Changes in Topics During COVID-19 Within the Korean Sociology Studies

[a,1]

[a]*Department of Sociology, Korea University.*

Abstract. COVID-19 has completely changed people's lives. Academia seems to be no exception to this, so this study explored and analyzed changes in topics after COVID-19 by sociology researchers in Korea. As a result, it can be seen that sociology researchers in Korea have been conducting research that is more personal and emphasizes the welfare of minorities after COVID-19. In addition, research on national security has also increased. Conversely, research on macro-theories and ideologies, and macroeconomics and management, has declined. However, it is not known whether this trend is proportional to the intensity of COVID-19.

Keywords. Sociology, topic modeling, latent dirichlet allocation(LDA), COVID-19

1. Introduction

In the field of sociology, the study of topics is one of the interesting parts, because sociology covers many fields. One of the most interesting and extensive studies analyzing the topics of sociology is the study of Seol at al. (2018)[1], which investigated how the research topic of the “Korean Journal of Sociology” changed with the times. In this study, it is revealed that the main research topics of sociology have changed to urban and rural (1964-1979), social structure (1980-1999), culture and gender, and immigration (2000-2017).

Especially in the context of COVID-19, many parts of the world have been greatly affected. The same was true of academia, where scholars in many fields tended to reduce the number of research papers or collaborators. However, this trend did not appear significantly in the field of social science (Kang & Kwon, 2021)[2], which can be called a social puzzle. Even if the number of papers or collaborators did not decrease, it can be assumed that there must have been a change in some areas of sociological research.

In this study, to determine how sociology research has changed due to COVID-19, using keyword data from sociology papers collected by KCI and Latent Dirichlet Allocation(LDA), pre-COVID (2018-2019) and post-COVID (2020-2021) of sociology research topics will be compared. Additionally, the study will compare the first and second half of 2021, with significant differences in the number of people infected with COVID-19, to see how sociological research topics have changed during the COVID-19 period.

[1] Corresponding Author,, Department of Sociology, Korea University, Seoul, Korea; E-mail:

2. Data Source

2.1 KCI-listed Papers Data

Korea Citation Index (KCI, 한국학술지인용색인) is a database of Korean research papers. KCI provides various paper metadata such as Korean and English titles, publishing organization names, and detailed fields as Excel files. The detailed properties are as follows.

- NO : The order of the downloaded records, not a unique key, but simply a column number.
- 논문 ID(paper ID) : The unique number of the article.
- 논문명/논문 외국어명(title, title in English)
- 저자/공동저자(author, co-author)
- 학술지 명/학술지 외국어 명(journal, journal in English)
- 발행기관 명/발행기관 외국어 명(publisher, publisher in English)
- 등재구분(listedness)
- 발행년(year)
- 권(호)/페이지(volume & number, pages)
- 키워드(한국어)/키워드(외국어)(Korean keywords, keyword in foregin language) : It should be noted that, depending on the paper and the publisher, various languages such as English, Japanese, and Chinese are included in the keyword column. Each keyword is separated by a comma.
- 주제분야(subject area) : Refers detailed field of paper.
- 피인용횟수(number of citations)
- URL, DOI

I used these metadata of papers, since KCI doesn't support unauthorized batch download of papers. So I downloaded metadata that belong to *Sociology*(사회학, N=1,986) and *Social Science in General*(사회과학일반, N=6,013). Therefore, in this data, papers classified separately from sociology such as business administration, tourism, education, anthropology, law, and so on are omitted. Also, I only used KCI-listed papers (including excellent listings), mainly to target only papers published in journals considered meaningful. In this data, as can be seen from the list above, the month in which the papers were published is missing, so I manually added the month column after downloading the papers by sorting them by time.

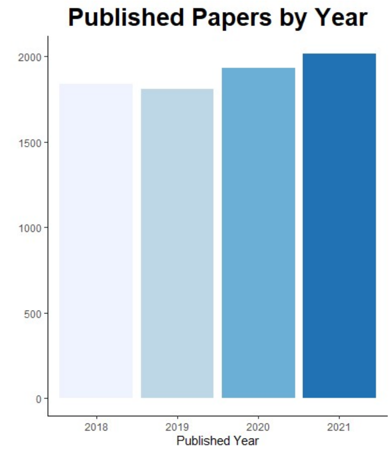


Figure 1. Published Paper by Year

During the study period, the number of published papers increased substantially, except for 2019. In 2019, the smallest number of papers was published, 1,810, and in 2021, the highest, 2,019 papers were published.

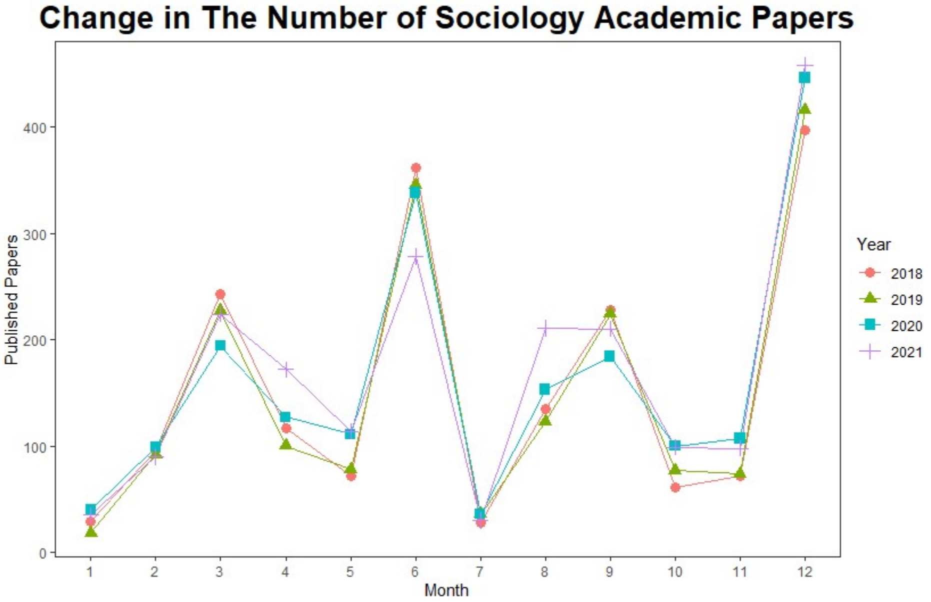


Figure 2. Change in the Number of Papers by Year and Month

Above Figure(Figure 2) shows changes of the number of papers in each year. There is a pattern in the publication of papers that gradually increases every three months. This point is consistent with the study of Kang & Kwon (2021, p. 196), who stated that the unit of the paper publication cycle was 3 months. In particular, when the publication of each paper is compared by 6 months, the number of publications increased during the first 3 months, decreased at the 4th month, similar or slightly decreased at the 5th month, and then significantly increased at the 6th month.

2.2 COVID-19 Data

To get COVID-19 data, I downloaded the COVID-19 confirmed and death cases csv file by country and date from the World Health Organization (WHO). The data columns are as follows.

- Date_reported : Year-month-day data.
- Country_code, Country, WHO_region : Regional data such as countries with cases.
- New_cases, Cumulative_cases : New or cumulative cases of COVID-19 confirmed infection.
- New_deaths, Cumulative_deaths : New or cumulative deaths by COVID-19.

In this data, I added year and month columns through date data, and extracted only data whose country is Republic of Korea (country code: KR). So I could get the Korean COVID-19 data which is grouped by year and month.

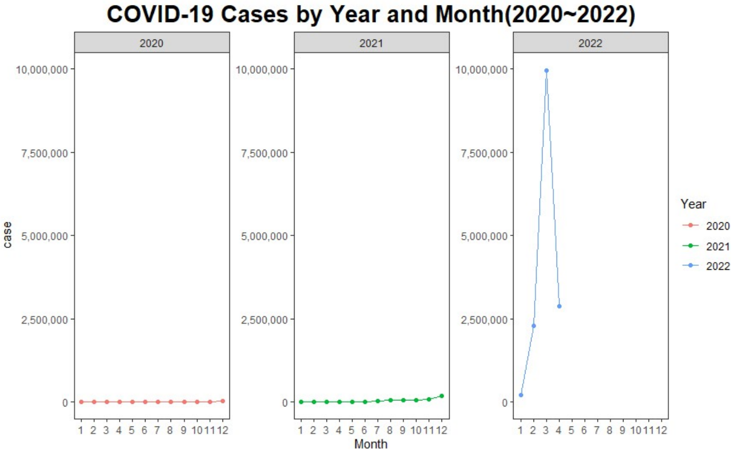


Figure 3. COVID-19 Cases by Year and Month(2020-2022)

As we can see, Data for 2022 is too large to make data for other years meaningless. In addition, since 2022 is still in progress, it is difficult to see the trend, so it should be removed.

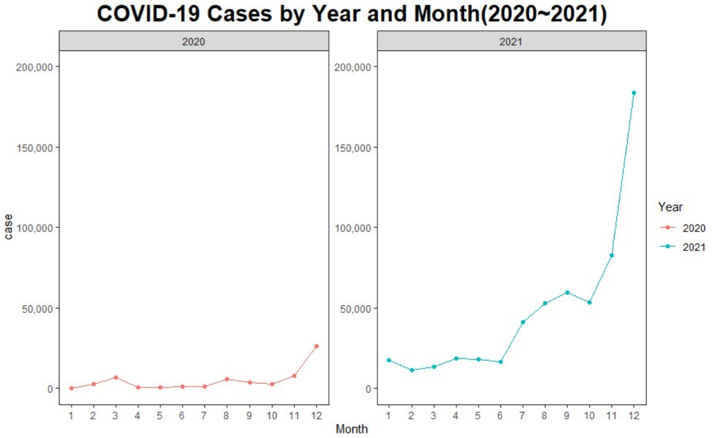


Figure 4. Published Paper by Year(2020-2021)

In this case, we can se COVID-19 is surged in the second half of 2021. In December 2021, the maximum for the above data, 183,608 people were infected with COVID-19. Conversely, 6,855 people infected with COVID-19 in March 2020 is a minimum. As of December 2021, the cumulative number of infected people is 630,835.

2.3 Data Preprocess & Methodology

Since there are a lot of non-Korean keywords in the Korean keywords column, I translated it to Korean. In this case, I used Python not R, because Python has a module with relatively wide quotas. Major translation APIs like Kakao and Google have so strict quota that I can't manage large amounts of records (at least for free). So, I use Googletrans and Numpy, Pandas module in Python to translate.

In addition, I removed all acronyms in the keyword column. Since acronyms are usually presented with the original word (like Korean Citation Index(KCI)), this is to avoid overestimating the word. So I used R to remove the acronyms and parentheses presented later when a keyword is given in the form of "original word (acronym)".

Finally, I splitted words according to commas, and used KoNLP, Sejong library and NIA dictionary to extract nouns to analyze topics. And used topicmodels library to find the optimal number of topics(four parameters, Griffiths2004, CaoJuan2009, Arun2010, Deveaud2014), lda library to conduct LDA(hyperparameters, α=.01, η=.001, Burn-in=500, Iteration=5,000), and LDAvis library to visualize it. In the process of making the topic model, the topic model of the entire data can also be used for the data of each year, but for this purpose, unless there is a larger amount of data, it seems difficult to properly analyze the topics of each year because there are many heterogeneous parts. Therefore, I decided to compare the keywords after making a model for each year. And for visualization, LDAvis provides visualization in the form of JavaScript. Since the createJSON function of the LDAvis library does not take non-English characters into account, I manually changed the encoding of each JSON file to UTF-8.

3 Results

3.1 Finding the Optimal Number of Topics

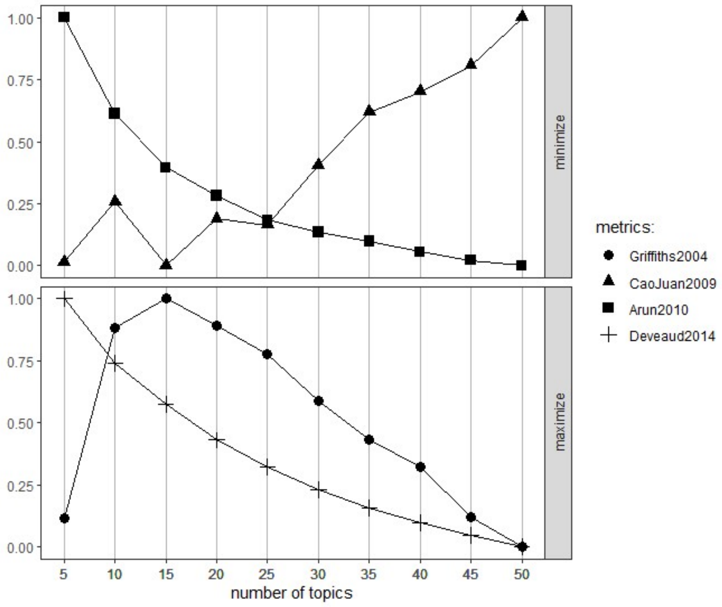


Figure 5. Metrics to Find the Optimal Number of Topics(2018-2021, 5-50)

First, when finding the optimal number of topics, I need to pick the point where Griffiths2004 and CaoJuan2009 are the lowest, and Arun2010 and Deveaud2014 are the highest, but there is a discrepancy between these values. So I decided to look more closely at the points where the number of topics was 5 to 15 in order to compromise and find concrete values.

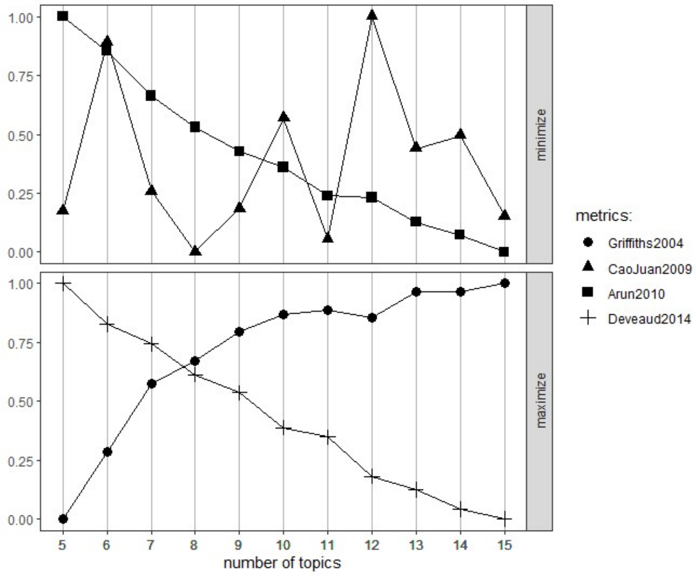


Figure 6. Metrics to Find the Optimal Number of Topics(2018-2021, 5-15)

As a result, there were still conflicts between the values, but since those values remain reasonably close to optimal at 7, I decided to set the number of topics for the 2018-2021 period to seven.

On the other hand, this trend was also the same between the other years. Therefore, the number of topics in each year was set to 7. It also facilitated comparative analysis between each year. In all cases, the number of topics was 5 to 50, and then 5 to 15 points were compared, and the result is as follows. Because the different metrics show constant variation (ie, when one value is optimized, the other is less optimized), I decided on the value, heavily influenced by CaoJuan2009, where the number of topics varied from 5 to 15 irregularly.

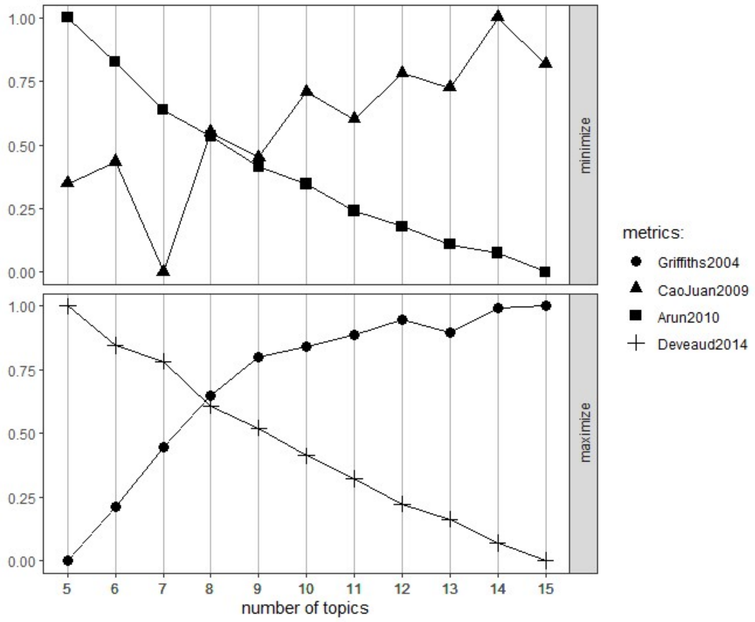


Figure 7. Metrics to Find the Optimal Number of Topics(2018-2019, 5-15)

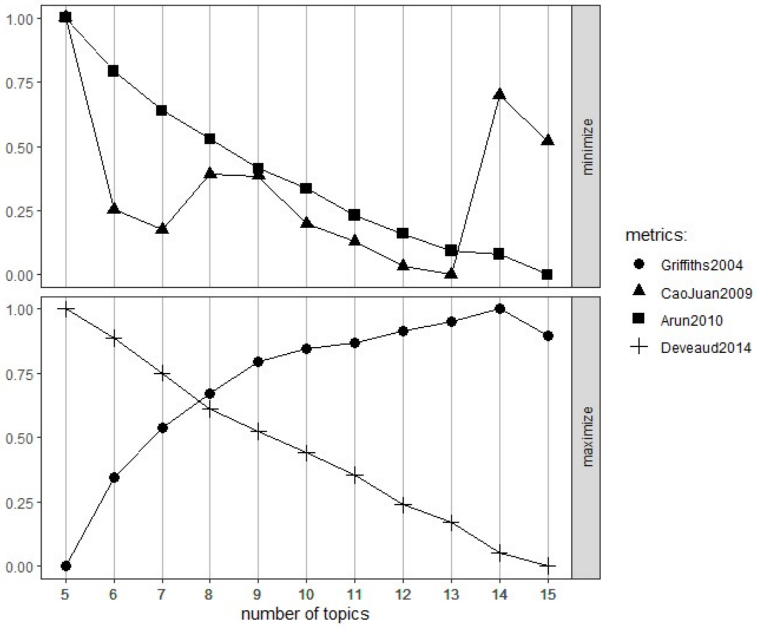
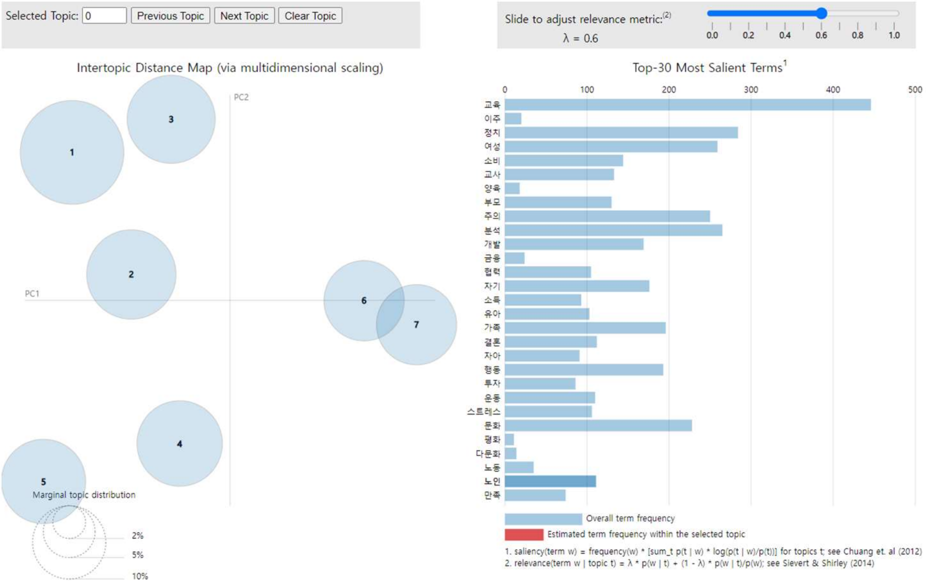


Figure 8. Metrics to Find the Optimal Number of Topics(2020-2021, 5-15)

3.2 Topics in the Whole Period and the Pre/Post-COVID Period

The overall topic trend from 2018 to 2021 is shown in the figure below. It is difficult to read all the results for each topic, so it can be organized as shown in the table below. Table 1 summarizes notable words among the 30 most salient words in each topic and arbitrarily determines topic names based on them..



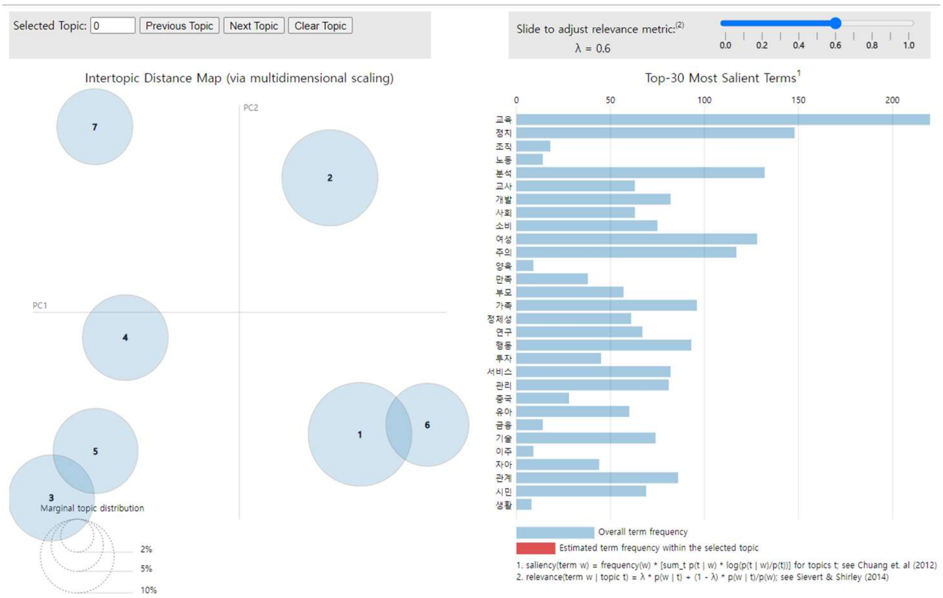
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 9. LDAvis Output(2018-2021)

Table 1. Topics of 2018~2021. Notable Words of the 30 Most Salient Words(λ=.6).

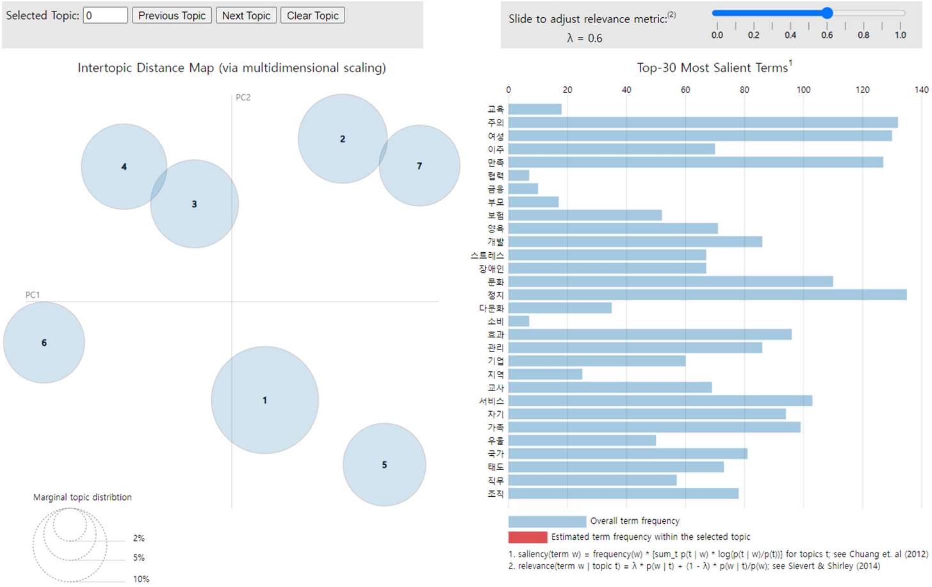
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**
percentage	20	14.8	14.4	13.6	13.2	12	12
terms	교사, 양육	소득, 돌봄	이주, 여성	개발, 협력	소비, 금융	주의, 운동	평화, 정치
	부모, 유아	빈곤, 정책	다문화, 인권	디지털, 연구	투자, 재무	마르크스	테러, 북한
	교육, 가족	보험, 일자리	장애인, 북한	미디어, 산업	기업, 은행	시민, 연대	안보, 국제
Arbitrary Name	**Life Cycle & Education**	**Welfare Policy**	**Minorities**	**Industry**	**Economy & Business**	**Civil Politics & Ideology**	**Diplomacy & Security**

Meanwhile, how about topics before and after COVID-19? To find it, see Figures and Tables below.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 10. Pre-COVID LDAvis Output(2018-2019)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 11. Post-COVID LDAvis Output(2020-2021)

Table 2. Topics of Pre-COVID and Post-COVID(Ordered and Matched Arbitrary). Notable Words of the 30 Most Salient Words(λ=.6).

Pre-COV	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
percentage	19.7	16.9	13.6	13.4	13.2	12.6	10.6
terms	양육, 가족 부모, 행동 자아, 아동	소비, 투자 금융, 서비스 감사, 재무	정체성, 정치 주의, 민족 젠더, 국가	분석, 연구 과학, 데이터 미디어, 개발	소득, 난민 복지국가, 노동, 정책	교육, 교사 직무, 역량 조직, 리더십	사이버, 안보 원조, 비핵화 북한, 협력

Arbitrary Name	Life Cycle & Psychology	Economy & Business	Civil Politics & Ideology	Industry	Welfare & Labor	Education & Job	Diplomacy & Security
Post-COV	**Topic 1**	**Topic4**	**Topic 7**	**Topic 3**	**Topic 6**	**Topic 5**	**Topic 2**
percentage	21.1	13.4	12	14.2	12.1	12.1	14.6
terms	부모, 양육	기업, 투자	주의, 북한	소득, 소비	금융, 요양	이주, 여성	협력, 평화
	우울, 만족	플랫폼, 소득	페미니즘	공정성, 지방	보험, 의료	교육, 학교	국제, 원조
	노인, 청소년	재무, 비용	이데올로기	신뢰, 복지	관리, 은행	인권, 교사	안보, 외교
Arbitrary Name	**Life Cycle & Psychology**	**Economy & Business**	**Civil Politics & Ideology**	**Welfare Policy**	**Welfare**	**Education & Minorities**	**Diplomacy & Security**
Change	1.4%p↑	3.5%p↓	1.6%p↓	Replaced	Controversial	Terms Changed	4%p↑

First, there is 1.4%p increase of interest in life cycle and psychological topics(Topic 1, 1). This is not a huge number, but considering it was close to 20% before COVID-19, it can be considered a significant increase. In addition, the content on diplomacy and national security (Topics 7, 2) has increased significantly. This topic increased its ratio from 7th to 2nd with a 4%p increase.

In the business and economic sectors (Topics 2, 4), there was a decrease of 3.5%p, and in civil politics and ideology (Topics 3, 7) there was a slight decrease of 1.6%p. Also, Topic 4 of pre-COVID was about industry, and there was no content related to this among topics of post-COVID.

Regarding Welfare(Topic 5, 6), although the proportion of the similar topic has decreased, interest in welfare may actually have increased as the number of topics mentioning terms related to welfare has increased to two. And, in pre-COVID, welfare included a lot of terms related to labor, but in post-COVID, social values discussed in terms of fairness, trust, and policy, and content on universal welfare were more included.

Finally, with regard to education, the ratio itself did not change significantly with the drop of 0.5%p. However, before COVID-19, words related to education topics were mainly about jobs or capabilities, whereas after COVID-19, words related to social minorities in education have appeared a lot.

3.3 Change Within a Year(2021)

In the previous analysis, we studied changes before and after COVID-19. So, are these changes happening in real time even during COVID-19? In other words, will research topics change in proportion to the severity of COVID-19? To confirm this, I compared the research topics of the first half (January to June) and the second half (July to December) of 2021.

Table 3. Topics of First Half and Second Half of 2021(Ordered and Matched Arbitrary). Notable Words of the 30 Most Salient Words(λ=.6).

2021-1	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
percentage	16.4	15.2	14.9	14.2	13.7	13	12.5
terms	생활, 만족	산업, 가치	이주, 여성	교육, 청소년	정책, 시민	협력, 국제	서비스, 관리
	우울, 노인	관광, 구매	노인, 이민	다문화, 교사	정부, 업무	중국, 외교	감사, 공시
	청년, 청소년	문화, 에너지	코로나 19	교수, 장애	기관, 신뢰	미디어, 개발	여성, 은행
Arbitrary Name	**Life Cycle & Psychology**	**Industry**	**Minorities**	**Education & Minorities**	**Policy**	**Diplomacy**	**Economy & Business**
2021-2	**Topic 1**	**Topic 7**	**Topic 4**	**Topic 2**	**Topic 3**	**Topic 6**	**Topic 5**
percentage	17.6	12.2	14.2	14.4	14.4	13.2	14.1

terms	자아, 만족 아동, 노인 행동, 우울	종교, 폭력 전환, 에너지 공공성, 제도	돌봄, 이주 자녀, 여성 청년, 노동	교육, 학습 전문, 어린이 학부모, 유아	주의, 신뢰 이론, 담론 체제, 자유	개발, 주민 테러, 안보 북한, 글로벌	소비, 경제 보험, 공간 지방, 서비스
Arbitrary Name	**Life Cycle & Psychology**	**Civil Politics**	**Minorities**	**Education**	**Ideology**	**Diplomacy & Security**	**Economy & Business**
Change	1.3%p↑	Replaced	0.7%p↓	0.2%p↑	Replaced	0.2%p↑	1.6%p↑

The result, as shown in Table 3, did not reveal much. For most topics, the figures did not change significantly, and several topics were replaced with heterogeneous ones between the two periods. However, among them, the study on the life cycle and psychology slightly increased, similar to the research before and after COVID-19.

4 Conclusion

Through the study, it was found that the topics before and after COVID-19 show the following trends.

- Increased interest in individual life and psychological factors.
- Decreased interest in economic activities such as bus iness, investment and labor.
- Interest in real politics, welfare policy, diplomacy, and national security has increased.
- On the other hand, interest in abstract grand theories and ideologies has declined.
- Interest in education did not change, but in pre-COVID, it was associated with occupational activity, and in post-COVID, it was associated with minorities.

Therefore, after COVID-19, sociologists have been doing more research on personal lives or people who do not enjoy basic well-being due to COVID-19, and conversely, less research on theories and ideologies from a macro perspective. In addition, research on international relations and national security has increased, which can be seen as a tendency to keep in mind conflicts between countries.

However, this trend did not appear in the comparison within 2021, when COVID-19 was severe, either because there was not enough data or because the types of papers published in the first and second half of the year were different. Also, this result may have occurred because the impact of COVID-19 can be expected to affect papers published after a certain period of time, rather than papers currently being published.

This study has the following limitations. First, in considering COVID-19 as a variable, only approximate trends and time variables were used. To contrast the changes, it is necessary to establish an additional study period prior to COVID-19 to compare. It is also necessary to control the variables that can influence the topics. Second, since I used LDA as my research methodology, domain knowledge was used to name the topic. This part basically overcame the limitations of LDA, which is unsupervised learning, but at the same time, the possibility that the topic was set according to my discretion cannot be excluded. Finally, instead of comparing the years by setting topics according to the entire period, a different model was created for each year, because the number of keywords was too small compared to the heterogeneity of the keywords in the papers.

Therefore, in the future, it is necessary to improve this point to extend the study period, to include research in other fields of social science, or to analyze the main text of the paper beyond keyword metadata.

References

[1] D.H. Seol, J.H. Ko, & S.H. Yoo, Korean Sociological Association and Sociological Research: Changes in the Areas of Sociology in Korea 1964-2017, *Korean Journal of Sociology* 52 (2018), 153-213.

[2] J.H. Kang, & E.R. Kwon, The Effect of COVID-19 Pandemic on Research Productivity in South Korea: A Comparative Analysis of Korean Journal Articles across Academic Fields, Korean Journal of Sociology 55 (2021), 179-199.