve the performance of morpheme analysis using user dictionaries.

Scikit-Learn is a widely used machine learning library for Python. Various machine learning algorithms for classification, regression, and clustering are provided and

the algorithms for dimensionality reduction, model selection, and pre-processing are also included in the library. Scikit-Learn was installed using *pip*, which is the package installer for Python included in the Python language package. It can be also easily installed in the conda environment. Fig. 2 shows the resulting environment structure of Korean text analysis.

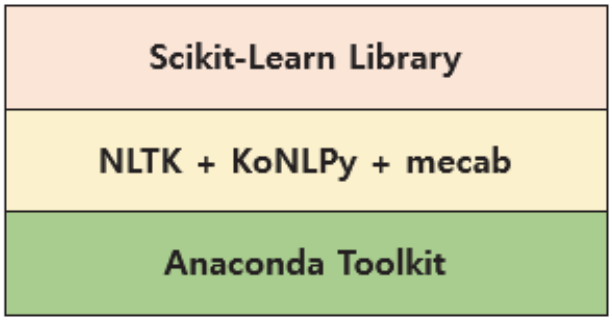

Figure 2 Environment stack of Korean text analysis

The transcribed text obtained as the output of the speech recognition is pre-processed using NLTK, KoNLPy, and MeCab morpheme analyser, then text analysis, such as sentiment analysis, and keyword analysis, is performed using Scikit-Learn's machine learning algorithms for extracting or generating necessary information from the text.

4.2.2 Keyword Analysis

Keyword analysis (also known as keyword detection or keyword extraction) is a text analysis technique that automatically finds the frequently appeared words and phrases from a given document. By summarizing the content of texts and identifying the main topics being discussed using these words and phrases, valuable insights into the topics customers are talking about can be identified. The goal of keyword analysis is simple: to find out what the user is searching or wants. In general, keyword analysis helps users to determine what terms people are searching for in relation to products, services, areas of business or interests, and so on. For this reason, keyword analysis is often connected to search engine optimization (SEO). In this paper, the term keyword analysis is used to refer to the process of finding important words in a given text.

Keyword analysis is commonly used on all kinds of texts, from regular documents and business reports, replies in social network services, to various kinds of reviews and more. Companies use keyword analysis to extract the most mentioned attributes about their products and services from customer reviews or to follow conversations on social media to understand their audience, make improvements, or respond quickly to situations that could potentially result in a PR crisis. Keyword analysis can be a powerful ally for brand monitoring. The most important words and phrases mentioned in customer feedbacks or during customer support conversations can be easily identified. With the help of identified keywords, companies understand customer's responses or thoughts and obtain interesting insights and keys for product or service improvement. In terms of our proposed SDAP, the keyword analysis function supports counsellors in identifying the situation and needs of counselees from the conversations held during consultations.

There are many approaches to keyword analysis. Well-known statistical approaches include word frequency, word collocation and co-occurrences, term frequency and inverse document frequency (TF-IDF), and rapid automatic keyword extraction (RAKE) [22]. Keyword analysis methods often make use of linguistic information about the documents, such as morphological or syntactic information, discourse markers, and semantic information [23]. In [24], the authors explored a new graph-based approach for extracting key phrases related to the major topics within a text. Support vector machines (SVM) [25] and deep learning [26] are also used to extract the most relevant keywords in a text. Recently, Bidirectional Encoder Representations from Transformers (BERT) [27] has risen as a popular model for most natural language processing applications due to its high performance. In [28], the authors improved keyword analysis performance using BERT. BERT was used to extract key sentences from the given text as information that supplements the original text. Then, statistical methods were applied to the extended text (the original text plus the key sentences) to extract keywords [28].

The method for keyword analysis applied to our proposed SDAP simply extracts frequently occurring words to determine the keywords, without applying more complex machine learning based methods. This simple method was chosen because checking the words that appear frequently in conversation is generally sufficient for identifying major issues during counselling conversations and figuring out what the counselee needs. Based on the issues identified by the keywords, the counsellor can quickly suggest an appropriate solution to the counselee. For instance, in medical consultations, the counsellor can easily recognize the client's situation from the extracted keywords presented along with their frequency of appearance and suggest the appropriate area of medical treatment.

The SDAP's keyword analysis function processes the transcribed text as follows. NLTK provides the sentence tokenization function, *sent_tokenizer*(), which separates what the counsellor and the counselee said in the speech recognition result text file into individual sentences. MeCab morpheme analyser then performs morphological analysis on each sentence and tags the words with part-of-speeches. When the morphological analysis of the entire text is completed, only the nouns are extracted, and their appearances in the text file are counted. The words are sorted in the order of their occurrence frequency, and words whose number of occurrences is above threshold α and above top n are selected and presented as keywords.

4.2.3 Sentiment Analysis

Sentiment analysis, also known as opinion mining, ranges from detecting emotions (e.g., anger, happiness, fear) to sarcasm and intent (e.g., complaints, feedback, opinions). In its simplest form, sentiment analysis assigns a polarity (e.g., positive, negative, neutral) to a piece of text. It uses natural language processing and machine learning to interpret and classify the emotion and intention in subjective data [29].

Various approaches have been developed for performing sentiment analysis. Most early studies on

sentiment analysis were based on the Bayesian method. For example, Naïve Bayes algorithm was used for sentiment analysis of restaurant reviews [30]. Despite being very simple, the Bayesian method has been proven to be relatively accurate. Bayesian ensemble learning achieved high performance in sentiment analysis [31], and Bayesian networks classifiers showed competitive predictive results for sentiment analysis of twitter data when compared to support vector machines (SVMs) and random forest [32]. In addition, the relations among words can be identified by the resulting Bayesian networks.

In recent years, deep learning models have proved as a promising solution to the various natural language processing problems. Studies have suggested approaches to sentiment analysis using deep learning [33] and convolutional neural networks (CNN) [34, 35]. In [36], the authors used CNN to perform a sentiment analysis of Korean movie reviews and showed that different Korean text pre-processing methods (bigram, trigram, and unigram) could lead to different information outputs. Furthermore, the study achieved a higher performance of around 91% accuracy by combining the different information obtained from different pre-processing methods. Sentiment analysis using deep learning has also been performed with ensemble techniques that apply several ensembles of classifiers by integrating surface and deep features [37]. The results of the ensemble deep learning model performed on several sentiment datasets showed that the sentiment analysis performance of a deep classifier can be leveraged when using an ensemble of classifiers.

Although deep learning outperforms other machine learning methods for many complex problems, it also requires a lot of data and high-end hardware. In this paper, neural network models were constructed for sentiment analysis using Tensorflow 2.0 [38]. Due to the limitations in available counselling conversation data, the neural networks were trained using Naver's sentiment movie corpus [39]. Although Naver's sentiment movie corpus contains movie review data, the large size of the corpus makes it suitable for providing sufficient training for neural networks to perform sentiment analysis of consultations on exercise and fitness programs at fitness centres, which is the subject of the case study on sentimental analysis presented in this paper.

Sentiment analysis is widely used in various areas to gain information and insights for decision making. In business related applications, sentiment analysis can help organizations assemble the clients' needs and improve the quality of products and services based on the feedback collected from customers. Sentiment analysis can be applied as a part of trend analysis to predict market scenarios or trends [40]. Sentiment analysis can also play a role in fighting COVID-19 and infectious diseases. In [41], the authors performed sentiment analysis on the literature published over the last ten years to identify the most important texts, which were categorised into motivations related to disease mitigation, data analysis and challenges faced by researchers with respect to data, social media platforms, and community. This study emphasised the current standpoint and opportunities for research in this area and encouraged additional efforts on this topic.

Sentiment analysis can also be used advantageously in the medical domain [42]. There is much online information about healthcare that are not obtained methodically, such as personal blogs, social media, and websites about medical issues. Using sentiment analysis, it can be possible, for instance, to evaluate the expansive medical information available online to discern positive information that can be used to improve the quality of healthcare.

The proposed SDAP's sentiment analysis function can support counsellors in terms of giving suitable advice and recommendation that consider the emotive state of counselees. For instance, the SDAP's sentiment analysis can be used to assess how the counselee feels about the exercise he/she just completed so that the counsellor can provide appropriate encouragement and suggest better exercise plans.

4.3 User Interface of SDAP

Fig. 3 shows the user interface of the proposed SDAP's web-based system. Users could upload the voice recording file through this user interface, which sends the user's request to the server. The speech recognition and text analysis modules on the server analyse the file, then provide the transcribed text and visualized text analyses on the user interface.

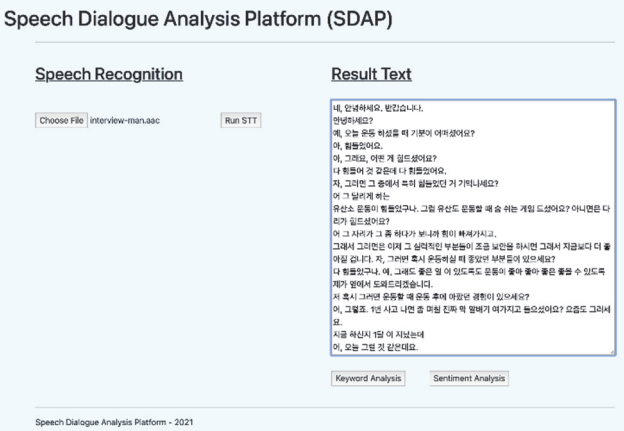

Figure 3 User interface of SDAP

The steps for uploading the voice recording file and receiving analysis results are as follows. The voice recording file could be uploaded by clicking on the "Choose File" button in the [Speech Recognition] part on the left side of the screen, then by selecting the voice recording file. When the user presses the "Run STT" button, the recording is processed with speech recognition. The result of speech recognition processing is presented in the [Result Text] part on the right side of the screen. Users could then either click on the [Keyword Analysis] button or the [Sentiment Analysis] button to perform the respective analysis on the speech-recognized text document.

5 EXPERIMENTAL RESULTS

This section describes how the neural network-based sentiment analysis model was built and demonstrates the usefulness of the proposed platform through case studies

using actual counselling conversations conducted at a dental clinic and a fitness centre.

5.1 Sentiment Analysis Model

Two types of neural network-based sentiment analysis models were constructed for the SDAP's sentiment analysis function: a vanilla neural network model and a long short-term memory (LSTM)-based neural network model. The vanilla neural network has a simpler structure, so it is easy to train, while the LSTM neural network can be more accurate because it can consider the context.

5.1.1 Experimental Environment

The hardware specifications of the machine that was used for sentiment analysis model building are shown in Tab. 1. Although the GPU is relatively low end, it was deemed sufficient for the parameters and complexities of the trial models. The Windows 10 operating system was installed on the machine.

Table 1 Hardware specifications for sentiment analysis model building

	Spec.
CPU	Intel® Core i7-6700, 4 core, 3.4 GHz
RAM	16 GB
GPU	NVIDIA GeForce GTX 960, 2 GB

Naver's sentiment movie corpus was used as the data for training and testing the sentiment analysis model. Naver's movie reviews are mostly one sentence long, with lengths up to 140 characters. The corpus consists of 150 000 training data and 50 000 test data, and each sentiment class was sampled equally. Movie reviews that give points from 1 to 4 for the movie were considered negative reviews, and those giving points from 9 to 10 were considered as positive reviews.

5.1.2 Model Building

The proposed SDAP model's sentiment analysis model was built through the following process. The building process began with data loading. 200,000 reviews from Naver's movie review data were loaded. First, word tokenization was performed using KoNLPy for both the vanilla neural network model and the LSTM-based model. Then, further text pre-processing was conducted differently for each model.

The vanilla network model requires part-of-speech information, so MeCab morpheme analyser was employed to tag each word in the data with part-of-speech. The total number of tokens in the training data was about 2160 K. Since using words with specific parts of speeches, rather than using all words, for sentiment analysis can render better results, only the words that are tagged as adjectives, adverbs, exclamations, modifiers, nouns, and verbs were extracted, which reduced the number of tokens to 1460 K. Then, NLTK's text module, which provides various functions for convenient document search, was used. The number of word occurrences was calculated using the *vocab*() function, then words with a high frequency of occurrence were extracted using the *most_common*() function. Through this process, word tokens were generated from the document set, and the number of each word token's occurrence was counted to generate a bag-of-words (BOW)-encoded vector that was used as input for the neural networks. The bag-of-words model is a simplifying representation used in natural language and information retrieval. In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity [43]. Since the existence of a specific word of a specific part-of-speech is important in determining sentiment, BOW encoding, which does not consider the order of words in a sentence, was used. The vocabulary was defined with the 10 K most frequent occurrences, meaning that a sentence of a movie review is represented by 10 000-dimension vector, and each element of the vector represents the number of times each word in the vocabulary appears in the sentence. Afterwards, the vector was converted to a *float32* type to be used as a neural network input.

The vanilla neural network model takes 25 words as input, and one hidden layer has 50 nodes. Since one movie review generally consists of one sentence, and about 95% of sentences in Naver's movie review data are 25 words or less, it was assumed that one sentence has a maximum of 25 words. The ReLU activation function was used for the input and hidden layers, and the sigmoid activation function was used for the output layer to compute the probability that the input sentence is positive. After trying various settings using different optimizers (RMSProp, SGD and Adam), different *validation_split*, and different *batch_size*, the model showing the best test accuracy was selected. Tab. 2 shows the structure and the training parameters for the chosen model network.

Table 2 Vanilla neural network's structure/training parameters

Structure			Training Parameters	
Layer	Node #	Param #	Optimizer	RMSprop
Input	25	250025	Loss function	binary crossentropy
Hidden	50	1300	Batch size	128
Output	1	51	Validation data	20%

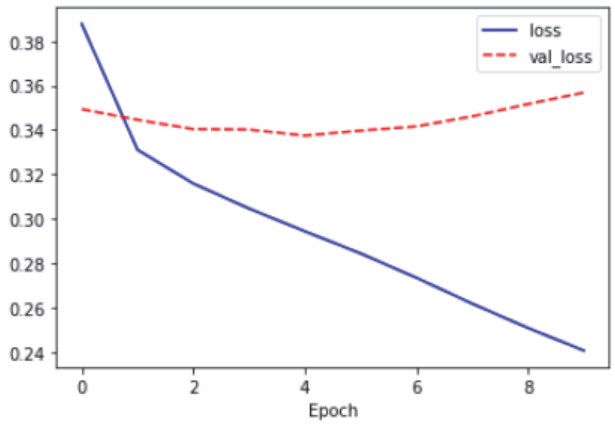

Figure 4 Training loss of the vanilla neural network model

Fig. 4 shows the changes in the vanilla neural network model's training loss over 10 epochs, which it took about 40 seconds. This is very little time compared to the time to pre-process the training data. It is possible to reduce the learning time by using a neural network with a simple structure, which can generate a high-performing model through learning in various settings. The figure says that

the training loss continues to decrease overall, but validation loss increases slightly after epoch 5.

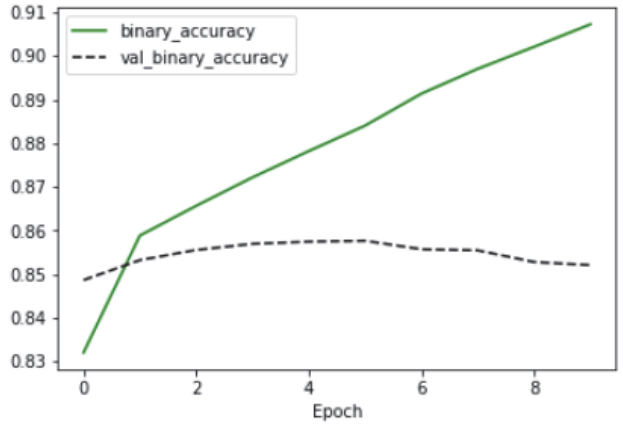

Figure 5 Training accuracy of the vanilla neural network model

Fig. 5 shows the vanilla neural network model's training and validation accuracy. Training accuracy continues to increase, but validation accuracy hardly changes. Since the validation loss begins to increase and validation accuracy begins to decrease after epoch 5, the model trained with epoch 5 was chosen. The chosen model's validation accuracy at epoch 5 was about 85%, and test accuracy was 84.6%, which is slightly lower than validation accuracy.

Next, in case of the LSTM-based model, the text pre-processing after word tokenization was done as follows. The part-of-speech information is not necessary for the LSTM-based model, because LSTM neural network considers word order, which requires all words to be input in the network. Instead, a cleaning process of removing inappropriate words, such as symbols, was performed. The maximum length of the word was adjusted to 5 to bring together meanings that can be dispersed across multiple words that are not divided by spaces. That is, the meaning of the words that are clumped together due to wrong word spacing were checked by taking only the first part of the words. Then, the lengths of all sentences were adjusted equally to 25 words using the *pad_sequence*() function so that the input sentences all have the same length. By using the *Tokenizer* class of the *tensorflow.keras.preprocessing.text* module, words were tokenized into a form suitable for use in the LSTM-based neural network. The vocabulary was defined with the 20 000 most frequent occurrences. A review sentence was represented by 25-dimension vector, each element representing the index of a word in the vocabulary. Each word was embedded with 300 dimensions using the *Embedding* layer. A word vector reflecting the relationship between words was generated through word embedding learning using the *Embedding* layer, enabling sentiment analysis reflecting the meanings of the words.

Like the number of hidden units in the vanilla neural network, the LSTM-based model has 50 units. Also, like the vanilla model, the network was made to have only one LSTM layer. The model gives two outputs, one representing positive probability and the other representing negative probability. The softmax activation function was used to represent the positive and negative probabilities on each output. Tab. 3 presents the structure of the LSTM-based neural network model.

Table 3 LSTM-based neural network's structure/training parameters

Structure			Training Parameters	
Layer	Node #	Param #	Optimizer	RMSprop
Embedding	25 * 300	6 000 000	Loss function	sparse categorical crossentropy
LSTM	50	70 200	Batch size	128
Output	2	102	Validation data	20%

Fig. 6 shows the changes in the LSTM-based neural network model's training loss over 5 epochs, which took about 6 minutes. The LSTM-based model's training time takes much longer because the number of training parameters is about 30 times. The figure shows that the training loss continues to decrease overall, but validation loss increases after epoch 1, indicating that the training is not progressing properly.

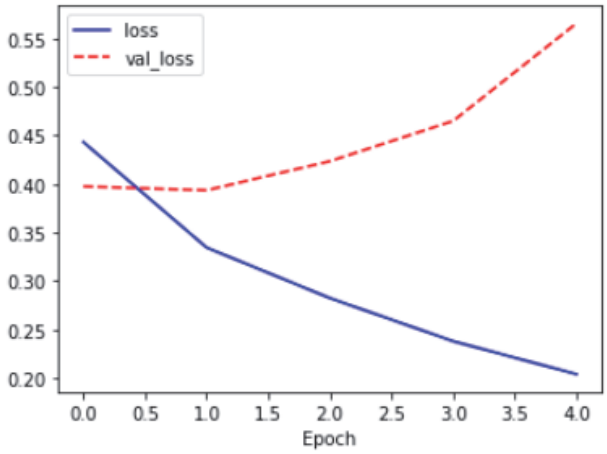

Figure 6 Training loss of the LSTM-based neural network model

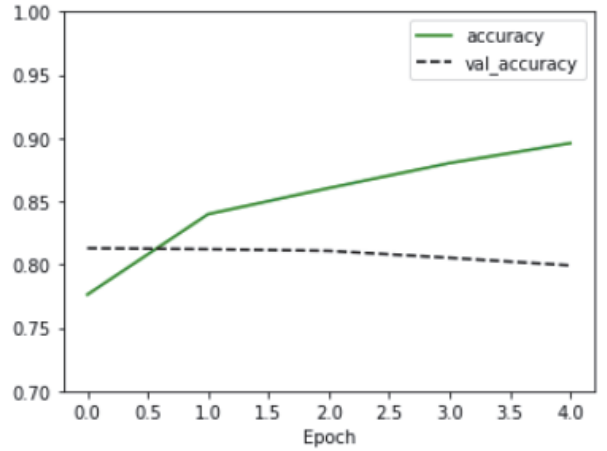

Figure 7 Training accuracy of LSTM-based neural network model

Fig. 7 shows the model's training and validation accuracies. Although the LSTM-based model's training accuracy is similar to that of the vanilla network model, its validation accuracy was about 81% at epoch 1 and 79.8% at epoch 5. Not only is the LSTM-based model's validation accuracy lower than that of the vanilla neural network model, but it also continues to decrease, which means that overfitting may occur. The model trained with epoch 1 was

chosen at this time, whose test accuracy was 79%, which is lower than that of the chosen vanilla model.

Another LSTM-based model with one output using the sigmoid function was also trained with different hyper-parameters for comparison. Tab. 4 shows the structure of the second LSTM-based neural network model. Because the processes for text pre-processing and embedding followed that applied to the previous LSTM-based model, the total number of parameters was almost the same, however, other optimizers and cost functions were applied.

Table 4 LSTM-based neural network (second)'s structure/training parameters

Structure			Training Parameters	
Layer	Node #	Param #	Optimizer	Adam
Embedding	25 *300	6 000 000	Loss function	binary crossentropy
LSTM	50	70 200	Batch size	128
Output	1	51	Validation data	20%

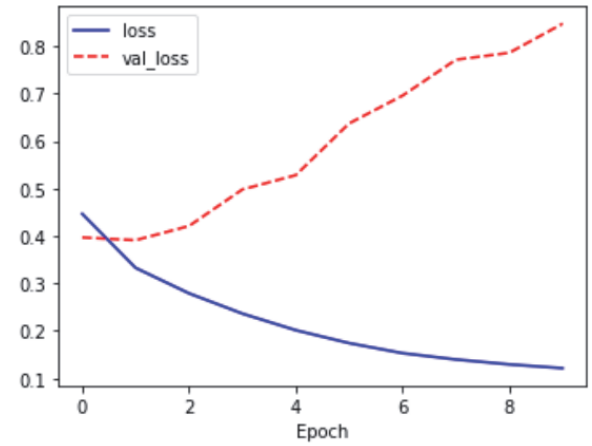

Figure 8 Training loss of the second LSTM-based neural network model

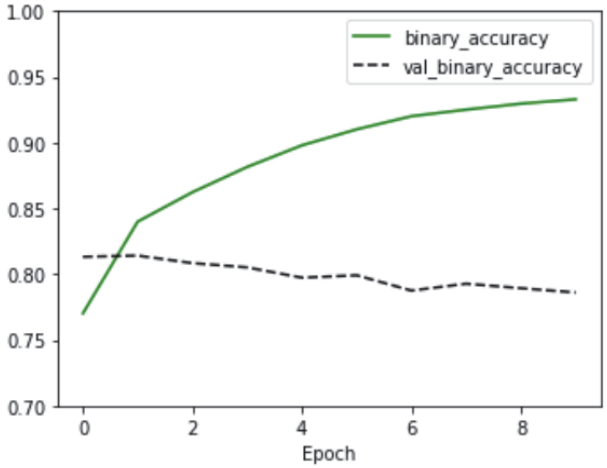

Figure 9 Training accuracy of the second LSTM-based neural network model

Figs. 8 and 9 show the changes in the second LSTM-based model's training loss and training/validation accuracies. In this case, training continued over 10 epochs, which took about 10 minutes. Better results may be expected by performing learning for a longer period of time. The training loss continues to decrease, ultimately to a lower value than the previous models, but validation loss continues to increase to a higher value. The second LSTM-based model's validation accuracy was 78.6%, and test accuracy was 78%, which was lower than expected.

From the above experiments, it can be judged that a simple neural network is sufficient for handling the relatively simple problem of classifying the given text into positive and negative. Considering the accuracy and training time, the vanilla neural network model trained over 5 epochs was selected as the proposed SDAP's model for sentiment analysis.

5.2 Fitness Centre Counselling

People go to the fitness centre and consult with experts to receive recommendations on appropriate exercises. Also, after exercising, counselling is conducted to evaluate how the customer performed the exercises and establish an effective exercise plan. The proposed SDAP can be used to support fitness centre counsellors in recommending appropriate exercises to customers based on the counselee's emotions and thoughts during the consultation. The counsellor can upload the recorded counselling conversation to the SDAP and perform the "Sentiment Analysis" to check the counselee's current exercise habits or basic attitude toward exercising.

When the counsellor clicks on "Sentiment Analysis" on the SDAP's user interface, the SDAP's sentiment analysis module in Fig. 1 receives the text output from the speech recognition module. Then, the sentiment analysis module encodes the words in each sentence of the text using the BOW encoding method. The encoded words in the sentences are input to the sentiment analysis model described in Section 5.1.2. The applied vanilla neural network model outputs the probability that the input sentence is positive. The positive probabilities of all of the sentences in the text are averaged to determine the overall probability that the text is positive. Eq. (1) gives the probability that the text is positive.

$$p_{\text{positive}}(D) = \frac{\sum_{s_i \in D}^{n} p_{\text{positive}}(s_i)}{n} \tag{1}$$

In Eq. (1), D is a given text (text file from the speech recognition module), s_i is a sentence in the text, $p(s_i)$ is the probability of a sentence s_i from the sentiment analysis model, and n is the total number of sentences in D.

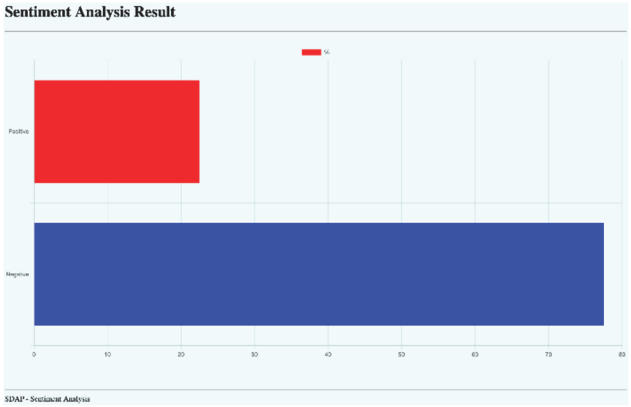

Figure 10 An example of sentiment analysis output (fitness centre consultation)

Fig. 10 shows an example of a sentiment analysis performed using the proposed SDAP on a counselling conversation at a fitness centre conducted after the

counselee (interviewee) finished exercising. The counsellor asked several questions about the exercise, such as how the exercise was, whether it was hard, and whether it was fun.

From the result presented in Fig. 10, the counsellor can easily identify that the counselee has negative emotions (77.5%) about the exercise he/she just finished. Based on this result, the counsellor can encourage the counselee and suggest exercise plans that are lower in intensity or reduce the duration of exercises.

5.3 Dental Clinic Counselling

Each dental patient at a dental clinic requires personalized treatment based on their individual conditions. When a person visits a dentist for the first time, a consultation takes place during which questions are asked to determine what kind of treatment is needed. The proposed SDAP can be used to analyse the counselling conversation and provide the frequently appearing keywords that can help the counsellor to discern the patient's condition and recommend appropriate treatment. Fig. 11 shows an example of a keyword analysis performed using the proposed SDAP on a counselling conversation conducted at a dental clinic.

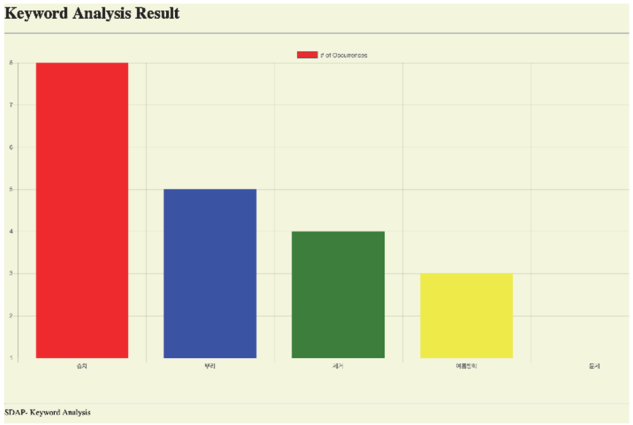

Figure 11 An example of keyword analysis output (dental clinic consultation)

As can be seen from Fig. 11, the word "cavity (in Korean, *chungchi*)" appeared the most, and words such as "root (*ppuli*)" and "removal (*jegeo*)" were also identified as keywords. Thus, the counsellor can easily judge from the keyword analysis results that the counselee needs caries treatment.

6 CONCLUSION

When counsellors are required to hold numerous consultations with customers, clients, or patients, they may find it difficult to provide efficient counselling or give adequate feedback over time. This paper proposed a SDAP composed of speech recognition and text analysis modules for analysing voice-based conversations to facilitate counsellors' understanding of counselees' individual situations. The SDAP's speech recognition module uses the Naver's CLOVA Speech service to ensure the latest performance, while the text analysis module was developed based on the Korean text analysis environment using Anaconda, NLTK, KoNLPy, MeCab morpheme analyser, and Scikit-Learn library. Text analysis is necessary to understand the content of the text. At present, the proposed SDAP provides keyword and sentiment analyses of counselling conversations that visually shows the counselee's psychology or state during the conversation without requiring the counsellor to think over the conversation. Thus, our SDAP enables counsellors to provide the proper feedback consistently and easily.

Further research and developments can be made to enhance the proposed SDAP, largely in terms of improving the accuracy of speech recognition, improving keyword analysis performance, and providing additional text analysis methods. The SDAP's speech recognition module depends on Naver's CLOVA Speech service, but post-processing for identifying words that are recognized incorrectly by speech recognition and correcting those mistakes can improve the SDAP's accuracy, especially that of the keyword analysis function. Providing a user dictionary can be another method to improve the SDAP's speech recognition. Text analysis, on the other hand, requires basic natural language processing (such as sentence tokenization and word tokenization) as well as in-depth language processing using machine learning). The Korean text analysis environment built for the SDAP is conducive to developing and adding new functions for processing Korean text, so the SDAP can be functionally enhanced by providing more text analysis options, such as text classification, intention identification, and named entity extraction.

Currently, the counsellors need to upload the voice recordings of counselling conversations to the SDAP, but a voice input module will be installed in the near future so that conversations could be directly sent to the SDAP using a microphone, eliminating the need to record conversations separately. Thus, the proposed SDAP is expected to become a genuine counselling support system for various areas requiring customer consultations.

Acknowledgements

This research was financially supported by Hansung University.

7 REFERENCES

[1] Fernandez, J. M. O. (2017). *Spoken Dialogue Systems: Architectures and Applications*. Doctoral Thesis.

[2] Chavan, R. S. & Sable, G. S. (2013). An Overview of Speech Recognition Using HMM. *International Journal of Computer Science and Mobile Computing*, *2*(6), 233-238.

[3] Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/ICASSP.2016.7472621

[4] Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/ICASSP.2017.7953152

[5] Black, A. W., Zen, H., & Tokuda, K. (2007). Statistical Parametric Speech Synthesis. *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing*. https://doi.org/10.1109/ICASSP.2007.367298

[6] Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2016). A Survey on Neural Speech Synthesis. https://doi.org/10.48550/arXiv.2106.15561

[7] Choi, H. S., Joo, S. H., Kim, D. C., Park, Y. C., Yeom, S., & Choo, H. S. (2016). Intelligent Classification and Context Analysis System of Voice Data. *The 23th KIPS Fall Conference*.
[8] Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V., & Wu, Y. (2020). Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition. *Neur IPS SAS 2020 Workshop.* https://doi.org/10.48550/arXiv.2010.10504
[9] Lee, S. K. & Kim, S.-D. (2021). Spoken Dialogue Analysis System for Supporting Effective Counseling. *Journal of Science and Engineering Management*, *3*(3), 1-8. https://doi.org/10.33832/jsem.2021.3.3.01
[10] Amazon EC2. https://aws.amazon.com/en/ec2
[11] Flask web development, one drop at a time. https://flask.palletsprojects.com
[12] CLOVA Speech. https://www.ncloud.com/product/aiService/clovaSpeech
[13] CLOVA Speech Recognition (CSR). https://www.ncloud.com/product/aiService/csr
[14] NLTK Documentation. https://www.nltk.org
[15] KoNLPy: Korean NLP in Python. https://konlpy.org
[16] Scikit-Learn, Machine Learning for Python. https://scikit-learn.org/stable
[17] CLOVA Speech API documentation. https://api.ncloud-docs.com/docs/ai-application-service-clovaspeech-clovaspeech
[18] ANACONDA. https://www.anaconda.com
[19] Java Downloads (Java SE Development downloads). https://www.oracle.com/java/technologies/downloads/
[20] JPype1 1.3.0. https://pypi.org/project/JPype1/
[21] Kudo, T. (2005). MeCab: Yet another Part-of-Speech and Morphological Analyzer.
[22] Huang, H., Wang, X., & Wang, H. (2020). NER-RAKE: An improved rapid automatic keyword extraction method for scientific literatures based on named entity recognition. *Proceedings of the Association for Information Science and Technology*, *57*(1), e374. https://doi.org/10.1002/pra2.374
[23] Dostal, M. & Je˘zek, K. (2011). Automatic Keyphrase Extraction based on NLP and Statistical Methods. *Proceedings of the Dateso 2011: Annual International Workshop on Databases, Texts, Specifications and Objects*, 140-145.
[24] Ying, Y., Qingping, T., Qinzheng, X., Ping, Z., & Panpan, L. (2017). A Graph-based Approach of Automatic Key phrase Extraction. *Procedia Computer Science*, *107*, 248-255. https://doi.org/10.1016/j.procs.2017.03.087
[25] Zhang, K., Xu, H., Tang, J., & Li, J. (2006). Keyword Extraction Using Support Vector Machine. *Advances in Web-Age Information Management*, *4016*. https://doi.org/10.1007/11775300_8
[26] Tensuan, J. P. & Azcarraga, A. (2013). Neural Network Based Keyword Extraction Using Word Frequency, Position, Usage And Format Features. *Research Congress 2012 De La Salle University*.
[27] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805
[28] Qian, Y., Jia, C., & Liu, Y. (2021). Bert-Based Text Keyword Extraction. *Journal of Physics: Conference Series*, *1992*. https://doi.org/10.1088/1742-6596/1992/4/042077
[29] Solangi, Y. A., Solangi, Z. A., Aarain, S., Abro, A., Mallah, G. A., & Shah, A. (2018). Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis. *The 5th International Conference on Engineering Technologies and Applied Sciences*.
[30] Kang, H., Yoo, S. J., & Han, D. (2012). Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, *39*, 6000-6010.
[31] Fersini, E., Messina, E., & Pozzi, F. (2017). Sentiment analysis: Bayesian ensemble learning. *Decision Support Systems*, *68*(2014), 26-38.
[32] Ruz, G. A., Henriquez, P. A., & Mascareno, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. *Future Generation Computer Systems*, *106*, 92-104. https://doi.org/10.1016/j.future.2020.01.005
[33] Dang, N. D., Moreno-Garcia, M. N., & la Prieta, F. D. (2020). Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics*, *9*(3).
[34] Zhang, Y. & Wallace, B. C. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. *International Joint Conference on Natural Language Processing.* https://doi.org/10.48550/arXiv.1510.03820
[35] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751. https://doi.org/10.3115/v1/D14-1181
[36] Kim, G. Y. & Lee, C. K. (2016). Korean Movie Review Sentiment Analysis Using Convolutional Neural Network. *Proceeding of the Korea Computer Congress*, 747-749.
[37] Araque, O., Corcuera-Platas, I., Sanchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, *77*, 236-246. https://doi.org/10.1016/j.eswa.2017.02.002
[38] Tensorflow 2.0. https://tenfsorflow.org
[39] Naver Movie Rating Data. https://github.com/e9t/nsmc/
[40] Mehta, P. & Pandya, S. (2020). A Review Sentiment Analysis Methodologies, Practices and Applications. *International Journal of Scientific & Technology Research*, *9*(12), 601-609.
[41] Mehta, P. & Pandya, S. (2020). A Review on Sentiment Analysis Methodologies, Practices and Applications. *International Journal of Scientific & Technology Research*, *9*(12), 601-609. https://doi.org/10.1016/j.eswa.2020.114155
[42] Abualigah, L., Alfar, H. E., Shehab, M., & Abu Hussein, A. M. (2019). Sentiment Analysis in Healthcare: A Brief Review. *Studies in Computational Intelligence*, *874*, 129-141. https://doi.org/10.1007/978-3-030-34614-0_7
[43] Bag-of-words model. https://en.wikipedia.org/wiki/Bag-of-words_model

Contact information:

Seok Kee LEE, Associate Professor
(Corresponding author)
School of Computer Engineering, Hansung University,
Samseongyo-ro 16gil 116, Seongbuk-gu, Seoul, Republic of Korea
E-mail: seelee@hansung.ac.kr

Sung-Dong KIM, Professor
School of Computer Engineering, Hansung University,
Samseongyo-ro 16gil 116, Seongbuk-gu, Seoul, Republic of Korea
E-mail: sdkim@hansung.ac.kr