

Analysis of depression in social media texts through the Patient Health Questionnaire-9 and natural language processing

Digital Health
Volume 8: 1–17
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2055207622114204
journals.sagepub.com/home/dhj



Nam Hyeok Kim¹, Ji Min Kim², Da Mi Park², Su Ryeon Ji¹ and Jong Woo Kim³

Abstract

Objective: Although depression in modern people is emerging as a major social problem, it shows a low rate of use of mental health services. The purpose of this study was to classify sentences written by social media users based on the nine symptoms of depression in the Patient Health Questionnaire-9, using natural language processing to assess naturally users' depression based on their results.

Methods: First, train two sentence classifiers: the Y/N sentence classifier, which categorizes whether a user's sentence is related to depression, and the 0–9 sentence classifier, which further categorizes the user sentence based on the depression symptomology of the Patient Health Questionnaire-9. Then the depression classifier, which is a logistic regression model, was generated to classify the sentence writer's depression. These trained sentence classifiers and the depression classifier were used to analyze the social media textual data of users and establish their depression.

Results: Our experimental results showed that the proposed depression classifier showed 68.3% average accuracy, which was better than the baseline depression classifier that used only the Y/N sentence classifier and had 53.3% average accuracy.

Conclusions: This study is significant in that it demonstrates the possibility of determining depression from only social media users' textual data.

Keywords

Depression, Patient Health Questionnaire-9, natural language processing, machine learning, deep learning, social media

Submission date: 3 October 2021; Acceptance date: 30 June 2022

Introduction

Depression is a disease that threatens the mental health of modern people and is recognized as a problem that needs to be solved, but there is a lack of understanding and agreement on the proper treatment for depression. Depression leads to a decline in the functions of daily life, with its main symptoms of being demotivated and feeling sad or unhappy.¹ The World Health Organization (WHO) expects depression to be the most burdensome disease for humans in 2030, with more than 264 million people in the world suffering from it.² WHO also notes that, globally, depression is a major cause of disability and contributes to

the burden of disease on people.² According to the (US) National Institute of Mental Health (NIMH), 17.3 million people, accounting for 7.1% of the adult population in the United States, have had at least one major depressive

¹Department of Mathematics, Hanyang University, Seoul, Republic of Korea

²Business Administration, Hanyang University, Seoul, Republic of Korea

³School of Business, Hanyang University, Seoul, Republic of Korea

Corresponding author:

Jong Woo Kim, School of Business, Hanyang University, 222, Wangsimni-ro, Seongdong-gu, Seoul 04763, Republic of Korea.

Email: kjw@hanyang.ac.kr.



episode.³ Additionally, 3.2 million American teenagers, which account for 13.3% of the population between the ages of 12 and 17 years, suffer from the same symptoms.³ What is more concerning is that 35% of adults and 60% of adolescents in the United States who suffer from depression are not receiving proper treatment, even though depression occurs in various age groups.³ As such, the mental health status of modern people has emerged as a major social problem and has begun to be perceived as a problem that can no longer be ignored.

Despite concerns about depression, there is no definitive treatment for mental illness in many countries. In 2016, Mental health service utilization rates for those diagnosed with mental illness were 43.1% in the United States, 46.5% in Canada, 34.9% in Australia, 35.5% in Spain, and 22.2% in Korea,⁴ indicating that mental health service utilization is significantly lower than 50% worldwide. According to Andrade et al.,⁵ the three main reasons for the low utilization of mental health services are low perceived need, structural barriers, and attitudinal barriers. Low perceived need is the lack of awareness of mental health issues and means the patient himself/herself thinks no help is needed. Structural barriers represent concerns about money, lack of time, accessibility, insurance coverage, etc. Attitudinal barriers include the idea that the mental disorder will improve itself, prejudice against mental health services, and distrust of treatment effects, the result of which is that patients fail to use the service. Unlike with physical diseases, patients suffering from mental illness often do not understand the extent of the disease, often do not receive treatment due to low motivation, and often do not know that problems can be improved by using mental health services. In general, mental illness is fully treatable by early intervention, but the later the treatment, the more serious the disorder can be,⁶ so proper awareness and early detection of mental illness are important and necessary steps in treating the disease. Furthermore, recognizing the disease and knowing its exact name increases the probability of early detection and enhances positive treatment effects.⁷

One of the important steps in treating depression is correct self-awareness of this condition. Self-diagnosis of depression allows people to check their degree of depression themselves, and there are many self-diagnosis tests for depression. Examples of self-diagnosis instruments for depression include the Beck Depression Inventory (BDI),⁸ the Center for Epidemiologic Studies Depression Scale (CES-D),⁹ the Patient Health Questionnaire-9 (PHQ-9),¹⁰ and the Geriatric Depression Scale (GDS).¹¹ There are various self-diagnosis tables for depression, but it is difficult for people with mental disabilities to identify their condition through this method for the same reasons as the low utilization rate of mental health services. Thus, an alternative could be a system that automatically (without specific patient involvement) identifies depression levels in such patients.

In fact, there have been many attempts to predict or detect depression through various techniques.^{12–18} In recent years, the expansion of social media such as Twitter and Facebook has raised interest in automatic depression detection techniques.¹² As social media has become an integral part of modern life, much data is produced, suggesting that considerable textual data are available for mental health analysis. This can be seen as a valuable resource for depression and mental disability assessment through text, the direction our research aims to pursue. Existing studies on text analysis for depression or mental disorder^{12–16} were conducted by establishing a classifier to determine whether the text is related to symptoms of depression, and to further assess the degree of concern for depression. Sentence classification has been carried out through techniques such as naïve Bayes classification (NBC), latent Dirichlet placement (LDA), support vector machine (SVM), and logistic regression, by building vocabularies with relevant experts. In particular, in a study conducted by Yazdavar et al.,¹³ by whose work we were most inspired, the PHQ-9 was used as a text classification criterion. In our study, the performance of sentence classification was improved by applying a better natural language processing (NLP) method to that preceding study, and the resulting sentence classification was expanded to a model that can judge depression based on the sentence classification results. Therefore, the purpose of our study was to associate textual data with the nine symptoms of depression in the PHQ-9 through NLP techniques, and to identify users' depression based on their results.

The remaining sections are organized as follows. Related research section introduces the diagnosis of depression and various depression self-diagnosis tables. We also analyze the applications of NLP in health care and reference prior studies on online depression detection. Depression classification model section introduces our depression classification model and describes the model details and mechanisms. Experiments section describes the experiments to evaluate our model and analyze its results. The academic and practical significance of this study are described in Discussion section, and Conclusion section concludes with a summary of key findings and future research proposals.

Related research

Diagnosis of depression

The diagnosis of depression usually occurs as follows: psychiatrists identify symptoms through consultation with the patients, conduct the necessary tests, and comprehensively analyze and diagnose the results; diagnostic assessment tools are used in this process. In real-world clinical situations, the use of appropriate diagnostic assessment tools can significantly help diagnosis because depression

symptoms can be overlooked if they are not clear.¹⁹ The most widely used diagnostic evaluation tools are the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5)²⁰ and the *10th International Classification of Diseases* (ICD-10).^{21,22}

DSM-5 is published by the American Psychological Association (APA) and has 20 categories of mental illness classification, providing comprehensive diagnostic criteria for mental illness, including major depressive disorders. According to DSM-5, the symptomatic criteria for diagnosis of major depressive disorders include (1) depression, (2) reduction of interest, (3) eating disorders, (4) sleep disorders, (5) impatience, (6) fatigue, (7) self-blame, (8) decreased concentration, and (9) thoughts of self-harm or suicide. If five symptoms (including either (1) or (2)) persist for 2 weeks, depression is diagnosed. ICD-10 is an international disease classification system developed by WHO in 1994 and updated regularly, which deals with mental and behavioral disorders in Chapter 5. ICD-10 is similar to DSM-5 in that it assesses the severity of depression according to the number of symptoms.²³

In general, depression screening tools are used prior to precise clinical diagnosis. For basic epidemiological studies in psychiatry, simple depression screening tools are used to classify depression symptoms into positive and negative groups before precise diagnostic testing is performed, which reduces the human and economic burden.²⁴ Examples of depression screening tools are the BDI,⁸ SDS,²⁵ CES-D,⁹ PHQ-9,¹⁰ and GDS.¹¹

The PHQ-9 is a self-reported test developed in 1990 by Robert L. Spitzer et al., which focuses on major depression and evaluates its severity.¹⁰ The questionnaire consists of nine questions asking about various symptoms of depression, such as depressed emotions, appetite, and suicidal thoughts, and calculates the score by evaluating how often the symptoms have occurred in the past 2 weeks. The score analysis criteria are— 0–4 points are not depression, 5–9 points are mild depression, 10–14 points are moderate depression, 15–19 points are treatment-needed depression, and 20–27 points are severe depression that requires active treatment.

In a study comparing widely used depression screening tools in primary care,²⁶ the SDS and BDI were compared with the PHQ-9. The PHQ-9 has proven to be a stable and reasonable tool for measuring depression, with 88% sensitivity and 88% specificity compared to other scales. It is also considered easy to apply, taking less time to complete, and being easier to score than conventional depression screening tools.²⁷ In this study, a version of the PHQ-9 translated into Korean and standardized was used; the reliability and validity of this version have been verified by previous studies.^{28,29}

NLP

In pre-deep learning NLP studies, sentence classification through machine learning was mainly based on NBC,

SVMs,³⁰ and the random forest classification algorithm. With deep learning architectures and algorithms making significant contributions in the field of computer vision and pattern recognition, studies on deep learning-based NLP have begun to emerge.³¹ Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are representative deep learning techniques that are applied to various fields such as image recognition, object tracking, and autonomous driving, as well as NLP. Although CNNs were originally used as a key technology in image analysis, studies of applying CNNs to NLP have begun since Collobert and Weston's 2008 research³² and research by Collobert et al. in 2011,³³ and many have published results that CNN also excels in text processing.^{34–41} Unlike traditional artificial neural networks wherein the signal flow is directed only in the output direction, RNNs are utilized for sequential data because the signal flow has a circular structure and can have information about the past. RNNs have proven to be an effective method for statistical language modeling⁴² and has also been shown to be suitable for many NLPs, such as language modeling, machine translation, and speech recognition.³¹

Long short-term memory (LSTM)⁴³ is a technique that complements the limitations of gradient vanishing in RNNs and is actively used in NLP research with RNNs.^{44–47} Recently, in addition to RNNs and LSTM, pre-trained language models such as bidirectional encoder representations from transformers (BERT),⁴⁸ generative pre-trained transformer (GPT)-2,⁴⁹ and GPT-3⁵⁰ have been used in NLP research.

Application of artificial intelligence in the medical field

With the development of artificial intelligence (AI), the development of AI-based medical technologies using information and communication technology (ICT) convergence technologies and medical “big data” is actively underway.⁵¹ AI-based medical technology reduces uncertainty in clinical decision making and can analyze and process vast amounts of medical data to provide customized medical services for patients and medical staff. Currently, there are three areas of medical technology based on AI: electronic health records (EHRs) and medical data, medical and pathological imaging, and physiological signal monitoring.

Unlike conventional paper records, the application of AI to EHRs and medical data refers to the electronic storage of all medical information concerning a patient's past and present health status and health care.⁵² With the introduction of EHRs in hospitals and clinics, the role of AI as a tool for extensive clinical care and research has become important.⁵³ EHRs are characterized as being vast, heterogeneous, incomplete, noisy, and generated for purposes other than research, so studies that apply NLP to extract relevant medical information from these vast data troves

are actively underway.^{54–59} EHR studies based on NLP have shown that NLP can be applied by extracting meaningful medical information from unstructured data such as clinical data. With the development of social media, NLP will become even more widely used and applied to extract meaningful information from more diverse textual data.

The second application is AI in medical and pathological imaging. In medical imaging, AI technology can be used for classification, image segmentation, automatic detection of lesions in computer-assisted detection methods, and simulated imaging.⁶⁰ Examples of these studies include AI-based detection of cancer spread to lymph nodes in women with breast cancer,⁶¹ brain hemorrhage verification,⁶² skin cancer classification,⁶³ and the use of FDA-approved AI medical devices for diagnosis of breast cancer, stroke, and brain hemorrhage, as well as cardiac ultrasonic diagnosis and MRI heart analysis. AI increasingly supports various medical image analyses and diagnoses based on its consistency, certainty, and accuracy.⁶⁴

AI is also applied in physiological signal monitoring, which refers to obtaining signal information from a user through sensors attached to or worn on the body for medical purposes such as real-time response and 24-hour monitoring.⁶⁵ Physiological signals are meaningful data sources that help detect, treat, and rehabilitate diseases because they reflect the electrical activity of certain body parts; AI, such as deep learning, can be designed to utilize these signals to make health care decisions.⁶⁵ For example, monitoring of menstrual signals has already been commercialized and is used in hospital environments such as intensive care units, emergency rooms, and hospital rooms. Physiological signal monitoring is also applied to predict and prevent the worsening of disease courses that have been actively underway for years, and its utilization is expanding.⁶⁶ With the introduction of various AI technologies into clinical practice, there is a need for early thorough and systematic clinical verification,⁶⁰ as well as to compare imprecise medical data to standard data to create conditions to train AI.⁶⁷

Detection of depression from social media texts

Communication between doctors and patients is important in the diagnosis of depression, but patients are often reluctant to visit a hospital or clinic themselves because they resist accepting their depression. Due to these limitations, efforts are also being made to identify barriers to the use of mobile apps for depression treatment.⁶⁸ Recently, as technology has advanced, attempts have been made to detect depression based on electroencephalograms (EEGs), deep learning, and machine learning in various applications in the medical field.⁶⁹ For example, there have been attempts to predict depression using machine learning random forest, k-nearest neighbor algorithm, or

SVM.⁷⁰ In addition, the expansion of social networks such as Facebook, Twitter, and Imgur has increased interest in detecting depression using them.^{69,71} As social networks have become an integral part of modern life, tremendous numbers of posts are now generated in real time, which has led to attempts to analyze people's mental health conditions such as emotions and depression based on these textual data. Arguably, language reflects human thought, emotion, belief, behavior, and personality,⁷² so the generation of this abundant textual data has triggered research on the possibility of depression detection as well as emotional analysis through NLP. Examples of this are studies that parse out sentences written on social media to identify word-to-word similarities,¹² or that classify sentences into depression-related issues and then categorize them according to criteria in a self-diagnosis table.⁷³ Assuming that people with depression will tweet on their Twitter accounts and show symptoms of depression, studies have also been conducted to establish related vocabulary dictionaries to detect those symptoms.¹³ Furthermore, there have been attempts to use social networks to distinguish between suicidal ideation and depression.⁷⁴

While there is ample opportunity to detect depression in social media, to date there have been no significant, clear criteria established for emotional analysis. Therefore, this study created a model to determine whether social media-indicated depression conforms with actual depression self-diagnosis tables, which constitute a clear classification standard. In addition, the model was created with logistic regression, which can determine the relative effects of each variable. In addition to depression self-diagnosis tables, a variety of other diagnostic tables can be used to create a model specific to the symptoms if the model is trained based on the process in this study.

Depression classification model

The development process of a new depression classification model

We herein propose a three-step process for detecting and analyzing social media users' depression. In Figure 1, each of the three stages has two parts: training the model used for depression detection and then predicting depression using the model trained in the previous step. The "sentence classifier training phase" (SCT) and the "depression classifier training phase" (DCT) are in the first training phase, and the "user's depression classification phase" (UDC) is in the prediction phase. First, we train two sentence classifiers in the SCT phase. Two sentence classifiers are used to classify sentences based on symptoms of depression, the Y/N classifier, and the PHQ-9. The Y/N classifier determines whether or not a sentence is related to depression, and the 0–9 classifier determines, based on the symptoms given in the PHQ-9, which question(s) of the PHQ-9

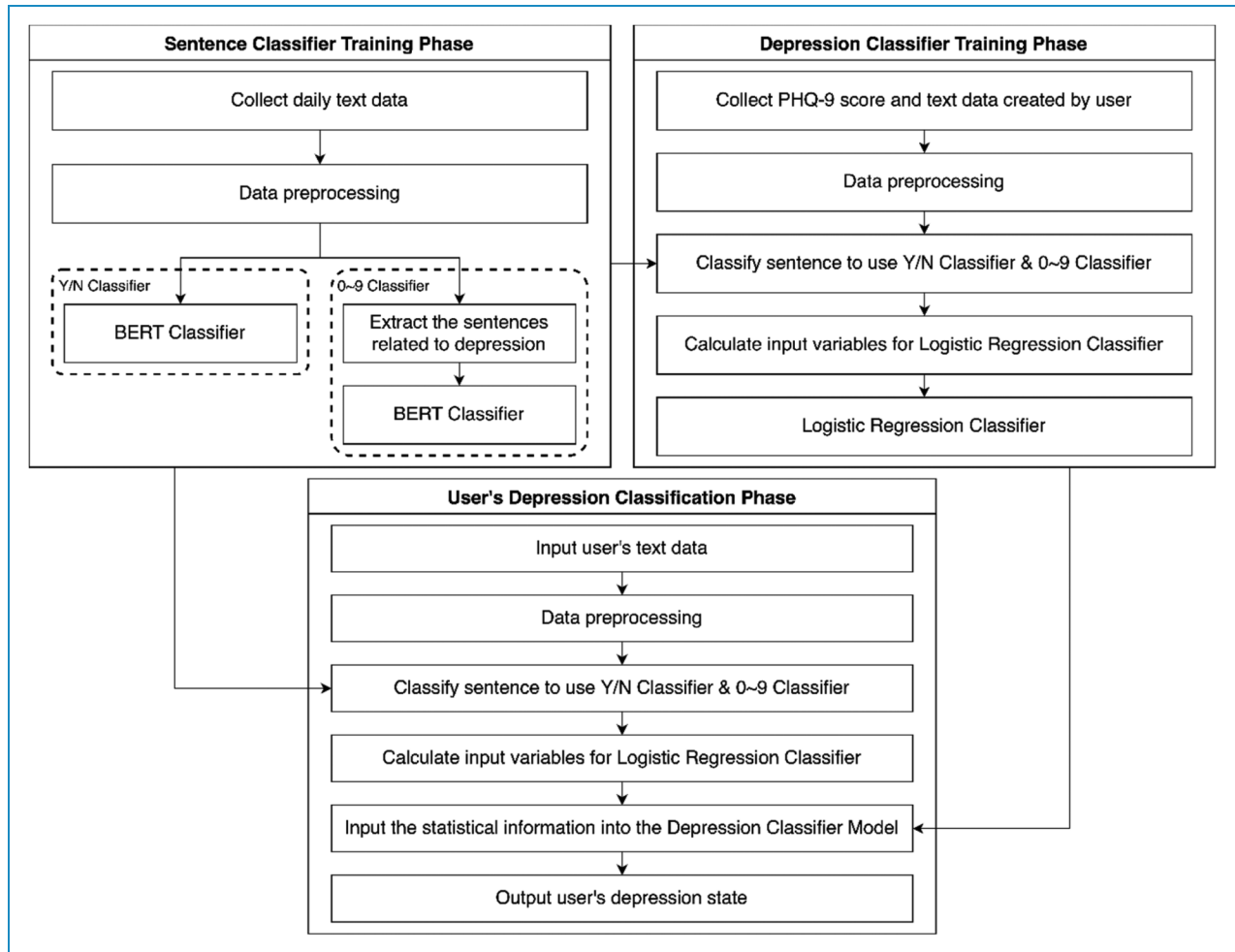


Figure 1. The entire process of training and using the depression classifier.

is/are related to the sentence. The 0–9 classifier classifies a sentence into one of 10 categories that correspond to each symptom in the PHQ-9, with 0 being a class that is not related to one of the PHQ-9 symptoms.

Next, in the DCT phase, a logistic regression classifier is trained to determine whether or not a user is depressed. To train the logistic regression classifier, a set of user-generated social media text data and a sample user's PHQ-9 score are necessary. The target variable of the logistic regression classifier is the likelihood that the user is depressed. The likelihood of depression is 1 if the user's PHQ-9 score is ≥ 5 ; otherwise it is 0. Finally, the UDC phase uses the previously trained sentence classifier and the depression classifier to predict one user's depression.

Sentence classifier training phase

To train sentence classifiers, social media texts that describe daily life were collected from the Internet. The collected text data were separated into sentences, which were then preprocessed by removing stop words and spell checking.

Then three people who received data labeling training read each sentence individually and assigned a Y/N label, that is, a "Y" when the sentence was related to depression, and "N" when it was not. For each sentence related to depression ("Y"), a label(s) of "0" through "9" were assigned for each of the PHQ-9 symptom(s) it reflected. If the assigned labels were different among the three people, the final labels were assigned through discussion among them.

To train the sentence classifiers, BERT, Word2Vec, and Unicode word embedding methods were used, and the applied classifiers were NBC, SVM, RNN, LSTM, CNN, and BERT. Finally, the model with the highest accuracy was selected as the final model.

Depression classifier training phase

To train the depression classifier, user PHQ-9 scores and social media text data for 2 weeks were necessary. We collected social media text data from 30 adults who, based on their PHQ-9 scores, were judged to have depression and 30

adults who were not. The collected textual data were pre-processed in the same way as in the SCT phase, and were then classified using the trained Y/N classifier and the 0–9 classifier. The ratio of each user's classified sentences was calculated based on the number of sentences classified Y/N and 0–9 and the total number of sentences the user had written. Then a logistic regression classifier was trained with depression as a dependent variable and the ratio of each label (Y/N and 0–9) as an independent variable. To determine the final logistic classifier, statistically significant coefficients were selected stepwise based on the variability inflation factor (VIF) and p -value for each variable. Due to the small number of users (60 adults), we performed five-fold cross-validation to improve the reliability of the results.

User's depression classification phase

Each user's social media text data for 2 weeks was used for input. This data was preprocessed the same way as in the SCT phase, and were then classified using the trained Y/N classifier and the 0–9 classifier. Based on those classification results, input variables of the logistic regression classifier were calculated. Through the logistic classifier, the user was categorized as being depressed or not.

Experiments

Experimental designs

The experiment was approved by Hanyang University Institutional Review Board (IRB) and the approval number was HYUIRB-202008-001. Participants were informed of the detailed experimental purpose and procedure, and written consent was obtained.

Training the sentence classifier based on depression symptoms: The sentences to be labeled were collected from the representative Korean blog sites Naver Blog, Naver Cafe, and Daum Cafe. Sentences not related to depression were collected through the daily Naver Blog, and sentences deemed to be depressed were collected through Naver's and Daum's depression-related cafes.

For each user posting, we collected the user ID, URL, upload time, title, and content. We collected 23,115 documents and separated them into sentences using the Python library Korean Sentence Splitter (KSS). The total number of collected sentences was 249,103. The collected data were labeled in two steps. First, the sentences are labeled based on whether they are related to (Y) or not (N) with depression. When a sentence was labeled Y, the second labeling indicated which of the nine symptoms of the PHQ-9 corresponded to the sentence. We added a 0 label for those Y sentences that did not correspond to one of the PHQ-9's 1–9 symptoms. According to the above rules, it was labeled independently by three workers who had basic knowledge of NLP and learned PHQ-9. Each

worker read the given sentence, labeled Y/N according to whether it was related to depression, and if Y, labeled it in a category between 0 and 9 according to the criteria of PHQ-9. When there exists inconsistency in the labeled results, it was labeled as ground truth according to the majority vote between the three workers for the same sentence. If all workers were labeled differently, the ground truth was determined through discussion between the three workers.

Once the collected data were labeled, we found there was a significant data imbalance: there were significantly fewer sentences reflecting depression than those that did not. A more severe imbalance was found in sentences pertaining to the PHQ-9 symptoms: among the sentences tagged Y for depression, very few were labeled with symptoms other than 0, 1, or 9. To resolve these data imbalances, under-sampling was performed on the sentences that were not related to depression (tagged N). The details of the final dataset after under-sampling are shown in Table 1.

Training the depression classifier: For this experiment, we recruited blog users who had written daily online articles in the 2-week selection period. A total of 60 adults over the age of 19 were selected, 30 who were considered depressed and 30 who were not. Whether a person was currently

Table 1. Number and proportion of sentences according to depression and the Patient Health Questionnaire-9 (PHQ-9) symptoms.

Label Y/N	The PHQ-9 label 0–9	Count	Percentage	
Y	0	2156	15.07%	49.99%
	1	2132	14.90%	
	2	534	3.73%	
	3	1358	9.49%	
	4	685	4.79%	
	5	809	5.65%	
	6	1986	13.88%	
	7	2027	14.17%	
	8	591	4.13%	
	9	2030	14.19%	
Total		14,308	100.00%	
N		14,313	50.01%	
Total		28,621	100.00%	

experiencing depression was determined based on their PHQ-9 test results (≥ 5 = depressed) acquired in the application stage. Since PHQ-9 diagnoses 5 points as mild depression, we also determined depression based on 5 points. And, to ensure the reliability of the PHQ-9 results, we administered a second PHQ-9 test three days after the first one. Considering that PHQ-9 diagnoses depression based on symptoms of 2 weeks, we collected all textual data from the users' blogs over the 2 weeks prior to the PHQ-9 test date. The data collected in this way were divided into sentence units and preprocessed and organized by experimenter.

These data were used to train and evaluate the logistic regression classifier with fivefold cross-validation. First, a baseline depression classifier using only the Y/N classifier was created, without the 0–9 classifier, to compare the performance of the proposed logistic regression classifier. Then, after training each logistic regression classifier, the accuracy of the two models was compared.

The reason for using fivefold cross-validation here is that the number of data was too small. We used cross-validation to overcome this and ensure reliability of the model's

performance. Also, we used fivefold cross-validation instead of commonly used 10-fold cross-validation to secure enough test dataset. If the 10-fold cross-validation was used, the size of the test dataset would have been too small at 6, but by using fivefold cross validation, we obtain relatively sufficient size of 12. The main experiment was conducted with fovefold cross-validation, and experiments on 10-fold cross-validation and 3-fold cross-validation were also conducted. The results of the experiment are presented in the Appendix.

Experimental results

Performance of the sentence classifier: To find the best-performing sentence classifiers, we conducted experiments with various embeddings and classification algorithms. NBC, SVM, RNN, LSTM, BiRNN, and BiLSTM classifiers were trained with Word2Vec embedding, and CNN 1D and CNN 2D were trained with Unicode embedding. In the case of BERT, the classifier was implemented by adding a linear layer to the last layer of KoBERT, a Korean BERT released by SKT. As shown in Figure 2,

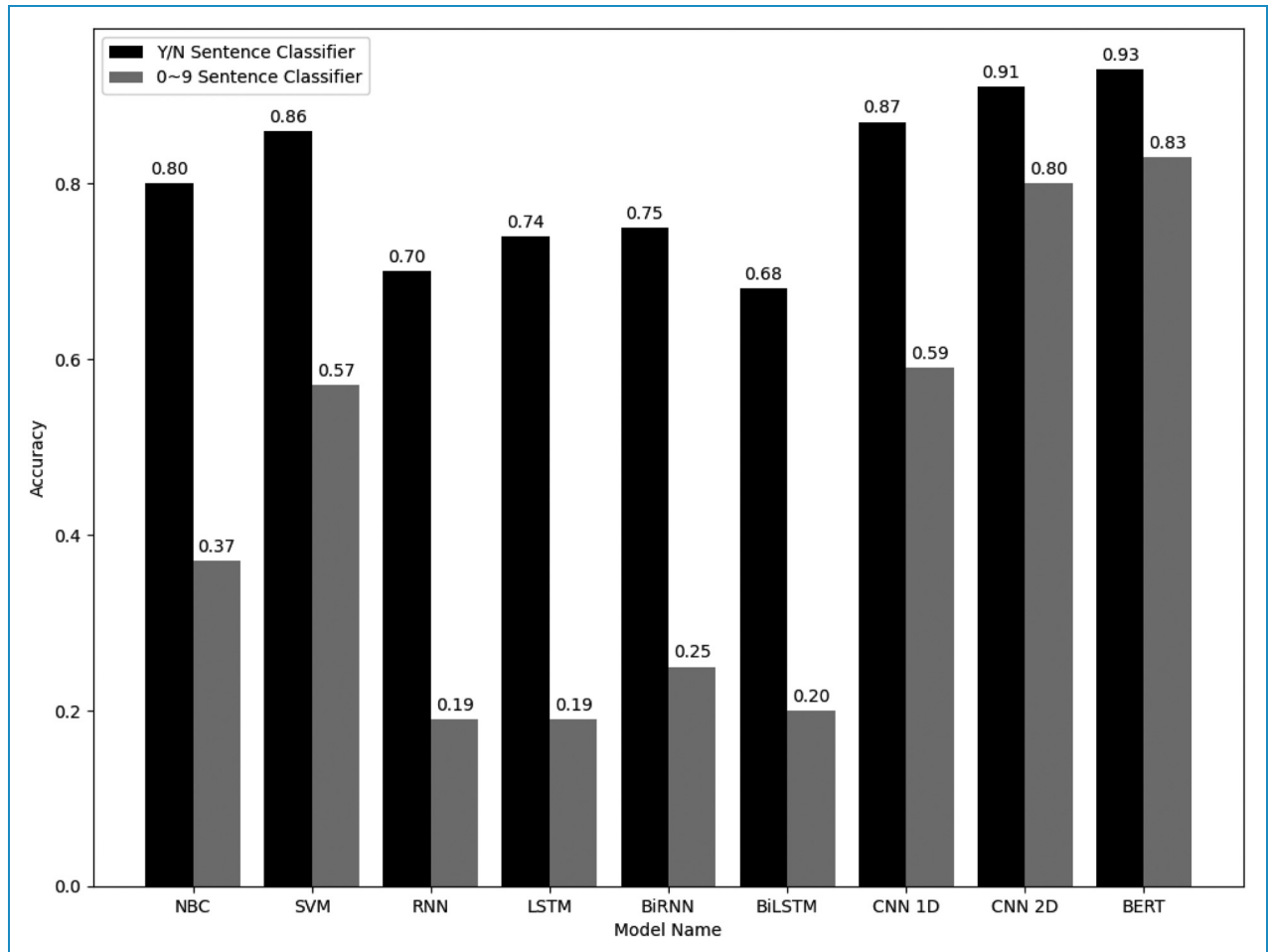


Figure 2. Performance comparison of the Y/N and 0–9 sentence classifiers according to algorithm.

Table 2. Performance of bidirectional encoder representations from transformers (BERT)-based Y/N sentence classifier.

Y/N sentence classifier			
Class	Precision	Recall	F1-score
N	0.96	0.91	0.93
Y	0.92	0.96	0.94
Accuracy			
0.9368			

the experimental results showed that BERT classifiers were the best for both Y/N and 0–9 sentence classification.

In Table 2, the accuracy of the BERT-based Y/N sentence classifier was 93.68%. The precision of N was 96%, which was greater than the precision of Y, while the recall of Y was 96%, which was greater than that of N (91%).

In Table 3, the accuracy of the BERT-based 0–9 sentence classifier was 83.29%. However, because the accuracy of the Y/N sentence classifier was 93.68%, the actual accuracy was $93.68\% \times 83.29\% = 78.02\%$.

In addition, when viewed through Figure 3, the F1-scores were different among the symptoms because this is influenced by how distinct the symptom-specific features are. In the case of symptom 0, various symptoms were mixed and they did not correspond to numbers 1 through 9, so the characteristics were not clear, resulting in the lowest F1-score. Symptom 1 related to depression, symptom 2 related to reduction of interest, and symptom 5 related to psychomotor agitation or retardation also had low F1-scores because they were difficult to extract from textual data. On the other hand, F1-scores were high for symptom 3, which related to significant weight loss, symptom 4 related to insomnia, and symptom 9 related to thinking about or attempting suicide or death, because they were relatively easy to identify from textual data.

Performance of the depression classifier: The baseline depression classifier had two variables: S (the number of sentences) and Ratio_D (the ratio of the sentences related to depression to the total sentences). When viewed through Table 4, among five folds, Ratio_D was selected three times after variable selection. Even though in the two cases, Ratio_D was not selected as a significant variable, if the number of data is sufficient, Ratio_D can certainly be chosen as a significant variable.

On the other hand, the proposed depression classifier had three variables: S (the number of sentences), Y (the number of depression-acknowledged sentences), and Ratio_n (the number of sentences classified as nth symptoms/total

Table 3. Performance of the 0–9 sentence classifier.

0–9 sentence classifier			
Class	Precision	Recall	F1-score
0	0.73	0.69	0.71
1	0.77	0.79	0.78
2	0.67	0.73	0.79
3	0.90	0.96	0.93
4	0.94	0.97	0.95
5	0.80	0.82	0.81
6	0.89	0.81	0.85
7	0.86	0.85	0.86
8	0.82	0.89	0.85
9	0.90	0.92	0.91
Accuracy			
0.8329			

number of sentences). Variable selection steps are performed for five folds. The results show that Ratio_1, Ratio_2, Ratio_3, and Ratio_6 are significant variables. The coefficients of Ratio_1 and Ratio_6 are positive, and that of Ratio_2 and Ratio_3 are negative. That is, higher proportion of sentences on symptoms 1 and 6 among the entire sentences increases the probability of depression. On the contrary, higher proportion of sentences of symptoms 2 and 3 reduce the chance of depression. It can also be seen that the absolute values of the Ratio_2's coefficients are significantly larger than that of other variables, which means the ratio of sentences on symptom 2 is more sensitive than that of other symptoms.

The average accuracy of the proposed depression classifier was 68.3%, which was 15% higher than that of the baseline depression classifier (53.3%). In Figure 4, for all cases in the fivefold cross-validation, the accuracies of the proposed depression classifier were always higher than those of the baseline depression classifier. Therefore, the user's depression could be more accurately classified when adding label-specific ratios obtained through the 0–9 sentence classifier, rather than only using the Y/N sentence classifier.

Discussion

This study aimed to determine whether a user's depression can be predicted based on text written on social media. Our

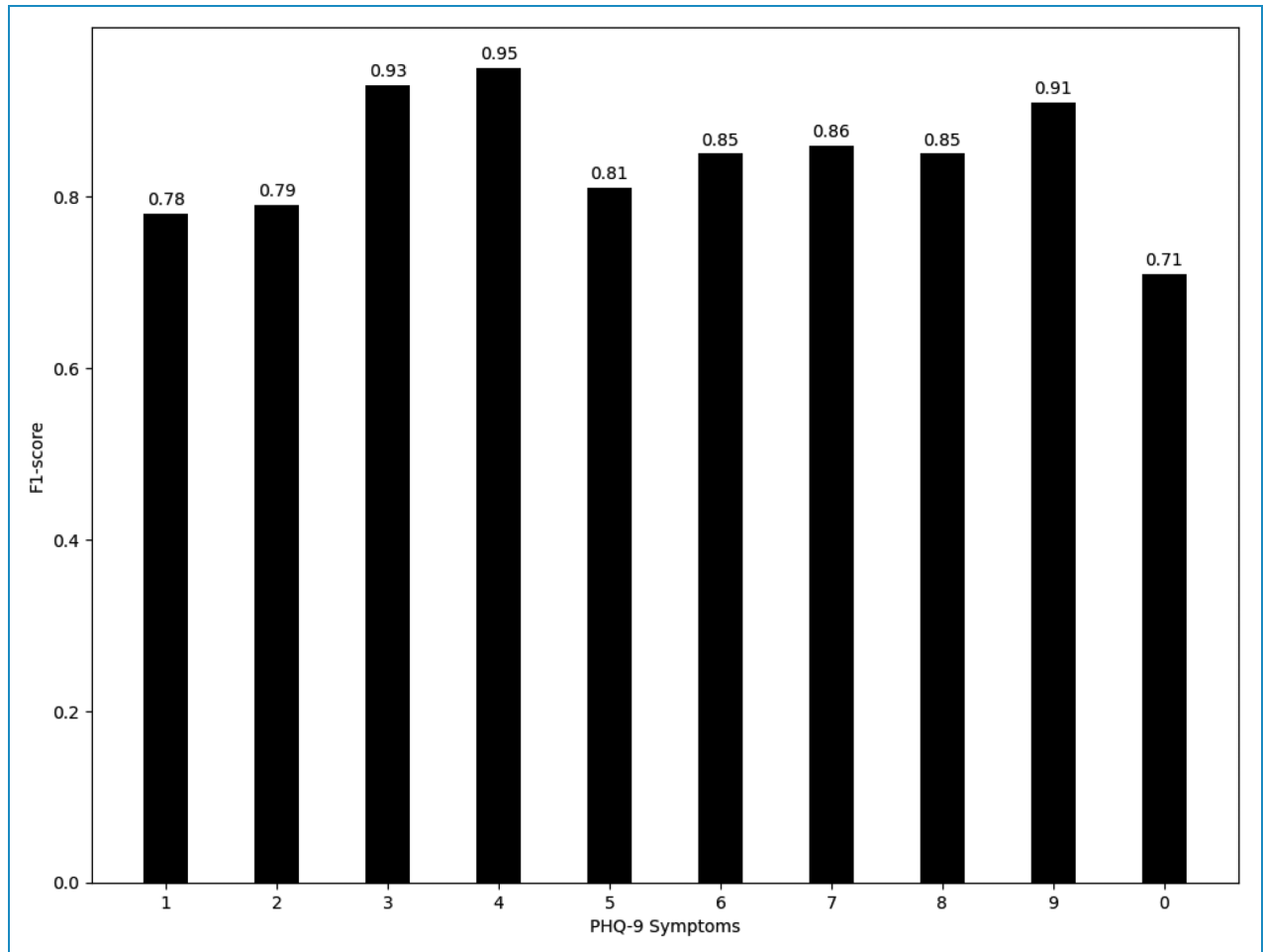


Figure 3. F1-scores of the 0-9 sentence classifier.

study results indicated that this is possible, using NLP and machine learning techniques. This study contributes to early depression identification, which is a significant step in the treatment of depression. And the methodology described here can be applied without the conscious participation of the user.

There are currently many mental health online applications (“apps”) that can automatically analyze users’ emotions and detect mental disorders, and the model proposed here can be included in many of them. In the case of mental illness, it is important to constantly scan for mental conditions and get professional help to prevent mental deterioration. Therefore, our model can be used for mental health care services and apps for people suffering from mental illness. Although it is necessary that some users provide their PHQ-9 scores and their social media text for the training purpose, after the training phase, the classifiers can be used to determine whether other users are depressed or not solely based on their social media text.

In addition, if there were systematic disease indicators for various diagnoses, more diverse mental disorders could be analyzed online in similar ways. For example, in this study, we used the PHQ-9 but we might also be able to create other models using BDI,⁸ SDS,²⁵ CES-D,⁹ and GDS.¹¹ In addition to depression, self-diagnoses of panic disorder, anxiety disorder, stress, bipolar disorder, etc. can be established from social media texts.

This study simply tries to classify whether a user has depression or not. However, in future research, we can extend our model by combining various technologies, which will be more helpful for the early detection of depression and preventing it from worsening. Another possible future avenue is “explainable artificial intelligence” (XAI), which is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms.⁷⁵ If the development of technology allows us to identify the causes of depression through XAI, it would

Table 4. Logistic regression results of baseline and proposed depression classifier by fivefold cross-validation.

Baseline depression classifier					Proposed depression classifier			
K-fold	Variable	Estimate	Std. Error	Pr (> t)	Variable	Estimate	Std. Error	Pr (> t)
1	Intercept	0.3800	0.0963	2.71e-4	Intercept	0.449	0.079	1.07e-6
	Ratio_D	3.0651	1.6601	0.0712	Ratio_1	15.219	6.328	0.0204
					Ratio_2	-61.892	37.617	0.1069
2	Intercept	0.3593	0.1034	0.0011	Ratio_1	20.172	7.287	0.0082
					Ratio_2	-56.379	29.164	0.0598
	Ratio_D	3.0429	1.6253	0.0675	Ratio_3	-30.373	16.583	0.0732
					Ratio_6	20.586	11.849	0.0894
3	Intercept	0.5000	0.0729	1.35e-8	Ratio_1	26.240	9.821	0.0106
					Ratio_2	-78.131	29.953	0.0125
	Ratio_D	0.5000	0.0729	1.35e-8	Ratio_3	-34.988	17.794	0.0557
					Ratio_6	24.628	11.798	0.0428
4	Intercept	0.5000	0.0729	1.35e-8	Ratio_1	21.768	7.997	0.0093
					Ratio_2	-63.253	32.416	0.0575
	Ratio_D	0.5000	0.0729	1.35e-8	Ratio_3	-33.660	16.825	0.0517
					Ratio_6	27.675	12.397	0.0308
5	Intercept	0.3775	0.1021	5.78e-4	Ratio_1	21.261	7.994	0.0109
					Ratio_2	-67.880	34.380	0.0548
	Ratio_D	2.7240	1.6182	0.0990	Ratio_3	-36.262	18.687	0.0589
					Ratio_6	25.842	14.083	0.0734

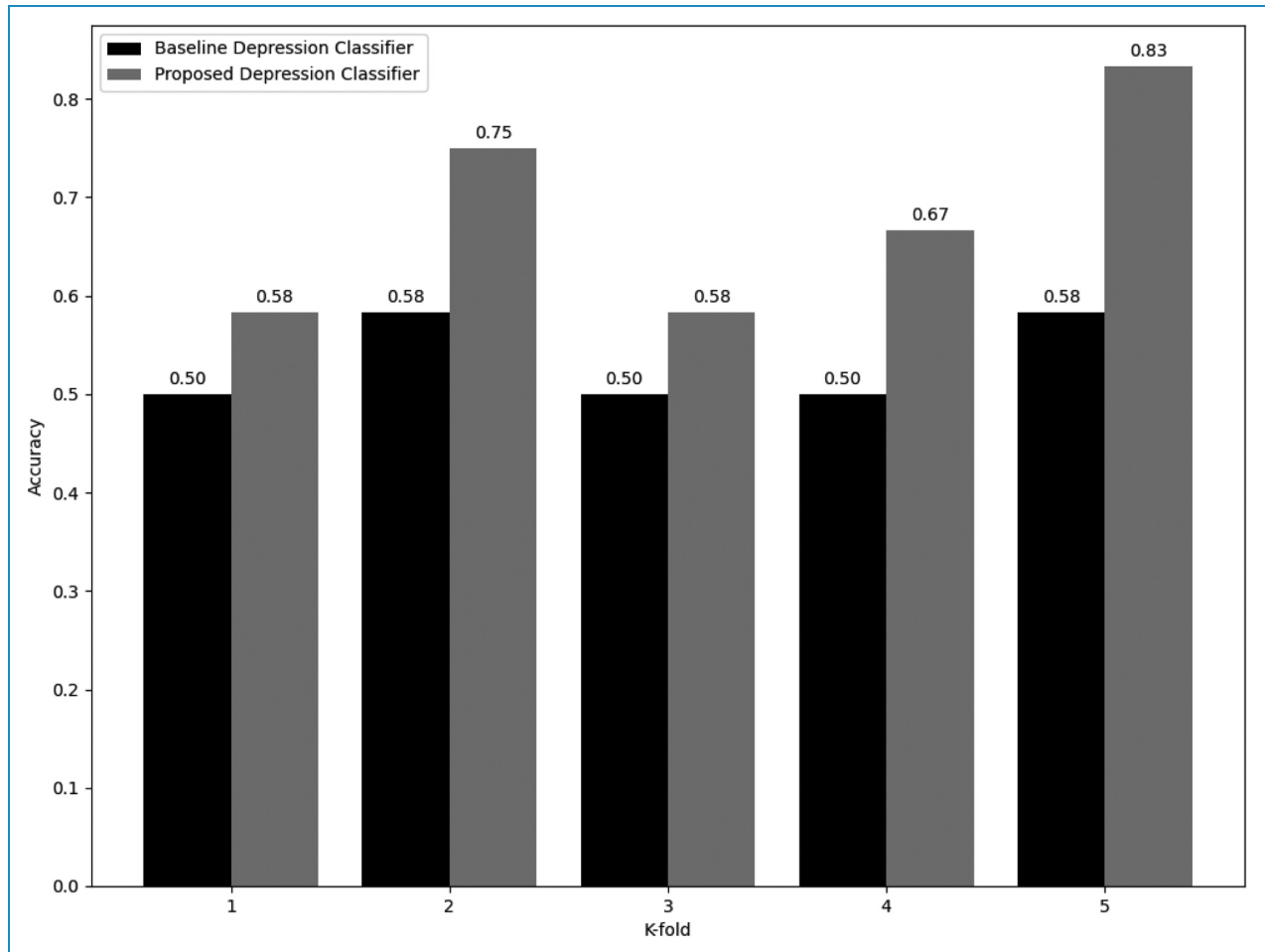


Figure 4. Comparison of the accuracy of the baseline and the proposed depression classifier by fivefold cross-validation.

contribute to the improvement of mental health through customized treatment and emotional management.

Conclusion

In this study, we created a model to determine whether or not social media users are depressed, by analyzing their past social media texts. The proposed model consists of three classifiers: the Y/N sentence classifier which determines whether or not a text sentence is related to depression, the 0–9 sentence classifier which classifies a text sentence according to the depression symptomology in the PHQ-9, and the Depression classifier, which ultimately establishes whether or not a social media user is potentially depressed. To improve the sentence classification accuracy, we tried various text classification algorithms; among them, BERT-based classifiers showed the best performance for both the Y/N and 0–9 sentence classifiers. In particular, the accuracy of the sentence classifier of Yazdavar et al.,¹³ which is the basis of this paper, was 68%, whereas our sentence classifier showed 83.29% accuracy,

which is approximately 15% higher performance. Of course, since it was not compared with the same dataset, it is difficult to compare the performances directly. In addition, it is necessary to verify the proposed approach with other data sets. However, currently, there are no available open depression data sets to perform such verification. Lastly, the depression classifier, which is a logistic regression classifier, also showed that sentence classification based on the PHQ-9 is helpful to improve prediction accuracy.

The most significant limitation of this study was that the social media textual data of only 60 users were used in the Depression classifier's training. To overcome this, fivefold cross-validation was performed, but with more data, it would have been possible to train the model more stably and achieve certain results without having to use k-fold cross-validation. There was also a limitation in that the proposed model is a binary classification of whether or not a user is depressed. Finally, the contribution of this paper at the methodological level is limited because the main purpose of this study is to improve the performance of

the Depression Classifier on users' social media text data by applying the state-of-the-art NLP techniques. Although there was a significant improvement in the performance of the depression classifier from the proposed approach, future studies can improve the performance using emerging advanced techniques.

Acknowledgments: The authors would like to thank Seok Hyeon Kim, MD, PhD and Kounseok Lee, MD, PhD at the Department of Neuropsychiatry, Hanyang University Medical Center for their advice of medical perspectives in this research.

Contributorship: The study was conceived and conducted with collaboration of five authors. The study was under the guidance of JWK, and NHK led the study. Five authors designed the research together and NHK, JMK, DMP, and SRJ implemented machine learning classifiers and performed experiments. NHK wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: The Hanyang University Institutional Review Board (IRB) approved this study (approval number: HYUIRB-202008-001).

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: 'Business Laboratory Project Semester' under University Innovation Support Project at Hanyang University.

Guarantor: JWK.

ORCID iD: Jong Woo Kim  <https://orcid.org/0000-0001-9484-5968>

References

1. Depressive disorder. <http://www.snulh.org/health/nMedInfo/nView.do?category=DIS&medid=AA000353> (2020, accessed 6 June 2020).
2. Depression. <https://www.who.int/news-room/fact-sheets/detail/depression> (2020, accessed 12 July 2020).
3. Major Depression. <https://www.nimh.nih.gov/health/statistics/major-depression.shtml> (2019, accessed 11 September 2020).
4. Park EA. Health promotion research brief. *Korea Health Promotion Institute* 2017; 8: 8–9.
5. Andrade LH, Alonso J, Mneimneh Z, et al. Barriers to mental health treatment: results from the WHO world mental health surveys. *Psychol Med* 2014; 44: 1303–1317.
6. Hong JP. The survey of mental disorders in Korea. *Samsung Medical Center* 2016; 391: 1–501.
7. Hong HS, Kim SJ, Song JH, et al. A survey on the mental health knowledge of general adults in Korea. *National Mental Health Center* 2008; 1: 2–3.
8. Beck AT, Ward CH, Mendelson M, et al. An inventory for measuring depression. *Arch Gen Psychiatry* 1961; 4: 561–571.
9. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. *Appl Psychol Meas* 1977; 1: 385–401.
10. Spitzer RL, Kroenke K and Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Jama* 1999; 282: 1737–1744.
11. Yesavage JA, Brink TL, Rose TL, et al. Development and validation of a geriatric depression screening scale: a preliminary report. *J Psychiatr Res* 1982; 17: 37–49.
12. Burdisso SG, Errecalde M and Montes-y-Gómez M. A text classification framework for simple and effective early depression detection over social media streams. *Expert Syst Appl* 2019; 133: 182–197.
13. Yazdavar AH, Al-Olimat HS, Ebrahimi M, et al. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, July 2017, pp. 1191–1198.
14. Guntuku SC, Yaden DB, Kern ML, et al. Detecting depression and mental illness on social media: an integrative review. *Curr Opin Behav Sci* 2017; 18: 43–49.
15. De Choudhury M, Gamon M, Counts S, et al. Predicting depression via social media. In Seventh international AAAI conference on weblogs and social media, June 2013.
16. Aldarwish MM and Ahmad HF. Predicting depression levels using social media posts. In 2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS) March 2017, pp. 277–280. IEEE.
17. Stasak B, Epps J and Goecke R. Automatic depression classification based on affective read sentences: opportunities for text-dependent analysis. *Speech Commun* 2019; 115: 1–14.
18. Esposito A, Esposito AM, Likforman-Sulem L, et al. On the significance of speech pauses in depressive disorders: results on read and spontaneous narratives. In: *Recent advances in nonlinear speech processing*. Cham: Springer, 2016, pp.73–82.
19. Jung SW, Lee EJ, Choi YK, et al. A guide of diagnostic evaluation for depression -focused on assessment instrument for depression. *The Korean Journal of Psychopathology* 2009; 18: 21–38.
20. American Psychiatric Association. *Depressive disorders: DSM-5® selections*. Washington, DC: American Psychiatric Pub, 2015.
21. World Health Organization. *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research (vol. 2)*. Geneva: World Health Organization, 1993.
22. Moon ES, Sakong JK, Jung SW, et al. Development of clinical guideline for the diagnosis and evaluation of depression : focused on diagnosis guideline. *J Korean Neuropsychiatr Assoc* 2014; 53: 15–23.
23. Lee MS. Current Status and diagnosis of recent depression. *Journal of Korean Society of Health-System Pharmacists* 2013; 30: 505–511.
24. Cho MJ, Bae JN, Suh GH, et al. Validation of geriatric depression scale, Korean version (GDS) in the assessment of DSM-III-R Major depression. *J Korean Neuropsychiatr Assoc* 1999; 38: 48–63.

25. Zung WW. A self-rating depression scale. *Arch Gen Psychiatry* 1965; 12: 63–70.
26. Kroenke K, Spitzer RL and Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001; 16: 606–613.
27. An JY, Seo ER, Lim KH, et al. Standardization of the Korean version of screening tool for depression(patient health questionnaire-9, the PHQ-9). *Korean Society of Biological Therapies in Psychiatry* 2013; 19: 47–56.
28. Choi HS, Choi JH, Park KH, et al. Standardization of the Korean version of patient health questionnaire-9 as a screening instrument for major depressive disorder. *Journal of the Korean Academy of Family Medicine* 2007; 28: 114–119.
29. Han C, Jo SA, Kwak JH, et al. Validation of the patient health questionnaire-9 Korean version in the elderly population: the ansan geriatric study. *Compr Psychiatry* 2008; 49: 218–223.
30. Cortes C and Vapnik V. Support-vector networks. *Mach Learn* 1995; 20: 273–297.
31. Young T, Hazarika D, Poria S, et al. Recent trends in deep learning based natural language processing. *IEEE Comput Intell Mag* 2018; 13: 55–75.
32. Collobert R and Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning, 2008, pp. 160–167.
33. Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. *J Mach Learn Res* 2001; 12: 2493–2537.
34. Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.
35. Kalchbrenner N, Grefenstette E and Blunsom P. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188, 2014.
36. Zhang Y and Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820, 2015.
37. Sharma AK, Chaurasia S and Srivastava DK. Sentimental short sentences classification by using CNN deep learning model with fine tuned Word2Vec. *Procedia Comput Sci* 2020; 167: 1139–1147.
38. Ruder S, Ghaffari P and Breslin JG. Insight-1 at semeval-2016 task 5: Deep learning for multilingual aspect-based sentiment analysis. arXiv preprint arXiv:1609.02748, 2016.
39. Severyn A and Moschitti A. Modeling relational information in question-answer pairs with convolutional neural networks. arXiv preprint arXiv:1604.01178, 2016.
40. Tu Z, Hu B, Lu Z, et al. Context-dependent translation selection using convolutional neural network. arXiv preprint arXiv:1503.02357, 2015.
41. Zhou Y, Xu R and Gui L. A sequence level latent topic modeling method for sentiment analysis via CNN based diversified restrict boltzmann machine. In 2016 International Conference on Machine Learning and Cybernetics (ICMLC), 2016, Vol. 1, pp. 356–361. IEEE.
42. Grachev AM, Ignatov DI and Savchenko AV. Compression of recurrent neural networks for efficient language modeling. *Appl Soft Comput* 2019; 79: 354–362.
43. Hochreiter S and Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9: 1735–1780.
44. Fu X, Liu W, Xu Y, et al. Long short-term memory network over rhetorical structure theory for sentence-level sentiment analysis. In *Asian Conference on Machine Learning* 2016; 63: 17–32.
45. Li D and Qian J. Text sentiment analysis based on long short-term memory. In 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI), 2016, pp. 471–475. IEEE.
46. Bao L, Lambert P and Badia T. Attention and lexicon regularized LSTM for aspect-based sentiment analysis. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, July 2019, pp. 253–259.
47. Luo X and Chen Z. English Text quality analysis based on recurrent neural network and semantic segmentation. *Future Gener Comput Syst* 2020; 112: 507–511.
48. Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.
49. Radford A, Wu J, Child R, et al. Ilya sutskever. *Language Models are Unsupervised Multitask Learners* 2019; 1: 1–9.
50. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners, 2020; 33: 1877–1901.
51. Lee YC. Artificial intelligence (AI)-based medical devices Status and issues (1). *Korea Health Industry Development Institute* 2018: 1–20.
52. Kirch W. Electronic health record (EHR). In: Kirch W (ed) *Encyclopedia of public health*. Dordrecht: Springer, 2008, pp. 46–46.
53. Juhn Y and Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol* 2020; 145: 463–469.
54. Datta S, Bernstam EV and Roberts K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *J Biomed Inform* 2019; 100: 1–15.
55. Savova GK, Danciu I, Alamudun F, et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. *Cancer Res* 2019; 79: 5463–5470.
56. Savova GK, Tseytlin E, Finan S, et al. DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records. *Cancer Res* 2017; 77: e115–e118.
57. Agaronnik N, Lindvall C, El-Jawahri A, et al. Challenges of developing a natural language processing method with electronic health records to identify persons with chronic mobility disability. *Arch Phys Med Rehabil* 2020; 101: 1739–1746.
58. Forsyth AW, Barzilay R, Hughes KS, et al. Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. *J Pain Symptom Manage* 2018; 55: 1492–1499.
59. Lindvall C, Lilley EJ, Zupanc SN, et al. Natural language processing to assess end-of-life quality indicators in cancer patients receiving palliative surgery. *J Palliat Med* 2019; 22: 183–187.
60. Park SH. Artificial intelligence in medicine: beginner's guide. *Journal of the Korean Society of Radiology* 2018; 78: 301–308.
61. Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* 2017; 318: 2199–2210.
62. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, et al. Advanced machine learning in action: identification of

- intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ digital Medicine* 2018; 1: 1–7.
63. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542: 115–118.
64. Jeong GH. Trends in artificial intelligence-based medical image analysis technology. *Weekly ICT Trends* 2018; 1863: 2–13.
65. Faust O, Hagiwara Y, Hong TJ, et al. Deep learning for healthcare applications based on physiological signals: a review. *Comput Methods Programs Biomed* 2018; 161: 1–13.
66. Lee YC. Artificial intelligence (AI)-based medical devices Status and issues (2). *Korea Health Industry Development Institute* 2018: 1–20.
67. Park JW. AI Healthcare: expect new high value-added services to be created. *Korea Institute of Science and Technology Information* 2016; 10: 1–6.
68. Mugica F, Nebot À, Bagherpour S, et al. A model for continuous monitoring of patients with major depression in short and long term periods. *Technol Health Care* 2017; 25: 487–511.
69. Xiaowei L, Xin Z, Jing Z, et al. Depression recognition using machine learning methods with different feature generation strategies. *Artif Intell Med* 2019; 99: 101696.
70. Byeon H. Developing a random forest classifier for predicting the depression and managing the health of caregivers supporting patients with Alzheimer's disease. *Technol Health Care* 2019; 27: 531–544.
71. Aldarwish MM and Ahmad HF. Predicting Depression Levels Using Social Media Posts. 2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS), 2016. DOI; 10.1109/ISADS.2017.41.
72. Tung C and Lu W. Analyzing depression tendency of web posts using an event-driven depression tendency warning model. *Artif Intell Med* 2016; 66: 53–62.
73. Schwartz HA and Ungar LH. Data-driven content analysis of social media: a systematic overview of automated methods. *Ann Am Acad Pol Soc Sci* 2015; 659: 78–94.
74. O'dea B, Wan S, Batterham PJ, et al. Detecting suicidality on twitter. *Internet Interv* 2015; 2: 183–188.
75. What is explainable AI? www.ibm.com/watson/explainable-ai. (2021, accessed 9 September 2021).

Appendix

Table A1. Logistic regression results of baseline and proposed depression classifier by threefold cross-validation.

Baseline depression classifier					Proposed depression classifier			
K-fold	Variable	Estimate	Std. Error	Pr (> t)	Variable	Estimate	Std. Error	Pr (> t)
1	Intercept	0.500	0.080	2.37e-7	Intercept	0.447	0.085	6.53e-6
					Ratio_1	15.759	6.064	0.013
					Ratio_2	-66.203	37.013	0.081
2	Intercept	0.326	0.111	0.005	Intercept	0.406	0.105	4.88e-4
					Ratio_1	20.623	7.953	0.013
	Ratio_D	3.714	1.706	0.035	Ratio_2	-65.304	29.789	0.035
					Ratio_3	-29.131	16.752	0.091
					Ratio_6	26.269	12.593	0.044
3	Intercept	0.500	0.080	2.37e-7	Intercept	0.363	0.111	0.0023
					Ratio_1	20.257	11.702	0.0922
					Ratio_2	-81.680	43.931	0.0714
					Ratio_3	-43.838	21.637	0.0504
					Ratio_6	37.081	15.827	0.0249

Table A2. Accuracy of baseline and proposed depression classifier by threefold cross-validation.

K-fold	1	2	3	Average
Baseline depression classifier	0.50	0.50	0.50	0.50
Proposed depression classifier	0.70	0.65	0.70	0.683

Table A3. Logistic regression results of baseline and proposed depression classifier by 10-fold cross-validation.

Baseline depression classifier					Proposed depression classifier			
K-fold	Variable	Estimate	Std. Error	Pr (> t)	Variable	Estimate	Std. Error	Pr (> t)
1	Intercept	0.371	0.094	2.64e-4	Intercept	0.408	0.088	2.65e-5
					Ratio_1	19.611	7.194	0.008
	Ratio_D	3.045	1.573	0.058	Ratio_2	-56.506	28.866	0.055
					Ratio_3	-26.151	16.321	0.115
					Ratio_6	20.434	11.728	0.087
2	Intercept	0.355	0.095	4.64e-4	Intercept	0.405	0.087	2.45e-5
					Ratio_1	20.959	7.055	0.004
	Ratio_D	3.381	1.599	0.039	Ratio_2	-69.264	27.726	0.015
					Ratio_3	-36.069	15.941	0.028
					Ratio_6	28.988	11.463	0.014
3	Intercept	0.388	0.095	1.59e-4	Intercept	0.411	0.088	2.47e-5
					Ratio_1	20.968	7.303	0.006
	Ratio_D	2.648	1.600	0.103	Ratio_2	-66.677	32.394	0.044
					Ratio_3	-31.352	16.221	0.059
					Ratio_6	24.542	12.778	0.061
4	Intercept	0.500	0.068	1.6e-9	Intercept	0.386	0.089	6.98e-5
					Ratio_1	22.385	7.441	0.004

(continued)

Table A3. Continued.

Table 7.67 continued.

Baseline depression classifier					Proposed depression classifier			
					Ratio_2	-64.583	31.222	0.043
					Ratio_3	-35.105	16.774	0.041
					Ratio_6	26.204	12.151	0.035
5	Intercept	0.381	0.095	2.04e-4	Intercept	0.3696	0.084	5.87e-5
					Ratio_1	11.627	6.531	0.081
	Ratio_D	2.881	1.629	0.082	Ratio_2	-58.225	27.590	0.039
					Ratio_6	27.298	12.053	0.026
6	Intercept	0.366	0.097	4.26e-4	Intercept	0.395	0.088	4.44e-5
					Ratio_1	20.764	7.025	0.004
	Ratio_D	3.024	1.593	0.063	Ratio_2	-63.907	27.663	0.025
					Ratio_3	-31.458	16.053	0.055
					Ratio_6	25.317	11.328	0.030
7	Intercept	0.500	0.068	1.6e-9	Intercept	0.441	0.089	9.04e-6
					Ratio_1	19.753	7.636	0.012
					Ratio_2	-70.562	28.502	0.016
					Ratio_3	-36.055	15.904	0.027
					Ratio_6	24.206	11.409	0.038
8	Intercept	0.500	0.068	1.6e-9	Intercept	0.434	0.075	4.87e-7
					Ratio_1	11.592	6.231	0.068
9	Intercept	0.500	0.068	1.6e-9	Intercept	0.449	0.075	2.3e-7
					Ratio_1	15.401	5.892	0.011
					Ratio_2	-65.636	29.208	0.029

(continued)

Table A3. Continued.

Baseline depression classifier					Proposed depression classifier			
10	Intercept	0.500	0.068	1.6e-9	Intercept	0.375	0.088	9.2e-5
					Ratio_1	14.537	6.835	0.038
					Ratio_2	-52.649	27.772	0.063
					Ratio_6	19.591	11.213	0.086

Table A4. Accuracy of baseline and proposed depression classifier by 10-fold cross-validation.

K-fold	1	2	3	4	5	6	7	8	9	10	Average
Baseline depression classifier	0.50	0.50	0.50	0.50	0.50	0.33	0.50	0.83	0.33	0.50	0.50
Proposed depression classifier	0.66	0.50	0.66	0.50	0.66	0.66	0.66	0.83	0.66	0.83	0.66