ORIGINAL ARTICLE
eISSN 2005-7571
Korean J Biol Psychiatry 2020;27(1):18-26
https://doi.org/10.22857/kjbp.2020.27.1.003

텍스트 분류 기반 기계학습의 정신과 진단 예측 적용

순천향대학교 서울병원 정신건강의학과
백두현 · 황민규 · 이민지 · 우성일 · 한상우 · 이연정 · 황재욱

Application of Text-Classification Based Machine Learning in Predicting Psychiatric Diagnosis

Doohyun Pak, MD, Mingyu Hwang, MD, Minji Lee, MD, Sung-Il Woo, MD, Sang-Woo Hahn, MD, Yeon Jung Lee, MD, Jaeuk Hwang, MD

Department of Psychiatry, Soonchunhyang University Seoul Hospital, Seoul, Korea

Objectives The aim was to find effective vectorization and classification models to predict a psychiatric diagnosis from text-based medical records.

Methods Electronic medical records (n = 494) of present illness were collected retrospectively in inpatient admission notes with three diagnoses of major depressive disorder, type 1 bipolar disorder, and schizophrenia. Data were split into 400 training data and 94 independent validation data. Data were vectorized by two different models such as term frequency-inverse document frequency (TF-IDF) and Doc2vec. Machine learning models for classification including stochastic gradient descent, logistic regression, support vector classification, and deep learning (DL) were applied to predict three psychiatric diagnoses. Five-fold cross-validation was used to find an effective model. Metrics such as accuracy, precision, recall, and F1-score were measured for comparison between the models.

Results Five-fold cross-validation in training data showed DL model with Doc2vec was the most effective model to predict the diagnosis (accuracy = 0.87, F1-score = 0.87). However, these metrics have been reduced in independent test data set with final working DL models (accuracy = 0.79, F1-score = 0.79), while the model of logistic regression and support vector machine with Doc2vec showed slightly better performance (accuracy = 0.80, F1-score = 0.80) than the DL models with Doc2vec and others with TF-IDF.

Conclusions The current results suggest that the vectorization may have more impact on the performance of classification than the machine learning model. However, data set had a number of limitations including small sample size, imbalance among the category, and its generalizability. With this regard, the need for research with multi-sites and large samples is suggested to improve the machine learning models.

Key Words Text-classification · Electronic medical record · Vectorization · Machine learning · Present illness · Psychiatric diagnosis.

Received: November 11, 2019 / **Revised:** December 31, 2019 / **Accepted:** March 9, 2020
Address for correspondence: Jaeuk Hwang, MD
Department of Psychiatry, Soonchunhyang University Seoul Hospital, 59 Daesagwan-ro, Yongsan-gu, Seoul 04401, Korea
Tel: +82-2-709-9959, **Fax:** +82-2-709-9938, **E-mail:** hju75@schmc.ac.kr

서 론

전자의무기록(electronic medical records)의 분석은 최근 활발하게 연구되고 있는 분야 중 하나로 개별 환자의 의무기록을 분석하여 질병의 위험군을 분류하거나 치료 경과를 예측하는 데 활용되고 있다.[1)2)] 이러한 연구들은 대부분 환자의 활력 징후 또는 검사 수치 등 객관적인 데이터의 변화를 분석하고 있으나 정신과의 진단은 임상적 관찰에 의해 기술된 비정형적인 설명적 텍스트 데이터(descriptive text data)를 통해 하기 때문에 진단을 표준화하는 것이 상대적으로 쉽지 않다.[3)]

비정형적인 텍스트 데이터로부터 유용한 정보를 추출해내

는 과정을 텍스트 마이닝(text mining)이라 하며 최근 이에 대한 연구 또한 활발히 진행되고 있다.[4] 텍스트 마이닝은 주제에 따라 문서를 나누는 텍스트 분류(text classification), 원하는 정보를 가진 문서를 찾는 정보 검색(information retrieval), 문서로부터 필요한 정보를 찾는 정보 추출(information extraction), 문서 작성자의 심리, 감정 상태 등을 파악하는 감정 분석(sentiment analysis) 등 여러 분야에서 활용되고 있다.[5]

최근의 의학 연구자들은 인공지능(artificial intelligence)을 활용하여 방대한 양의 의료 데이터를 컴퓨터에게 수학적인 방법으로 학습시켜 분석하고 있으며, 이를 통하여 진단 및 치료반응 예측을 시도하는 등 의료 전반에 인공지능의 활용도가 증가하고 있다.[6-8] 이러한 연구 중, 텍스트 마이닝에 활용되는 기계학습 기반의 연구들이 소개되고 있으며 최근에는 딥러닝(deep learning) 기반의 텍스트 분류가 영상판독 리포트 등 의료 데이터의 분류에도 좋은 성과를 내고 있다.[9][10] 정신과학 분야에서도 환자의 외래 내원 초진 기록을 분석하여 우울, 불안 등 11가지의 정신병리 상태를 추론하는 연구가 진행되어 63%의 예측률을 보이기도 하였다.[11]

정신질환은 다양한 검사 결과를 활용하는 내외과적 질환과는 다르게 병력 청취를 통해 환자의 상태를 기술한 의무기록을 바탕으로 진단을 한다.[12] 앞서 설명한 텍스트 분류 기법을 의무기록에 적용할 경우 각기 다른 진단명에 따른 분류를 시도할 수 있다. 이에 본 연구에서는 비정형적인 텍스트 데이터인 정신과 입원 환자의 현병력을 자연어 처리(natural language process) 및 벡터화(vectorization) 후 텍스트 분류 기반의 기계학습(machine learning)으로 분석하여 각기 다른 세 가지 정신질환에 대한 진단 분류 정확도를 평가하였다.

방 법

연구 대상

2008년 1월 1일~2018년 12월 31일까지 주요우울장애, 제1형 양극성 장애, 조현병으로 진단하 순천향대학교 서울병원 정신건강의학과에 처음 입원한 환자들의 전자의무기록을 대상으로 후향적 연구를 시행하였다. 국제 질병 분류(International Statistical Classification of Diseases and Related Health Problems) 10판의 한국번역 판인 한국표준질병·사인분류(Korean Standard Classification of Diseases 7)상 상기 진단에 해당하는 F200, F201, F202, F203, F203, F206, F208, F209, F250, F258, F259, F301, F311, F312, F313, F314, F315, F316, F318, F319, F321, F322, F323, F328, F329, F331, F332, F333, F339 진단 코드로 입원한 환자의 입원 초기 의무기록을 수집하였다. 정신건강의학과 전문의에 의하여 정신장애의 진단 및 통계 편람 제4판[13] 및 제5판[14] 기준으로 입원 당시 주요우울장애(n = 300), 제1형 양극성장애(n = 91), 조현병(n = 103)이 진단된 494명의 환자를 연구 대상으로 하였다. 표 1은 각 군의 대상자들의 사회인구학적 특징을 나타내었다. 진단의 객관성을 확보하기 위해, 입원 시 담당 전공의가 반구조화된 면담을 통해 작성한 현병력에 대해 4명의 전문의가 의견을 종합하는 논의 과정을 거쳤다. 이후 치료 경과 중에 담당 전문의가 가장 가능성이 높다고 판단한 진단명을 기계학습의 레이블(label)로 사용하였다. 세 가지 정신질환의 분류만을 목적으로 하였기 때문에 명시(specifier) 및 동반 이환(comorbidity)의 종류가 다를 경우 이에 대한 세부적인 구분은 시행하지 않았다. 문서의 평균 글자 수는 2904자, 표준편차는 697.62였다. 입원 당시 여러 개의 배제 진단이 내려진 환자의 경우 퇴원 후 외래 경과상에서 정신건강의학과 전문의의 판단으로 주 진단이 상기 기술한 3개의 진단 중 하나로 명확한 경우에만 해당 진단으로 본 연구에 포함하였다. 본 연구는 임상연구심의위원회의 승인을 받았다(승인번호 2018-09-003).

데이터의 분리(Data splitting)

의무기록에서 현병력을 추출하여 하나의 문단으로 만든 494개의 데이터 세트는 훈련용(n = 400)과 검증용(n = 94)으로 세 가지 진단명의 비율에 맞게 층화 무작위 배정(stratified randomization)을 하였다.

Table 1. Demographic information of the subjects

	MDD (n = 300)	BP1 (n = 91)	Schizophrenia (n = 103)
Age	48.55 ± 20.61	39.04 ± 15.94	38.45 ± 12.46
Sex, n (%)			
Male	86 (28.7)	32 (35.2)	40 (38.8)
Female	214 (71.3)	59 (64.8)	63 (61.2)
Marital status, n (%)			
Married	139 (46.3)	29 (31.9)	34 (33.0)
Unmarried	107 (35.7)	52 (57.1)	65 (63.1)
Divorced, widowed	54 (18.0)	10 (11.0)	4 (3.9)
Education, n (%)			
Elementary school	73 (24.3)	7 (7.7)	4 (3.9)
Middle school	44 (14.7)	3 (3.3)	8 (7.8)
High school	114 (38.0)	40 (44.0)	41 (39.8)
College	69 (23.0)	41 (45.1)	50 (48.5)

Data represent mean ± standard deviation for age. MDD : Major depressive disorder, BP1 : Type 1 bipolar disorder

의무기록의 전처리

비정형적인 텍스트 데이터를 컴퓨터가 이해할 수 있는 수치형 자료로 표현하기 위해서는 자연어 처리(natural language processing)라는 전처리 과정을 거쳐야 한다.[15] 텍스트 마이닝을 위해서는 연속된 문자열로 표현된 텍스트 데이터를 의미 표현의 기본 단위인 토큰(token)으로 나누어주는 토큰화(tokenization) 작업을 한다. 한글의 경우 일정한 의미를 지닌 가장 작은 말의 단위를 형태소(morpheme)라 하며 이렇게 의미를 나타내는 데 적합한 가장 기본적인 단위를 텍스트 마이닝에서는 토큰이라 한다. 토큰화를 위해서 파이썬 기반 한국어 정보처리 패키지인 Korean natural language processing in python(KoNLPy)에 내장된 한글 형태소 분석기를 활용하였다.[16] 현병력은 전문적인 의학 용어 및 영어, 약어 등이 배제된 채 일반용어로 서술되어, 한글 형태소 분석기를 적용하였다. 품사(part-of-speech) 태깅(tagging)은 형태소의 뜻과 문맥을 고려하여 그것에 표식(mark-up)을 하는 일로 각 토큰을 태깅하기 위하여 KoNLPy에 내장된 태그 패키지(tag package) 중 하나인 MeCab-ko(https://bitbucket.org/eunjeon/mecab-ko-dic)를 사용하였다. 각 데이터 세트를 토큰화 후 평균 형태소 개수는 1506개, 표준편차는 367.64였다. 진단명을 시사하는 특정한 단어를 포함하는 자료의 비율을 표 2에 나타내었다. 이러한 단어를 포함하고 있는 자료는 약 17.2%(n = 85)로 나타났다.

데이터의 수치화

태깅된 토큰을 키워드 추출하여 수치형 자료인 행렬로 벡터화하기 위하여 term frequency-inverse document frequency(TF-IDF) 모델과 Doc2vec 모델을 각기 이용하였다. TF-IDF 모델은 토큰이 여러 문서 중 특정 문서 내에서 얼마나 빈도수가 높은지를 평가하기 위한 통계적 기법으로 문서 유사도 측정에 활용된다.[17] 본 연구에서는 파이썬 기계학습 패키지인 scikit-learn의 TF-IDF vectorizer(https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html)를 활용하여 형태소 빈도수를 기반으로 문서를 통계적으로 벡터화하여 벡터 공간에 대해 유사도를 측정하였다. 본 연구와 같이 방대한 양의 문서에서 작업을 하는 경우 전체 문서에서 분류한 총 토큰의 수가 많아지는 만큼 벡터의 차원 또한 크게 늘어나게 된다. 각 문서의 벡터값은 각각의 문서에서 등장하지 않는 토큰에 대해 0으로 채워지게 되므로 TF-IDF로 표현된 벡터는 고차원(high-dimentional)의 성긴(sparse) 구조를 지니게 된다.[18]

Doc2vec 모델은 문장, 단락, 문서와 같은 가변 길이의 텍스트를 비지도 학습하여 고정 길이의 벡터로 표현하기 위해 제안되었다.[19] Doc2vec은 문서의 의미가 반영된 유사도에 기반하여 각각의 문서를 하나의 고유한 벡터로 생성하는 알고리즘으로서, 문서에 포함된 토큰의 개수와 상관없이 문서 자체를 하나의 고정된 크기의 벡터로 표현한다. Doc2vec은 동일한 범주에 속한 문서들에 대해 유사한 문서 벡터를 생성하므로 레이블과 실제 데이터가 필요하며, 본 연구에서는 주요

Table 2. The number of cases which had the words indicating psychiatric disorders in each subgroup

	MDD (n = 300)	BP1 (n = 91)	Schizophrenia (n = 103)	Total (n = 494)
Number of cases which had only a single word, n (%)				
'우울증'	43 (14.3)	7 (7.7)	6 (5.8)	56 (11.3)
'우울장애'	2 (0.7)	0 (0)	0 (0)	2 (0.4)
'조울증'	0 (0)	11 (12.1)	0 (0)	11 (2.2)
'조울병'	0 (0)	1 (1.1)	0 (0)	1 (0.2)
'양극성'	0 (0)	2 (2.2)	0 (0)	2 (0.4)
'분열병'	7 (2.3)	0 (0)	9 (8.7)	16 (3.2)
'조현병'	2 (0.7)	1 (1.1)	2 (1.9)	5 (1.0)
'스키조'	0 (0)	0 (0)	0 (0)	0 (0)
Number of cases which had two words concurrently, n (%)				
'우울증' and '조울증'	0 (0)	2 (2.2)	0 (0)	2 (0.4)
'우울증' and '우울장애'	1 (0.3)	0 (0)	0 (0)	1 (0.2)
'우울증' and '분열병'	1 (0.3)	0 (0)	1 (1.0)	2 (0.4)
'조울증' and '양극성'	0 (0)	1 (1.1)	0 (0)	1 (0.2)
Number of cases which had three words concurrently, n (%)				
'우울증' and '조울증' and '조현병'	0 (0)	1 (1.1)	0 (0)	1 (0.2)
Total number of cases which had any words, n (%)	52 (17.3)	17 (18.7)	16 (15.5)	85 (17.2)

MDD : Major depressive disorder, BP1 : Type 1 bipolar disorder

우울장애, 제1형 양극성장애, 조현병을 각기 0, 1, 2로 레이블링(labeling)하였다. TF-IDF를 이용하여 생성된 벡터의 크기는 평균적으로 13323였으며 Doc2vec 시행 시 5겹 교차 검증(five-fold cross-validation)을 통해, 초매개변수(hyper-parameter) 중 토큰의 분석 단위 개수(window)는 16, 총 토큰 수이자 최종 출력 벡터의 크기(vector size)는 100, 문서에 등장하는 토큰의 최소 빈도는 3으로 설정하여 기계학습 모델 성능에 최적화하였다.

기계학습

TF-IDF와 Doc2vec으로 벡터화하여 수치화한 데이터 세트를 각기 다른 학습 알고리즘(training algorithm)으로 기계학습을 시행하였다. TF-IDF로 벡터화한 데이터 세트를 확률적 경사 강하(stochastic gradient descent, SGD), 로지스틱 회귀(logistic regression, LR), 지지벡터기계(support vector machine, SVM)로 각기 기계학습을 시행하였다. Doc2vec으로 벡터화한 데이터 세트는 SGD, LR, SVM 및 딥러닝으로 각기 기계학습을 시행하였다. SGD, LR, SVM의 초매개변수는 모두 기본값으로 진행하였으며 딥러닝에 대해서 노드 개수(256)와 은닉층 수(5)에 대한 결과값을 산출하여 초매개변수값을 최적화하였다. TF-IDF에서는 문서에 나타난 총 토큰의 수만큼 고차원의 성긴 구조의 벡터를 생성하기 때문에 딥러닝을 시행하기에 부적합하다고 판단하였다.

딥러닝의 심층 신경망(deep neural network)은 기존 인공신경망(artificial neural network)에서 발전되어 온 기계학습법 중의 하나로서 인공 신경망보다 더 깊고(deep) 넓은(wide) 구조를 가지고 있다.[20] 심층 신경망은 은닉층(hidden layer)의 수를 증가시켜 일반적인 인공 신경망보다 더 정밀한 분류가 가능해지도록 발전하였으며 본 연구는 256개의 노드로 이루어진 총 5개의 은닉층을 적용하였다. 심층 신경망 구조에 정류 선형 유닛(rectified linear unit) 함수를 활성화 함수로 적용하여 가중치 재추정이 더욱 용이하게 만들었다. 본 연구에서는 전체 데이터 대신 일부 조그마한 데이터의 모음(mini-batch)에 대해서만 오차 함수(error function)를 계산하는 확률적 기울기 하강(stochastic gradient descent) 방법 중 하나인 아담 옵티마이저(Adam optimizer)를 사용하였다.[21] 기계학습 모델의 일반화를 위하여 5겹 교차 검증을 시행하였다. 데이터 전처리 및 학습과 관련된 연구 방법은 그림 1과 같이 요약하였다.

기계학습 모델 간의 성능 평가

이범주 분류를 위한 알고리즘을 다범주 분류 문제에 적용하기 위하여 세 군 간의 분류는 one versus rest 방법으로 진행하였다.[22] 이 방법은 k개의 범주 중 한 범주를 선택하여 그 범주는 양의 범주로, 선택하지 않은 다른 모든 범주는 음의 범주로 지정하여 이범주로 이루어진 자료 k개를 만든다. 각 레이블에 이범주 분류 알고리즘을 적용하여 각 레이블이 가지는 양의 범주에 대한 확률을 음의 범주에 대한 확률과 비교하였다. 다음과 같은 4가지 값을 분류 성능을 평가하기 위한 수치적 방법으로 활용하였다. 먼저 정확도[accuracy = 참양성 + 참음성/(참양성 + 참음성 + 거짓 양성 + 거짓 음성)]는 전체 자료 중 정확히 분류된 자료의 비중을 뜻한다. 정확도가 분류 성능 평가의 한 기준이 되지만 분류 성능을 정확히 평가하기 위해서는 다른 지표들도 확인해 볼 필요가 있다.[23] 정밀도[precision = 참양성/(참양성 + 거짓 양성)]는 양으로 분류된 자료들 중 참값이 실제로 양인 자료의 비중을 의미한다. 민감도[sensitivity, recall = 참양성/(참양성 + 거짓 음성)]는 참값이 양인 자료들 중 분류 결과 양으로 식별된 자료의 비중을 의미한다. 참값이 양인 자료를 정확히 분류하는 것이 중요할 때에는 이들 지표 중에서 정밀도와 민감도가 상대적으로 중요한 지표가 되며, 이 둘의 조화평균이 F1-score [2 × (정밀도 × 민감도)/(정밀도 + 민감도)]이다. F1-score는 데이터 레이블이 불균형 구조일 때, 모델의 성능을 정확하게 평가할 수 있다. 여러 범주를 가진 데이터의 평균을 구할 때는 마이크로 평균(micro-average)과 매크로 평균(macro-average)을 사용하며, 균형 잡힌 평균을 구하기 위하여 매크로 평균보다 마이크로 평균을 이용하는 것이 좋다. 매크로 평균은 각 범주의 평균 합/범주 개수의 합으로 범주를 구별하여 평가하기 위해 사용하며 마이크로 평균은 개별 수치의 합/전체 개별 수치의 개수로 각 개별 수치를 구별하여 평가하기 위해 사용된다.

결 과

5겹 교차 검증을 한 훈련용 데이터 세트(train data set)에 대하여 TI-IDF로 벡터화를 한 후 SGD, LR, SVM를 시행하여 정확도는 SVM에서 0.79, 정밀도는 SVM에서 0.79, 민감도는 SVM에서 0.79, F1-score는 SVM에서 0.79로 제일 높게 측정되었다. Doc2vec으로 벡터화를 한 후 SGD, LR, SVM 및 딥러닝을 시행하여 정확도는 딥러닝에서 0.87, 정밀도는 딥러닝에서 0.87, 민감도는 딥러닝에서 0.87, F1-score는 딥러닝에서 0.87로 제일 높게 측정되었다. 검증용 데이터 세트(test data set)에 대하여 TI-IDF로 벡터화를 한 후 SGD, LR, SVM을 시행하여 정확도는 SGD 및 SVM에서 0.78, 정밀도는 SGD 및 SVM에서 0.78, 민감도는 SGD 및 SVM에서 0.78, F1-score는 SGD 및 SVM에서 0.78로 제일 높게 측정되었다.

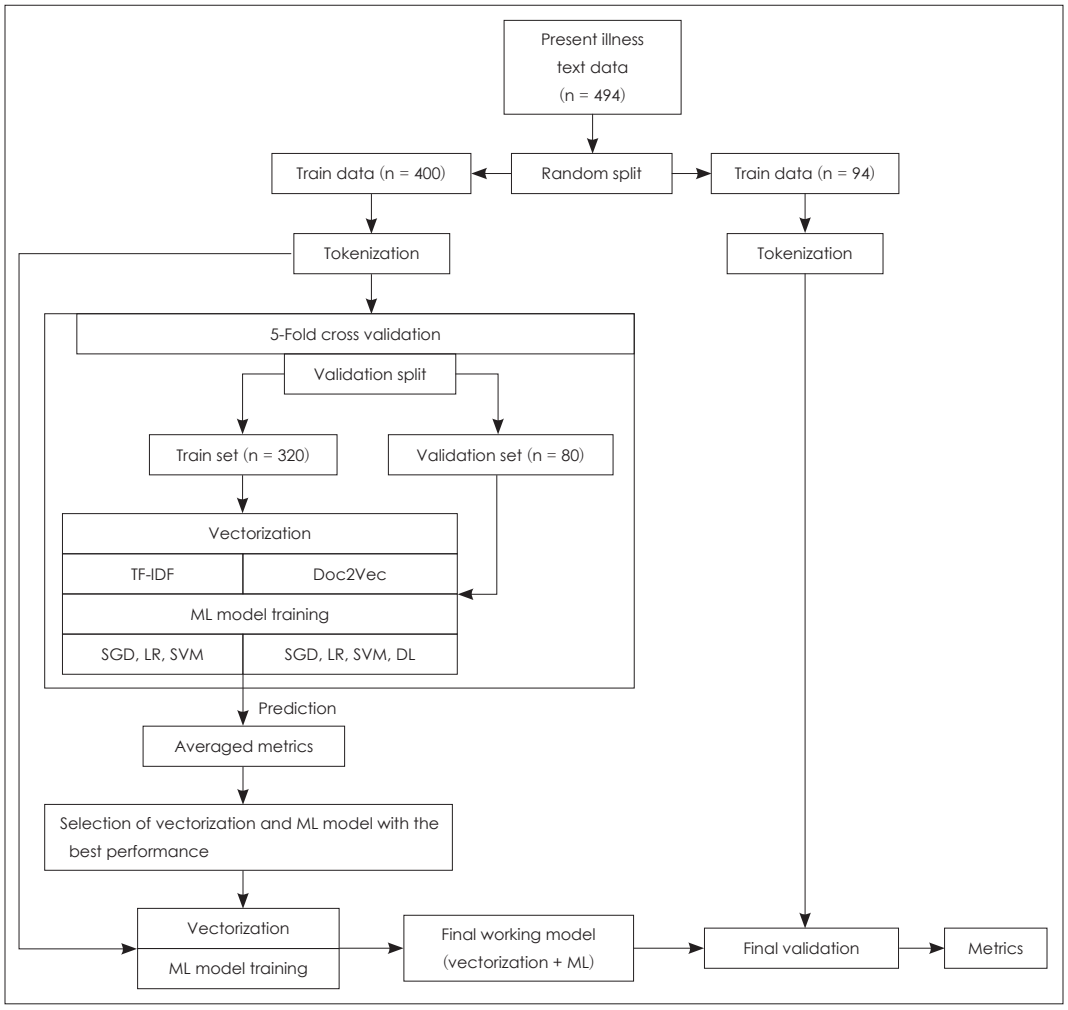

Fig. 1. Flow diagram of text-classification based machine learning for psychiatric diagnosis using present illness of admission records. ML : Machine learning, SGD : Stochastic gradient descent, LR : Logistic regression, SVM : Support vector machine, DL : Deep learning, TF-IDF : Term frequency-inverse document frequency.

Doc2vec으로 벡터화를 한 후 SGD, LR, SVM 및 딥러닝을 시행하여 정확도는 LR 및 SVM에서 0.80, 정밀도는 LR 및 SVM에서 0.80, 민감도는 LR 및 SVM에서 0.80, F1-score는 LR 및 SVM에서 0.80으로 제일 높게 측정되었다. TF-IDF로 벡터화를 하고 LR로 기계학습을 시행한 경우 훈련용 데이터 세트와 검증용 데이터 세트 모두에서 제1형 양극성장애 및 조현병을 주요우울장애로 예측하여 정밀도, 민감도 및 F1-score가 0으로 측정되었다. 본 연구의 데이터는 레이블 0, 1, 2의 숫자가 각기 300, 91, 103으로 불균형 구조를 이루고 있어 F1-score를 기준으로 분류 모델의 성능을 평가하였다. 훈련용 데이터 세트에 대하여 Doc2vec으로 벡터화를 한 후 딥러닝을 시행하였을 때 가장 높은 성능을 나타내었고 검증용 데이터 세트에 대하여 Doc2vec으로 벡터화를 한 후 LR와 SVM를 시행하였을 때 가장 높은 성능을 나타내었다. 표 3은 TF-IDF와 Doc2vec으로 데이터를 수치화한 후 훈련용 데이터 세트와 검증용 데이터 세트에서 각기 기계학습을 시행한 결과값을 나타낸다. 표 4는 주요우울장애, 제1형 양극성장애 및 조현병의 검증용 세트에 대한 실제값을 LR, SVM, 딥러닝으로 분류한 관측값과 비교하여 혼동 행렬(confusion matrix)의 형태로 나타내었다.

고 찰

본 연구는 한국어로 기록된 전자의무기록인 정신과 입원 환자의 현병력을 토대로 하여 기존 텍스트 분류에 사용되어 온 벡터화와 기계학습 두 과정의 각기 다른 모델 간의 성능을 비교하였다. 같은 조건에서 F1-score를 비교 시 Doc2vec을 적용한 모델이 TF-IDF를 적용한 모델보다 높은 성능을 산출하는 것을 확인할 수 있었다. 이것은 장문의 문서를 벡터화할 시 TF-IDF로 생성한 벡터가 Doc2vec으로 생성한 벡

Table 3. Metrics of each vectorization method and machine learning model

	Metrics	ML Models						
		TF-IDF			DOC2VEC			
		SGD	LR	SVM	SGD	LR	SVM	DL
Train set (5-fold CV)	Accuracy	0.73	0.61	0.79	0.82	0.85	0.84	0.87*
	Precision							
	MDD	0.73	0.61	0.79	0.84	0.87	0.89	0.90*
	BP1	0.52	0	0.70	0.91*	0.89	0.84	0.85
	Schizophrenia	0.74	0	0.87	0.76	0.78	0.71	0.83*
	Micro-average	0.73	0.61	0.79	0.82	0.85	0.84	0.87*
	Recall							
	MDD	0.98	1.00	0.95	0.97*	0.97*	0.93	0.96
	BP1	0.31	0	0.47	0.53	0.63	0.66	0.78*
	Schizophrenia	0.37	0	0.62	0.67	0.70	0.72*	0.72*
	Micro-average	0.73	0.61	0.79	0.82	0.85	0.84	0.87*
	F1-score							
	MDD	0.83	0.75	0.86	0.89	0.92*	0.91	0.93
	BP1	0.38	0	0.56	0.66	0.73	0.73	0.81*
	Schizophrenia	0.47	0	0.72	0.71	0.73	0.71	0.76*
	Micro-average	0.73	0.61	0.79	0.82	0.85	0.84	0.87*
Test set	Accuracy	0.78	0.61	0.78	0.78	0.80*	0.80*	0.79
	Precision							
	MDD	0.83	0.61	0.83	0.89	0.89	0.91*	0.91*
	BP1	0.69	0	0.75	0.79*	0.77	0.73	0.79*
	Schizophrenia	0.67*	0	0.63	0.54	0.58	0.60	0.54
	Micro-average	0.78	0.61	0.78	0.78	0.80*	0.80*	0.79
	Recall							
	MDD	0.91	1.00*	0.91	0.84	0.89	0.86	0.86
	BP1	0.65*	0	0.53	0.65*	0.59	0.65*	0.65*
	Schizophrenia	0.50	0	0.60	0.70	0.70	0.75*	0.70
	Micro-average	0.78	0.61	0.78	0.78	0.80*	0.80*	0.79
	F1-score							
	MDD	0.87	0.75	0.87	0.86	0.89*	0.88	0.88
	BP1	0.67	0	0.62	0.71*	0.67	0.69	0.71*
	Schizophrenia	0.57	0	0.62	0.61	0.64	0.67*	0.61
	Micro-average	0.78	0.61	0.78	0.78	0.80*	0.80*	0.79

* : The best metric among the models. TF-IDF : Term frequency-inverse document frequency, ML : Machine learning, SGD : Stochastic gradient descent, LR : Logistic regression, SVM : Support vector machine, DL : Deep learning, CV : Cross-validation, MDD : Major depressive disorder, BP1 : Type 1 bipolar disorder

터보다 고차원의 성긴 구조를 지니고 있으며, 고차원의 성긴 구조의 자료를 학습하기 위해서는 대량의 학습 데이터가 요구되어(차원의 저주)[24] 본 연구와 같이 제한된 수의 자료로 학습하는 경우 모델의 성능이 낮게 나올 수밖에 없는 것으로 생각된다. 이 연구 결과에 기반하여, TF-IDF보다 Doc2vec을 활용하여 현병력을 벡터화하는 것이 진단 분류 성능 향상에 효과적임을 알 수 있다.

본 연구는 기존 의학 외 분야에 활용되고 있는 기계학습을 통한 텍스트 분류 기법을 정신과 진단 분류에 활용한 국내 최초의 연구이다. 특히 한국어 텍스트 분류를 위한 최적의 전처리 방법과 기계학습 모델이 정형화되지 않은 만큼 이를 의무기록에 적용하여 분류 성능을 측정하는 것은 기존 의학 연구에서는 시도되지 않았던 접근 방법이다. 기존의 인공지능을 활용한 국내의 의학 연구에서는 각종 수치화된 자료를 활용하여 왔으나 본 연구에서는 비정형적인 텍스트 데이터만을 분석하여 정신과 진단 분류가 가능함을 밝혀내었다. 이는 영상 자료를 분석하여 의사의 영상 판독 및 진단을 보조하는 컴퓨터 보조 진단(computer-aided detection and diagnosis)[25] 시스템을 텍스트 데이터를 활용하여 정신과 분야에 적용시킬 수 있는 가능성을 제시하였다.

Table 4. Confusion matrices of LR, SVM, and DL

Actual labels	Predicted labels			
	MDD	BP1	Schizophrenia	Actual total
LR				
MDD	51	1	5	57
BP1	2	10	5	17
Schizophrenia	4	2	14	20
Predicted total	57	13	24	94
SVM				
MDD	49	2	6	57
BP1	2	11	4	17
Schizophrenia	3	2	15	20
Predicted total	54	15	25	94
DL				
MDD	48	2	7	57
BP1	1	12	4	17
Schizophrenia	4	2	14	20
Predicted total	53	16	25	94

LR : Logistic regression, SVM : Support vector machine, DL : Deep learning, MDD : Major depressive disorder, BP1 : Type 1 bipolar disorder

기존 한국어 텍스트 분류 성능 비교를 위한 연구는 의학 외의 분야에서 다양하게 시도되었다. 생활 화학제품 관련 4800개의 뉴스 기사를 Doc2vec으로 벡터화한 후 각 기사에서 다루는 화학제품의 위해성 여부에 대하여 각기 다른 4개의 기계학습 알고리즘(LR, decision tree, naive bayesian, SVM)의 이진 분류 성능을 비교하였던 실험 결과, SVM의 F1-score가 0.83으로 가장 좋은 성능을 보였다.[26] 10000개의 신문기사 데이터를 Doc2vec 모델로 학습시킨 후 10개의 범주로 주제 분류 성능을 측정한 결과 딥러닝의 하나인 convolutional neural network(CNN)의 정확도가 0.82로 측정되었으며 Doc2vec과 Word2vec 모델을 함께 적용한 경우 분류율이 0.89로 향상되었다.[27] 20000개의 네이버 영화 평점을 15000개의 훈련용 세트와 5000개의 검증용 세트로 나눈 후 Word2vec으로 벡터화한 후 CNN 모델로 분류한 경우 F1-score가 0.73이 측정되었으며 recurrent neural network(RNN) 모델로 분류한 결과 분류율이 0.88로 향상되었다.[28] 본 연구는 평균 1000자 내외로 알려져 있는 신문기사[29] 및 1000자 이하로 글자 수 제한이 있는 네이버 영화 평점에 비하여 평균 2904자의 긴 문서를 분석 대상으로 삼았다. 또한 상기 연구들에 비해 494개의 비교적 적은 표본 수를 3개의 범주로 분류를 하였음에도 LR과 SVM의 F1-score가 0.80으로 측정되어 Doc2vec을 활용하는 것이 정신과 문서의 진단 분류에 효과적임을 국내에서 처음으로 검증하였다.

본 연구가 갖는 한계점과 추후 연구 제안은 다음과 같다. 첫째, 본원 입원 환자의 비율상 주요우울장애(n = 300)가 제1형 양극성장애(n = 91) 및 조현병(n = 103)보다 세 배 가량 많아 표본의 크기를 동일화하지 못하였다(imbalanced dataset). 이는 집단간 표본 개수의 차이가 커서 표본의 대부분을 표본의 수가 적은 소수 집단(minority class)보다 표본의 수가 많은 다수 집단(majority class)으로 예측하게 되는 집단 불균형 문제(class imbalance problem)를 야기한다.[30] 일반적인 기계학습 알고리즘은 각 집단의 표본 수가 비슷하다고 가정하기 때문에, 집단 불균형 문제가 있는 경우 전체 문서의 분류 성능이 높더라도 소수 집단의 문서를 제대로 예측하지 못하는 경우가 많다. 다수 집단 예측과 관련한 모델의 성능이 과대평가되는 점을 보완하기 위하여 F1-score를 모델 성능 평가에 사용하였다.[31] 그럼에도 불구하고, 본 연구 결과 해석에 집단 불균형의 문제를 반드시 고려해야 할 것으로 생각된다. 집단 불균형 문제를 완화하기 위해 소수 집단에 속하는 표본의 수를 늘리는 과대표집(oversampling) 기법과 다수 집단에 속하는 표본의 수를 줄이는 과소표집(undersampling) 기법 등이 연구되고 있다.[32] 둘째, 기존 다양한 분야에서 기계학습을 통한 한국어 문서분류에 활용되었던 표본의 수는 수천 개 이상으로 본 연구에서 사용한 494개의 현병력은 기계학습을 시행하기에 비교적 적은 수이다. 하지만 총 1000개의 평균 300자 이내의 정신과 외래 초진 현병력으로 11개의 각기 다른 진단 분류를 시도하였던 해외 연구[33]와 비교하여 한 범주에 속하는 평균 표본 수는 본 연구에서 더 많았다. 494개라는 적은 표본 수로 인하여 교차 검증 시 학습 세트보다 검증 세트에서의 기계학습의 성능(특히 딥러닝)이 저하되는 과적합(overfitting)[34]이 발생한 것으로 추정된다. 추후 다기관 협력 연구를 통하여 표본 수를 증가시킨다면 딥러닝을 포함한 기계학습의 성능이 개선될 것으로 기대된다. 셋째, 본 연구에서 사용한 현병력은 일 대학병원의 입원 환자의 기록을 대상으로 후향적으로 수집하였기 때문에 입원치료가 필요할 수 있는 3개의 주된 진단명으로 제한할 수밖에 없었으며, 진단명 내에서 명시 및 동반 이환의 종류가 다를 경우에 대한 세부적인 구분을 시행하지 못하였다. 구조적인 면담 도구를 활용하지 않고 임상진료시 내려진 진단명을 사용하였기 때문에 진단명의 타당성에 대한 근거가 부족한 부분이 본 연구의 또 다른 제한점이 될 수 있다. 추후 다기관 협력 연구를 통하여 구조화된 진단 도구를 이용하여 전향적으로 대량의 표본을 수집한다면, 이러한 제한점이 극복될 수 있을 것이다. 넷째, 현병력 작성 시, 해당 환자의 진단명 및 복용했던 약물명이 현병력에 포함될 가능성이 있으며, 이러한 진단명이 기계학습 성능에 영향을 주었을 수 있다. 본 연구에서 활용된 494개의 현병력의 17.2%에서 정신질환 진단명과 연관된 단어가 사용되었다(표 2). 그 외 약물명은 현병력에서

사용되지 않았다. 이러한 진단을 시사하는 단어가 제한적으로 포함된 경우, 학습성능에 미치는 영향에 대해서는 본 연구에서 확인하지 못한 것이 또 다른 연구의 제한점이다. 추후 연구에서 이러한 진단명과 정신병리 용어가 포함된 데이터를 배제한 경우와 포함시킨 경우 차이를 비교하여 분류 성능에 미치는 영향을 확인할 수 있을 것이다.

본 논문은 벡터화에 있어서 TF-IDF와 Doc2vec의 활용이 텍스트 분류 성능에 끼치는 영향 및 각기 다른 기계학습 알고리즘을 적용하였을 때 텍스트 분류 성능의 차이를 비교 검증하는 것을 목적으로 하였다. 한국어 문서의 분류를 위한 토큰화, 벡터화, 기계학습 과정에 있어 각기 다른 설정을 적용하여 분류율의 향상을 기대할 수 있으나 방법론의 조합의 수가 많아 본 연구에서 모두 다루지 못하였다. 토큰화 방법으로 어절 단위, 형태소 분석, word piece model 등 다양한 방법을 적용해 볼 수 있다. 벡터화에 있어 TF-IDF에서 일부 특징(feature)을 추출하는, 특징 선택(feature selection)을 통해 차원의 저주를 해결하기 위한 다양한 연구들이 제안되었다.[34)35)] 본 연구 자료에서 진단명별로 출현하는 단어의 빈도를 확인하여, 차이가 있는 단어를 문서별로 조합하여 학습에 사용하며 입력 데이터의 크기를 줄이고, 성김 정도를 줄이는 연구를 추후에 고려해 볼 수 있다. 그 외, Doc2vec의 성능 개선을 위하여 Word2vec과 함께 적용하는 시도 또한 있었다.[27)] 딥러닝 모델의 경우 매개변수 튜닝 및 신경망 층 구조 변경(CNN, RNN 등)을 통한 성능 향상이 기대되므로 서로 다른 토큰화, 벡터화 모델의 적용 및 최적의 딥러닝 모델을 찾는 후속 연구를 할 필요가 있다.

또한 K겹 교차 검증 과정에 있어 K가 작은 경우 분산(variance)은 감소하나 편향(bias)이 증가할 확률이 높다고 하며, K가 커질 경우 분산은 증가하나 편향이 감소한다고 알려져 있어[36)] 추후에 K겹 교차 검증에서 K값의 변화나 무작위 추출의 반복을 통하여 모델의 성능 평가를 현재와 비교해보는 연구도 필요할 것으로 보인다.

결론적으로 본 연구는 정신건강의학과 병동에 각기 다른 세 가지 진단명으로 입원한 환자의 현병력에 토큰화, 벡터화 및 기계학습을 시행하여 진단 분류 정확도를 확인하였다. TF-IDF보다 Doc2vec으로 벡터화를 한 경우 진단 예측률이 높았으며, 기계학습 모델 간 진단 예측에는 차이가 크지 않았다. 본 연구 결과를 통해 향후 정신건강의학 분야에서 인공지능을 활용하여 한국어로 된 텍스트 데이터를 분석하는 데 기여할 수 있을 것으로 기대한다.

중심 단어: 텍스트 분류 · 전자의무기록 · 벡터화 · 기계학습 · 현병력 · 정신과 진단.

Acknowledgments

본 연구는 소방청 현장중심형 소방활동지원 기술개발사업(MPSS-소방안전-2016-86) 및 한국연구재단을 통해 과학기술정보통신부의 뇌과학원천기술개발사업(2015M3C7A1028376)으로부터 지원받아 수행되었습니다.

Conflicts of interest

The authors have no financial conflicts of interest.

Author Contributions

Conceptualization: Jaeuk Hwang. Data curation: Doohyun Pak, Mingyu Hwang, Minji Lee. Formal analysis: Doohyun Pak, Jaeuk Hwang. Funding acquisition: Jaeuk Hwang. Investigation: Doohyun Pak, Mingyu Hwang, Minji Lee. Methodology: Doohyun Pak, Jaeuk Hwang. Project administration: Doohyun Pak. Resources: Doohyun Pak, Jaeuk Hwang. Supervision: Jaeuk Hwang. Validation: Mingyu Hwang, Minji Lee, Sung-Il Woo, Sang-Woo Hahn, Yeon Jung Lee. Visualization: Doohyun Pak, Jaeuk Hwang. Writing—original draft: Doohyun Pak. Writing—review & editing: Mingyu Hwang, Minji Lee, Sung-Il Woo, Sang-Woo Hahn, Yeon Jung Lee, Jaeuk Hwang.

ORCID iDs

Doohyun Pak	https://orcid.org/0000-0003-0455-8901
Mingyu Hwang	https://orcid.org/0000-0002-5805-2335
Minji Lee	https://orcid.org/0000-0003-4962-3869
Sang-Woo Hahn	https://orcid.org/0000-0003-1662-5438
Yeon Jung Lee	https://orcid.org/0000-0001-8953-5893
Jaeuk Hwang	https://orcid.org/0000-0003-0528-3305

REFERENCES

1) **Miotto R, Li L, Kidd BA, Dudley JT.** Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016;6:26094.
2) **Nguyen P, Tran T, Wickramasinghe N, Venkatesh S.** $\mathtt{Deepr}$: a convolutional net for medical records. IEEE J Biomed Health Inform 2017;21:22-30.
3) **Craddock N, Mynors-Wallis L.** Psychiatric diagnosis: impersonal, imperfect and important. Br J Psychiatry 2014;204:93-95.
4) **Weiss SM, Indurkhya N, Zhang T, Damerau F.** Text Mining: Predictive Methods for Analyzing Unstructured Information. New York, NY: Springer Science & Business Media;2010.
5) **Srivastava AN, Sahami M.** Text Mining: Classification, Clustering, and Applications. Boca Raton, FL: Chapman and Hall/CRC;2009.
6) **Deo RC.** Machine learning in medicine. Circulation 2015;132:1920-1930.
7) **Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V.** Artificial intelligence transforms the future of health care. Am J Med 2019;132:795-801.
8) **Forsting M.** Machine learning will change medicine. J Nucl Med 2017;58:357-358.
9) **Banerjee I, Ling Y, Chen MC, Hasan SA, Langlotz CP, Moradzadeh N, et al.** Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artif Intell Med 2019;97:79-88.
10) **Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, et al.** Deep learning to classify radiology free-text reports. Radiology 2018;286:845-852.
11) **Tran T, Kavuluru R.** Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks. J Biomed Inform 2017;75 Suppl:S138-S148.

12) **Sadock BJ, Sadock VA, Ruiz P.** Kaplan and Sadock's Synopsis of Psychiatry: Behavioral Sciences/Clinical Psychiatry. Philadelphia, PA: Wolters Kluwer Health;2014. p.192-211.
13) **American Psychiatric Association.** Diagnosis and Statistical Manual of Mental Disorders: DSM-IV. 4th ed. Washington, DC: American Psychiatric Association;1994.
14) **American Psychiatric Association.** Diagnostic and Statistical Manual of Mental Disorders: DSM-5. 5th ed. Arlington, VA: American Psychaitric Association;2013.
15) **Bird S, Klein E, Loper E.** Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Sevastopol, CA: O'Reilly Media, Inc.;2009.
16) **Park EL, Cho S.** KoNLPy: Korean natural language processing in Python. Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology;2014 Oct 10-11, Chuncheon, Korea.
17) **Ramos JA.** Using TF-IDF to determine word relevance in document queries. Proceedings of the First Instructional Conference on Machine Learning;2003 Dec 8-10, Rutgers, NJ, USA.
18) **Cataltepe Z, Aygun E.** An improvement of centroid-based classification algorithm for text classification. Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering;2007 Apr 17-20, Istanbul, Turkey.
19) **Le Q, Mikolov T.** Distributed representations of sentences and documents. Proceedings of the 31st International Conference on Machine Learning;2014 Jun 22-24, Beijing, China.
20) **LeCun Y, Bengio Y, Hinton G.** Deep learning. Nature 2015;521:436-444.
21) **Kingma DP, Ba J.** Adam: a method for stochastic optimization. Proceedings of the 3rd International Conference for Learning Representations;2015 May 7-9, San Diego, CA, USA.
22) **Kiers HA, Rasson JP, Groenen PJ, Schader M.** Data Analysis, Classification, and Related Methods. Heidelberg: Springer-Verlag;2000. p.181-186.
23) **Fawcett T.** An introduction to ROC analysis. Pattern recognition letters 2006;27:961-874.
24) **Bellmann R.** Dynamic Programming. Princeton, NJ: Princeton University Press;1957.
25) **Giger ML, Suzuki K.** Computer-aided diagnosis. In: Feng DD, editor. Biomedical Information Technology. Burlington, MA: Elsevier;2008. p.359-370.
26) **Jeong J, Jee M, Go M, Kim H, Lim H, Lee Y, et al.** Related documents classification system by similarity between documents. Journal of Broadcast Engineering 2019;24:77-86.
27) **Kim D, Koo M-W.** Categorization of Korean news articles based on convolutional neural network using Doc2Vec and Word2Vec. Journal of KIISE 2017;44:742-747.
28) **Kim J-M, Lee J-H.** Text document classification based on recurrent neural network using Word2vec. J Korean Inst Intell Syst 2017;27:560-565.
29) **Heo S-W, Sohn K-A.** Feature extraction to detect hoax articles. Journal of KIISE 2016;43:1210-1215.
30) **He H, Garcia EA.** Learning from imbalanced datas. IEEE Trans Knowl Data Eng 2009;21:1263-1284.
31) **Forman G, Scholz M.** Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM SIGKDD Explorations Newsletter 2010;12:49-57.
32) **Liu Y, Loh HT, Sun A.** Imbalanced text classification: a term weighting approach. Expert Systems with Applications 2009;36:690-701.
33) **Hastie T, Tibshirani R, Friedman J.** The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer Series in Statistics;2001.
34) **Guyon I, Elisseeff A.** An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157-1182.
35) **Tang J, Alelyani S, Liu H.** Feature selection for classification: a review. In: Aggarwal CC, editor. Data Classification: Algorithms and Applications. Boca Raton, FL: Chapman & Hall/CRC;2015. p/37-64.
36) **Geman S, Bienenstock E, Doursat R.** Neural Networks and the Bias/Variance Dilemma. Cambridge, MA: MIT Press;1992. p.1-58.