A Study on the Prediction of Depression using Semi-supervised Learning

EunSong Bang[1], Minsu Chae[1], Hwamin Lee[1]

[1] Department of Medical Informatics, College of Medicine, Korea University, Seoul 02841, Republic of Korea; bang08877@gmail.com, {misnuchae, hwamin}@korea.ac.kr

Abstract. The effects of depression extend from the individual to society. Predicting and treating depression is crucial for preserving people's mental health. This study aims to make a model for predicting depression. We investigate seven machine learning classifiers such as decision tree, random forest, SVM, logistic regression, Ada boost, XGB, and LGBM. Data include both labeled and unlabeled data. About 94% of the data in this set is unlabeled. We apply supervised machine learning and semi-supervised machine learning to train the model. Finally, we obtain improved precision, recall, and F1 score.

Keywords: Depression, Semi-Supervised Learning

1 Introduction

For years, the COVID-19 pandemic has been prevalent all over the world. People's lifestyles changed as a result of the COVID-19 pandemic, which had an impact on their mental health [1]. According to World Health Organization (WHO) reports, the Covid-19 pandemic triggers a 25% increase in depression [2]. Depression is a common mental disorder. However, depression can cause severely poor function in life quality. At worst, it can lead to suicide [3]. The nation must be concerned about the mental health of its citizens to avoid the worst. This study used Korea National Health and Nutrition Examination Survey (KNHHES) as a data source. This data source is conducted by the government every year. This data source is used to investigate the health of the public. However, both labeled and unlabeled data are present in this data. Therefore, we employed a semi-supervised learning algorithm to predict the patients with depression. Semi-supervised learning is a good option when you need to employ both labeled and unlabeled data. The performance of the machine learning model can be improved [4].

2 Related Work

There are many approaches to predict and diagnose depression using machine learning algorithms such as SVM, reinforcement, and transfer learning [5]. The results of semi-supervised models to predict depression are displayed in Table 1.

Table 1. Performance Result of Researches

Reference	Semi-Supervised Model	Data	Precision	Recall	F1-score
[6]	Statistical model	Tweets	72%	-	-
[7]	Graphic convolutional neural network model	Electroencephalogram (EEG)	89.246%	98.324%	-
[8]	Semi-supervised Graph Instance Transformer	Mobile Sensing	-	-	0.823

3 Material

The data used in this study is Korea National Health and Nutrition Examination Survey 2020 provided by the Korea Disease Control and Prevention Agency (KDCA) [9]. The data inform the national health level, health type, food, and nutrition status. Adults above the age of 19 are the primary focus of this study. Basic factors including sex, age, economic activity status, education level, household survey factors like household members and marital status, as well as self-rated health indicators like frequency of drinking, amount of daily exercise, stress level, and smoking, were employed in this study. The class was chosen based on the participant's clinical condition, including whether or not they were depressed. The survey respondents' options were "yes", "no", or "unknown". There were 5,386 participants in the study. Only 299 of them accurately responded to the class question. The dataset's description can be found in Table 2.

Table 2. Description of Data

Data	Description
Survey period	January 2020 – December 2020
Total participants	7,359
Used participants	5,386
Number of Features	17
Total Male	46.4%

Table3 shows the variables that we used. This variable consisted of basic factors, household survey factors, self-rated health indicators, and class.

Table 3. Variables

Variable	Description
Region	The area of residence
Sex	Male/Woman
Age	≥19 age
Income	Income quartile
Edu	Level of Education
EC1_1	Economic Activity
Cfam	Household member
Allownc	Beneficiary of National Basic Livelihood
House	Home ownership
Marri_2	marital status
D_1_1	self-health recognition
D_2_1	feel bad over the past two weeks
BD1_11	drinking frequency
BE3_75	daily exercise
Mh_stress	stress level
BS3_1	Smoking
DF2_pr	Depression / No Depression / Unknown

4 Results

We employed semi-supervised learning algorithms to use unlabeled data. We first trained supervised learning models using labeled data. The unlabeled data was then labeled using the trained supervised learning models. After we had the pseudo-labeled data with the highest probability of being correct, we repeatedly ran the model train until no more unlabeled data do not remain. Many classification models and evaluation were factored in through utilizing our data for choosing the best model. The results of supervised learning models are displayed in Table 4.

Table 4. The Results of Supervised Learning

Model	Type of source	Precision	Recall	F1 score
Decision Tree Classifier	Original source	0.63	0.64	0.64
	Oversampling	0.63	0.61	0.62
Random Forest Classifier	Original source	0.61	0.72	0.66
	Oversampling	0.59	0.6	0.6
SVM	Original source	0.65	0.76	0.7
	Oversampling	0.67	0.56	0.61

Logistic Regression	Original source	0.64	0.71	0.68
	Oversampling	0.69	0.59	0.63
Ada Boost Classifier	Original source	0.61	0.71	0.65
	Oversampling	0.63	0.71	0.67
XGB Classifier	Original source	0.61	0.72	0.66
	Oversampling	0.61	0.67	0.64
LGBM Classifier	Original source	0.62	0.71	0.66
	Oversampling	0.63	0.59	0.61

After training with labeled data, we selected the SVM. As previously said, we repeatedly trained the SVM. The progression of the model F1 score across the number of epochs is shown in Figure 1.

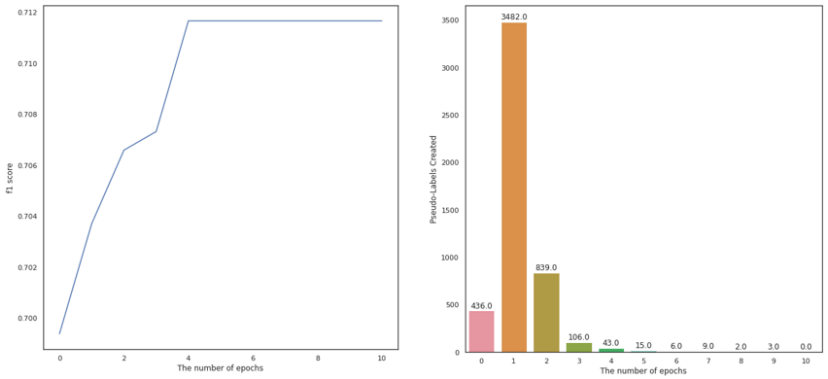

Figure 1. The Result of F1 score and Pseudo-labels following the number of epochs.

The result of the semi-supervised learning model is displayed in Table 5.

Table 5. The Result of Semi-Supervised Learning

Model	Type of Learning	Precision	Recall	F1 score
SVM	Supervised learning	0.65	0.76	0.7
	Semi-supervised learning	0.66	0.77	0.71

Precision, recall, and F1 score following semi-supervised learning were higher than those following supervised learning.

5 Conclusion

In this study, we can identify that unlabeled data could be helpful to improve the model performance. We improved model precision, recall, and F1 scores despite the fact that the survey is the data source that is impacted by participants' sincerity. We

anticipate that this study could help predict depression in people who have not yet a diagnosis. We will continue to improve this model and make it possible to predict other diseases.

Acknowledgments. This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) program(IITP-RS-2022-00156439) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation) and the Basic Science Research Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (NRF-2021R1A2C1009290).

Reference

1. Santomauro, D.F., et al.: Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. The Lancet. 398, 1700-1712 (2021)
2. World Health Organization.: World mental health report: transforming mental health for all. World Health Organization (2022)
3. World Health Organization. Regional Office for the Eastern Mediterranean.: Depression. World Health Organization. Regional Office for the Eastern Mediterranean (2019)
4. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine Learning 109, 373-440 (2019)
5. Aleem, S., Huda, N.u., Amin, R., Khalid, S., Alshamrani, S.S., Alshehri, A.: Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions. Electronics. 11, 1111 (2022)
6. Yazdavar, A.H., Al-Olimat, H.S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., Pathak, J., Sheth, A.: Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media. Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 1191-1198 (2017)
7. Wang, D., Lei, C., Zhang, X., Wu, H., Zheng, S., Chao, J., Peng, H.: Identification of Depression with a Semi-supervised GCN based on EEG Data. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2338-2345 (2021)
8. Guimin Dong, M.T., Lihua Cai, Laura E. Barnes, Mehdi Boukhechba: Semi-supervised Graph Instance Transformer for Mental Health Inference. 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp.1221-1228 (2021)
9. Korea Disease Control and Prevention Agency.: The seventh Korea National Health and Nutrition Examination Survey (KNHANES Ⅶ-2). (2020)