Analysis of Depression in News Articles Before and After the COVID-19 Pandemic Based on Unsupervised Learning and Latent Dirichlet Allocation Topic Modeling

Seonjae Been[1], Haewon Byeon[2]*
Department of Digital Anti-Aging Healthcare (BK21), Inje University, Gimhae 50834, South Korea[1]
Department of Medical Bigdata, Inje University, Gimhae 50834, South Korea[2]

***Abstract*—As of 2023, South Korea maintains the highest suicide rate among OECD countries, accompanied by a notably high prevalence of depression. The onset of the COVID-19 pandemic in 2020 further exacerbated the prevalence of depression, attributed to shifts in lifestyle and societal factors. In this research, differences in depression-related keywords were analyzed using a news big data set, comprising 45,376 news articles from January 1st, 2016 to November 30th, 2019 (pre-COVID-19 pandemic) and 50,311 news articles from December 1st, 2019 to May 5th, 2023 (post-pandemic declaration). Latent Dirichlet Allocation (LDA) topic modeling was utilized to discern topics pertinent to depression. LDA topic modeling outcomes indicated the emergence of topics related to suicide and depression in association with COVID-19 following the pandemic's onset. Exploring strategies to manage such scenarios during future infectious disease outbreaks becomes imperative.**

Keywords*—*COVID-19; depression; news articles; LDA topic modeling

I. INTRODUCTION

South Korea has consistently reported the highest suicide rate among OECD countries. Since 2004, South Korea has implemented a five-year suicide prevention plan called the Basic Plan for Suicide Prevention. However, the suicide rate in South Korea has remained the highest among OECD countries for an extended period, with a rate 2.2 times higher than the OECD average as of 2019 [1]. Previous studies have emphasized low treatment rates for depression as one of the causes of high suicide rates [2], indicating a significant number of individuals with depression in South Korea who do not receive treatment. Particularly, since the outbreak of COVID-19 in 2020, the prevalence of depression in South Korea has worsened. According to surveys conducted by Statistics Korea, while the pre-COVID-19 prevalence of depression was 20% [3], it increased nearly twofold to 36.8% after the COVID-19 pandemic [4].

The long-term COVID-19 pandemic has led to social changes in South Korea due to measures such as social distancing, reduced outings and gatherings, and economic difficulties [5]. As a result, issues such as social isolation due to disrupted social networks, economic hardships, weight gain, and concerns about COVID-19 infection spread have arisen [5]. These problems have had significant societal impacts in South Korea to an extent that gave rise to a new term called "corona blue."

As described above, depression in South Korea has been exacerbated socially due to the COVID-19 pandemic. According to a study by Lee et al., there was an increase in incidence rates of depression and prevalence rates for moderate and severe depression after the COVID-19 pandemic outbreak [6]. Furthermore, during periods of high transmission rates like during outbreaks or when Omicron variant spread internationally, significantly higher levels of excess mortality were observed [7]. In other words, the COVID-19 pandemic has had substantial effects not only on depressive symptoms but also on increased suicide rates among individuals. This study aims to analyze changes in keywords related to depression before and after the COVID-19 pandemic using news big data from South Korea through unsupervised learning algorithms such as Latent Dirichlet Allocation (LDA) topic modeling. By identifying hidden topics within collections of words or documents and grouping them based on documents or keywords using LDA topic modeling technique this research visualized key issues related to depressive symptoms before and after the COVID-19 pandemic.

II. RESEARCH METHODOLOGY

A. Data Collection

To analyze news articles related to depression before and after the COVID-19 pandemic, we utilized the BIGKinds service provided by the Korea Press Foundation. BIGKinds is a news big data analysis service that allows for the analysis of social phenomena using structured text data. In this study, we collected data from a total of 26 media outlets, including 11 national daily newspapers, eight economic daily newspapers, five broadcasting companies, and two specialized magazines (see Table I).

This study aimed to investigate depression before and after the COVID-19 pandemic by searching for keywords such as "depression," "feeling depressed," "depressive symptoms," and "depressive disorder." The search was divided into two periods: before and after the COVID-19 pandemic. The analysis period was determined considering the introduction of the COVID-19 pandemic in South Korea and the declaration of

*Corresponding Author.

an endemic. The period before the COVID-19 pandemic was set from January 1, 2016, to November 30, 2019. The period after the COVID-19 pandemic was set from December 1, 2019, to May 5, 2023 (the date of endemic declaration). After excluding duplicate articles and those unrelated to the topic, a total of articles used for analysis is shown in Table II.

TABLE I. MEDIA BIGDATA SET SUBJECT TO TABLE ANALYSIS

Category (Number)	Daily newspaper name
National daily newspaper (11)	Kyunghyang Shinmun, Kookmin Ilbo, Naeil Shinmun, Donga Ilbo, Munhwa Ilbo, Seoul Shinmun, Segye Ilbo, Chosun Ilbo, JoongAng Ilbo, Hankyoreh, Hankook Ilbo
Economic daily newspaper (8)	Maeil Economy, Money Today, Seoul Economy, Asia Economy, Aju Economy, Financial News, Korea Economy, Herald Economy
Broadcasting company (5)	KBS, MBC, OBS, SBS, YTN
Professional journals (2)	Digital Times, Electronic Newspaper

TABLE II. NUMBER OF ARTICLES USED FOR ANALYSIS

	Before the outbreak of COVID-19 pandemic	After the outbreak of COVID-19 pandemic
Total number of articles	47,905	52,707
Number of excluded articles	2,529	2,396
Number of analysis articles	45,376	50,311

B. Data Preprocessing

To investigate depression before and after the COVID-19 pandemic, we conducted keyword-based searches using combinations of terms such as "depression," "feeling depressed," and "depressive disorder." The analysis period was divided into two parts: pre-COVID-19 pandemic (from January 1, 2016, to November 30, 2019) and post-pandemic declaration (from December 1, 2019, to May 5, 2023), considering the timing of the COVID-19 outbreak in South Korea and the declaration of an endemic state. We excluded duplicate articles and those unrelated to the research topic.

C. Analysis Method

The analysis in this study was conducted using Python and involved frequency analysis, TF-IDF (Term frequency–inverse document frequency), and LDA (Latent Dirichlet Allocation) topic modeling.

First, we analyzed the frequency of word occurrences based on the collected articles. Then, we used TF-IDF to evaluate the relevance of words to the articles [8]. Subsequently, LDA topic modeling was applied to analyze the main topics within the collected articles.

LDA is a generative probabilistic model developed by Blei et al. and is one of the most well-known topic modeling techniques [9-11]. LDA analyzes data based on words used in documents or text data. It assumes that there are several topics and randomly selects a distribution of topics. Each word is then extracted from one of the topics, which is selected from the distribution specific to each document [11]. Therefore, LDA automatically discovers topics in documents and infers hidden semantic structures by outputting sets of words that have a high probability of co-occurrence within each topic [9-11].

III. RESULTS

A. Network Analysis

The network analysis of the relationship between depression and related keywords before and after the COVID-19 pandemic is shown in Fig. 1, Fig. 2 and also in Table III. The network analysis provided by BIGKinds is based on the top 100 articles with the highest accuracy among search results. It extracts noun phrases and applies Structured SVM (Support Vector Machine) to assign weights considering the number of related articles [12]. In Fig. 1, weights ranging from 4 to 34 were applied, while in Fig. 2, weights ranging from 6 to 51 were applied.

Fig. 1 represents the network analysis of depression-related keywords before the COVID-19 pandemic. The top keyword identified in this analysis is "program." This keyword is associated with programs related to depression operated by local governments and the central government. Additionally, the relationship diagram shows that it is related to Okcheon County due to its implementation of a project called "Zero Depression" [13]. The relationships associated with workers and Asan City are attributed to a depression screening conducted as part of mental health initiatives for workers by Asan City. The results showed that Asan Police Station had the lowest depression score, while call center counselors at Asan City Hall had the highest depression score [14].

Fig. 2 represents the network analysis of depression-related keywords after the COVID-19 pandemic. The top three keywords identified in this analysis are "COVID-19," "Mental Health Welfare Center," and "Ministry of Health and Welfare." This suggests that there has been an increase in experiences of depressive feelings and the prevalence of depression due to various social changes resulting from the COVID-19 pandemic.

B. Frequency Analysis Results

The frequency analysis of depression-related words before and after the COVID-19 pandemic revealed that "depressive disorder" had the highest frequency in both periods. However, there were additional keywords related to the COVID-19 pandemic (such as "COVID-19 outbreak," "corona," "coronavirus") in the post-pandemic period. It can be observed that the frequency and weight of keywords related to the COVID-19 pandemic are higher.

C. LDA Topic Modeling

1) LDA topic modeling design: Using Python, we calculated the coherence score, as shown in Fig. 3. The coherence score is a metric used to determine the optimal number of topics to be set for topic modeling. The number of topics with the highest coherence score is considered appropriate. In this study, based on the coherence scores, we set seven topics for the pre-COVID-19 pandemic period and six topics for the post-pandemic period.

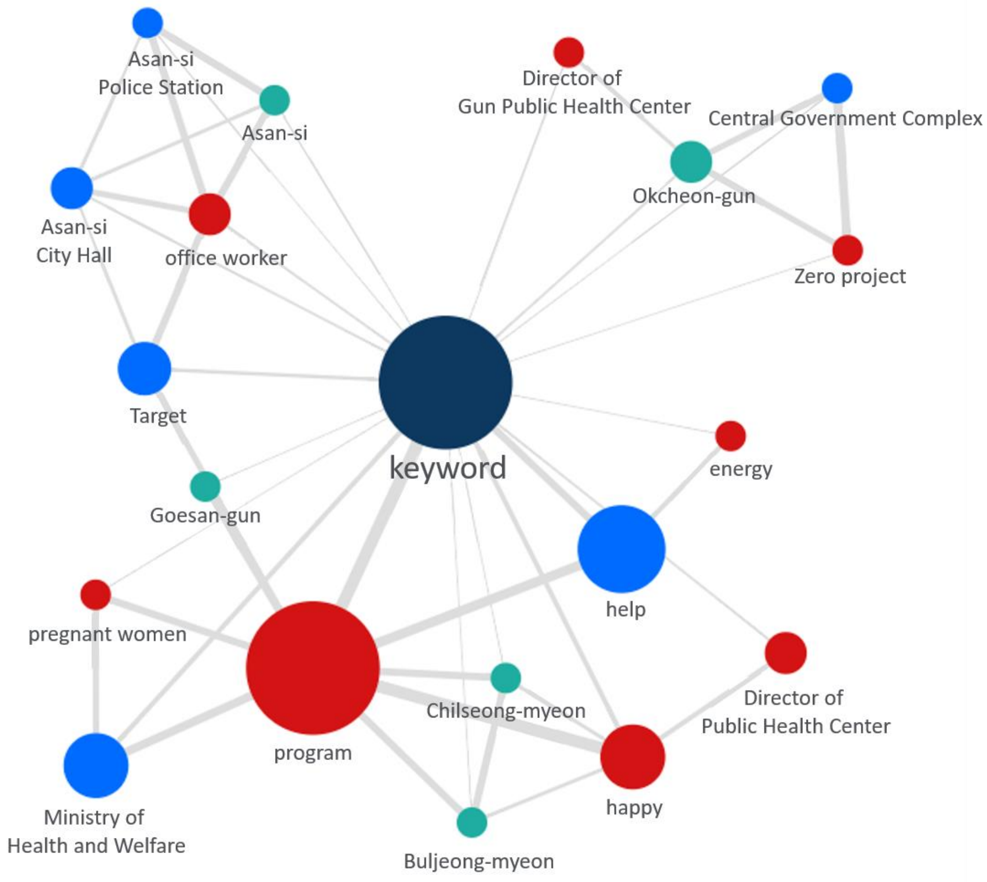

Fig. 1. Keyword network analysis of depression before the COVID-19 pandemic.

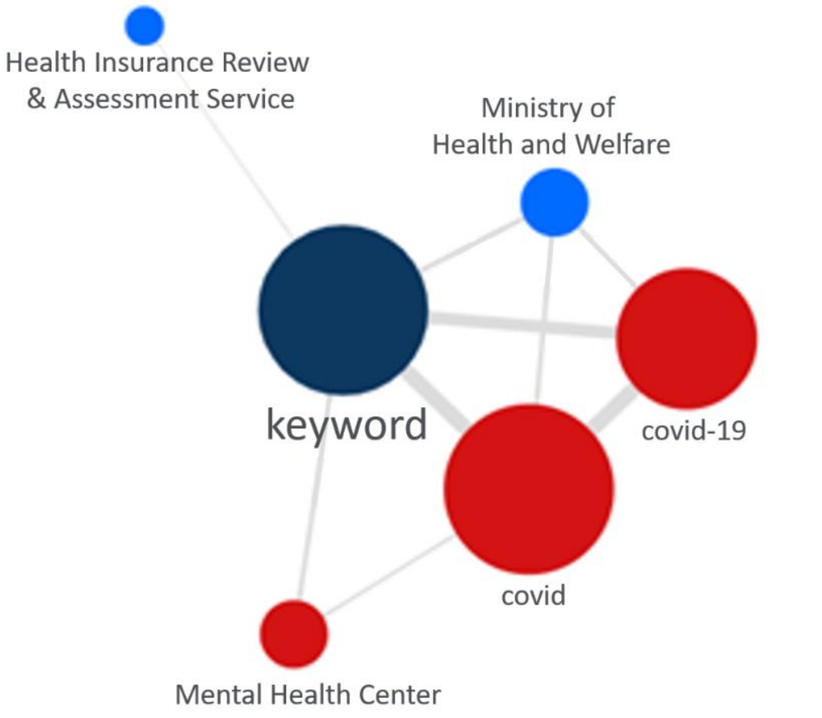

Fig. 2. Keyword network analysis of depression after the COVID-19 pandemic.

TABLE III. TOP 15 KEYWORDS BEFORE AND AFTER THE COVID-19 PANDEMIC

Before the outbreak of COVID-19 pandemic			After the outbreak of COVID-19 pandemic		
Word	Frequency	TF-IDF	Word	Frequency	TF-IDF
Depression	17733	16660.1469	Depression	13753	17836.1839
Seoul	7169	13227.3619	Covid-19 Pandemic	12413	17370.7382
USA	5668	11789.2733	Seoul	7745	14491.2603
Korea	4852	10846.0941	Corona Virus	6872	13679.5616
Depressive Syndrome	2558	7355.18821	USA	6009	12767.9171
Possibility	2474	7196.23002	Depressive Symptoms	5345	11982.8225
Victim	2423	7098.33421	Korea	4313	10594.2729
Japan	2280	6818.0422	Online	3773	9772.40313
Korea	2245	6748.09391	Police	3189	8795.90881
Person	2133	6520.54938	Victim	3188	8794.15013
Children	2125	6504.07473	Infectious Disease	2745	7982.6753
Police	2125	6504.07473	Select	2518	7539.80287
Myself	1899	6025.77282	Discovery	2388	7277.06911
Broadcast	1798	5803.49877	Possibility	2225	6937.58891
China	1786	5776.71894	Youtube	2001	6451.3782

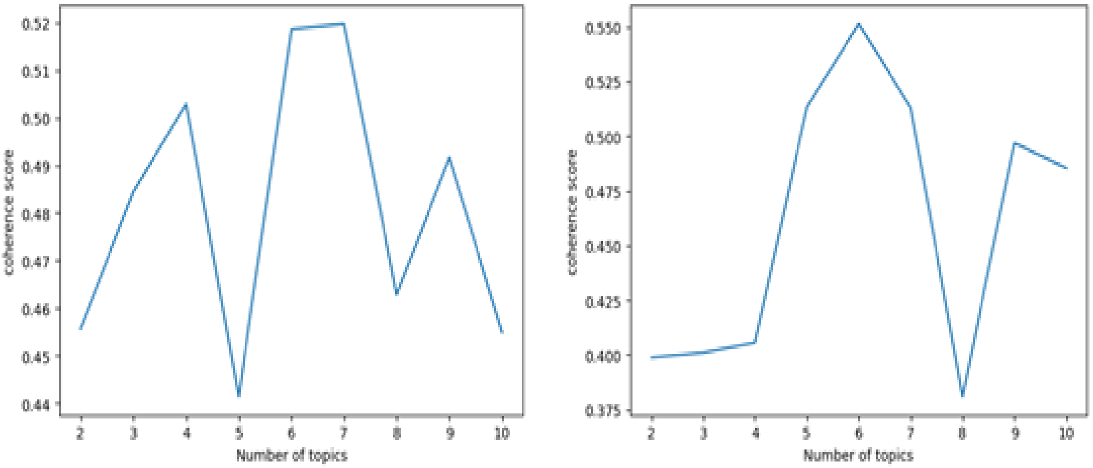

Fig. 3. Consistency score before and after COVID-19 pandemic outbreak.

TABLE IV. TOPIC MODELING BEFORE THE OUTBREAK OF COVID-19 PANDEMIC

Group	Topic name	Top 5 words
1	Depression of celebrity	Depression, Person, Broadcasting, Depression, Self
2	Depression and crime	Depression, Victim, Seoul, Court, Lawyer
3	Depression and female crime	Depression, Police, Investigation, Charges, Woman
4	Depression around the world	USA, Depression, Korea, UK, China
5	Overcoming depression	Confidence, Things, Friends, Viewers, Tears.
6	Depression in the general public	Seoul, Depression, Children, Korea, Office workers
7	Symptoms and causes of depression	Depression, Depression, Insomnia, Health, Stress

2) LDA topic modeling: In the LDA topic modeling results of this study, we have provided the top five words for each topic, as the weight of the 6th word in some topics was low. The results of topic modeling before the COVID-19 pandemic are presented in Table IV. The topics identified include famous individuals' confessions and overcoming depression, crime-related topics, global depression issues, and discussions on symptoms and causes of depression.

The results of the LDA topic modeling after the COVID-19 pandemic are presented in Table V. The differences between the topics before and after the COVID-19 pandemic can be observed, such as Topic 3 (COVID-19 pandemic and depression) and Topic 6 (depression and suicide), which specifically focus on the COVID-19 pandemic-related issues and suicide.

TABLE V. TOPICS MODELING AFTER THE OUTBREAK OF THE COVID-19 PANDEMIC

Group	Topic name	Top 5 words
1	Depression of celebrity	Depression, Agency, MBC (broadcasting), People, Fans
2	Covid-19 pandemic and depression around the world	United States, Depression, Korea, COVID-19 pandemic, Confirmed cases
3	Covid-19 pandemic and depressive symptoms	COVID-19 pandemic, Corona, Seoul, Depression, Coronavirus
4	Depression and crime	Depression, Seoul, Court, Victim, Defendant
5	Government and depression	Blue House, Immunity, Depression, Victims, Seoul
6	Depression and suicide	Police, Discovery, Selection, Investigation, Death

IV. DISCUSSION

In this study, we analyzed social issues related to depression before and after the COVID-19 pandemic using LDA topic modeling based on a large dataset of news articles (95,687 articles). We found that keywords related to "suicide" appeared prominently after the COVID-19 pandemic, unlike before the pandemic. The emergence of these keywords as potential key topics can be attributed to the prolonged duration of the COVID-19 pandemic, which has led to an increase in feelings of depression and suicidal ideation.

According to the Ministry of Health and Welfare, in 2020 when the initial wave of the COVID-19 pandemic occurred, there was a slight decrease in suicide deaths compared to 2019. This was attributed to factors such as a decrease in copycat suicides and social tension and cohesion resulting from a national disaster [15]. However, in 2021, both the number of suicide deaths and suicide rates increased by 1.2% compared to 2020 [16]. This suggests that with the prolonged duration of the COVID-19 pandemic, there has been an increase in feelings of depression and suicidal ideation [16], leading to an increase in youth suicide rates during the mid-term period of the COVID-19 pandemic [17].

The phenomenon of increasing youth suicide rates has also been reported overseas. In Japan, there was a 41.3% increase in youth suicides compared to 2019 [18], while overall adolescent suicides increased during the COVID-19 pandemic period in the United States [19]. Taking all these factors into account, it is speculated that there has been an increase in suicide deaths due to various social changes after the onset of the COVID-19 pandemic.

Another finding of this study is the discovery of topics related to the COVID-19 virus and depression after the onset of the COVID-19 pandemic. According to an online survey conducted by the Korea Health Promotion Institute targeting individuals aged 20 to 65, 40.7% of respondents reported experiencing "COVID-19 depression" [5]. In addition, a mobile survey conducted on individuals aged 15 and above in 17 metropolitan cities and provinces nationwide found that 47.8% of respondents reported experiencing depression and anxiety due to COVID-19 [20]. These survey results indicate that due to the prolonged duration of the COVID-19 pandemic, a significant number of people have experienced varying degrees of depression. Therefore, it is necessary for the government to explore measures to monitor mental health among all citizens on a national level, and prepare strategies to reduce depression and anxiety in future outbreaks or pandemics.

Although this study examined social changes related to depression before and after the onset of the COVID-19 pandemic using news big data and employed LDA topic modeling to identify issues, it has several limitations. Firstly, the data used in this study does not include time-series information, so we cannot examine trends over time. Therefore, future research should consider incorporating time-series analysis to observe changes in trends. Secondly, the collected data is limited to domestic news articles, so generalization to other cultural contexts may not be possible. Thirdly, the topic names derived from LDA are researcher-inferred labels; therefore, future research should consider applying algorithms that can complement this process.

V. CONCLUSION

In this study, we utilized BigKinds to analyze news articles related to depression before and after the onset of the COVID-19 pandemic using LDA topic modeling. We aimed to identify the main issues and differences between these periods. The results revealed that topics related to suicide emerged after the COVID-19 pandemic, as well as topics related to the COVID-19 virus and depression. Based on these findings, it is necessary for national-level policies to be developed in order to manage the mental health of citizens, such as depression, in future situations like infectious disease outbreaks.

ACKNOWLEDGMENT

This research Supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF- RS-2023-00237287, NRF-2021S1A5A8062526) and local government-university cooperation-based regional innovation projects (2021RIS-003).

REFERENCES

[1] “2022 White paper on suicide prevention”, Ministry of Health & Welfare, KOREA FOUNDATION FOR SUICIDE PREVENTION 2022.

[2] S. B. Hong, Restriction on SSRI (Selective Serotonin Reuptake Inhibitor) Antidepressant Prescription and Effort to Improve. Korean Journal of Family Practice, vol. 12, no. 4, pp. 212-216, 2022. doi: 10.21215/kjfp.2022.12.4.212.

[3] Gender general health examination mental health examination results by age by city and province, Statistics Korea, 2023. https://kosis.kr/statHtml/statHtml.do?orgId=350&tblId=DT_35007_N1180&conn_path=I2.

[4] Tackling the mental health impact of the COVID-19 crisis: An integrated, whole-of-society response, Organisation for Economic Co-operation and Development, 2021. https://www.oecd.org/coronavirus/policy-responses/tackling-the-mental-health-impact-of-the-covid-19-crisis-an-integrated-whole-of-society-response-0ccafa0b/

[5] 40.7% of Koreans said, "I experienced depression and anxiety due to COVID-19.”, Korea Health Promotion Institute, 2020. https://www.khepi.or.kr/board/view?pageNum=6&rowCnt=10&no1=55

3&linkId=1001456&menuId=MENU00907&schType=0&schText=&boardStyle=&categoryId=&continent=&country=&contents1=

[6] E. J. Lee, and S. J. Kim, Prevalence and Related Factors of Depression Before and During the COVID-19 Pandemic: Findings From the Korea National Health and Nutrition Examination Survey. Journal of Korean Medical Science, vol. 38, no. 10, pp. e74, 2023. doi: 10.3346/jkms.2023.38.e74

[7] National Medical Center, completes debate on proposed medical response directions to prepare for COVID-19 resurgence, National Medical Center, 2022. https://www.nmc.or.kr/nmc/bbs/B0000008/view.do?nttId=14764&menuNo=200394&pageIndex=1

[8] T. P. Hong, C. W. Lin, K. T. Yang, and S. L. Wang, 2013). Using TF-IDF to hide sensitive itemsets. Applied Intelligence, vol. 38, pp. 502-510, 2013. doi: 10.1007/s10489-012-0377-5

[9] L. He, D. Han, X. Zhou, and Z. Qu, The voice of drug consumers: online textual review analysis using structural topic model. International Journal of Environmental Research and Public Health, vol. 17, no. 10, pp. 3648, 2020. doi: 10.3390/ijerph17103648

[10] K. Bastani, H. Namavari, and J. Shaffer, Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Systems with Applications, vol. 127, pp. 256-271, 2019. doi: 10.1016/j.eswa.2019.03.001

[11] D. M. Blei, Probabilistic topic models. Communications of the ACM, vol. 55, no. 4, pp.77-84, 2012. doi: 10.1145/2133806.2133826

[12] Bigkinds. https://www.bigkinds.or.kr/

[13] B. H. Park: Okcheon-gun, Zero Depression Project Nationally Attended, Chungcheong Today, 2017. https://www.cctoday.co.kr/news/articleView.html?idxno=1106216

[14] S. H. Lee: 16% of Asan city officials 'at risk' of depression, asiatoday news, 2016. https://www.asiatoday.co.kr/view.php?key=20160824010012676

[15] 13,195 suicide deaths in 2020, down slightly from previous year, Ministry of Health & Welfare, 2021. https://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&CONT_SEQ=368016&page=3

[16] 13,352 suicide deaths in 2021, up slightly from last year, Ministry of Health & Welfare, 2022. https://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&page=2&CONT_SEQ=373035&SEARCHKEY=TITLE

[17] H. G. Woo et al., National Trends in Sadness, Suicidality, and COVID-19 Pandemic–Related Risk Factors Among South Korean Adolescents From 2005 to 2021. JAMA network open, vol. 6, no. 5, pp.e2314838-e2314838, 2023. doi: jamanetworkopen.2023.14838

[18] M. G. Kang, [Japan] COVID-19 and Suicide of Children and Youth. Changbi Children, vol. 19, no. 2, pp. 14-16, 2021.

[19] J. A. Bridge et al., Youth suicide during the first year of the COVID-19 pandemic. Pediatrics, vol. 151, no. 3, pp. e2022058375, 2023. doi: 10.1542/peds.2022-058375

[20] “Mental Health for the COVID-19 Generation!”, Gyeonggi Research Institute, 2020.