Asian Journal for Public Opinion Research - ISSN 2288-6168 (Online)
Vol. 10 No.1 February 2022: 4-22
http://dx.doi.org/10.15206/ajpor.2022.10.1.4

Topic Modeling of Korean Newspaper Articles on Aging via Latent Dirichlet Allocation[1]

So Chung Lee[2]

Namseoul University, South Korea

Abstract

The purpose of this study is to explore the structure of social discourse on aging in Korea by analyzing newspaper articles on aging. The analysis is composed of three steps: first, data collection and preprocessing; second, identifying the latent topics; and third, observing yearly dynamics of topics. In total, 1,472 newspaper articles that included the word "aging" within the title were collected from 10 major newspapers between 2006 and 2019. The underlying topic structure was analyzed using Latent Dirichlet Allocation (LDA), a topic modeling method widely adopted by text mining academics and researchers. Seven latent topics were generated from the LDA model, defined as *social issues*, *death*, *private insurance*, *economic growth*, *national debt*, *labor market innovation*, and *income security*. The topic loadings demonstrated a clear increase in public interest on topics such as *national debt* and *labor market innovation* in recent years. This study concludes that media discourse on aging has shifted towards more productivity and efficiency related issues, requiring older people to be productive citizens. Such subjectivation connotes a decreased role of the government and society by shifting the responsibility to individuals not being able to adapt successfully as productive citizens within the labor market.

Keywords: aging, social discourse, newspaper, topic modeling, latent Dirichlet allocation, South Korea

Background

[1] Funding for this paper was provided by Namseoul University year 2020.

[2] All correspondence concerning this article should be addressed to So Chung Lee at Namseoul University 91 Daehak-ro, Seobuk-gu, Cehonan-si, Chungcheongnam-do, South Korea or by email at snowvill@nsu.ac.kr.

Korea has experienced dramatic changes in population aging, along with rapid socioeconomic development. Declining fertility and increased life expectancy have had a major impact on population aging. In 2020, the proportion of the population aged 65 and over has reached 15.7% (Statistics Korea, n.d.), increasing each year at a rapid rate. With the aging of Korean baby boomers, born 1955-1963, population aging is gaining even more speed and the over-65 population is projected to reach 20% by 2025, according to Statistics Korea.

The issue of rapid population aging was raised only after the millennium, when the government became alarmed about its speed. The seriousness of the phenomenon accelerated the implementation of three five-year basic plans; nonetheless there was a lack of social consensus about how to address this issue. Looking at the precedents of developed countries that experienced population aging earlier, how the public recognizes aging could be extremely important in that not only could it lead to indifferent and unfavorable opinion towards policies but also could lead to serious social discordance such as intergenerational conflict and gerontophobia. Thus, this study aims to analyze how Korean society recognizes aging and how social discourse on aging has changed periodically with the purpose of suggesting priority issues and policy focus that could contribute to building a cooperative and inclusive super-aging society.

However, it is not an easy task to recognize public opinion and social discourse on aging because aging covers vast range of issues and perspectives.

One significant method to track public opinion is to analyze the media. In modern society, media plays a critical role in forming public perceptions of reality. In particular, newspaper articles are expected to maintain objectivity and as such, are often accepted as reality itself; public opinion may be affected and formed by discourses produced by newspapers (Jang, 2013). Thus, by analyzing the discourse on aging produced by the media, it is possible to uncover public opinion and the main elements of discourse on aging. Owing to technological development in an era of Industry 4.0, vast amounts of media texts have been digitized; the methodology by which fundamental insights and solutions may be drawn from such materials have been elaborated. This study aims to take advantage of such technological benefits and explore the structure of social discourse on aging by analyzing newspaper articles on aging, with the ultimate purpose of providing future policy implications.

Theoretical Background

Social Discourse and Media

Studies on discourse typically regard discourse as a social practice of participants communicating through linguistic and semiotic resources under certain contexts. Discourses are composed of vocabularies and rhetorical techniques as well as modes of thinking, grammars, rationalities, and even specific material practices that represent, interpret, and create new reality. Thus, discourses are seen as frameworks for understanding and directing different domains of social action.

Although various strands could be found within theories on discourse,[3] they generally articulate three problems: power, knowledge and subjectivity. Not only is discourse shaped by power structures, but it also contributes to objectifying and constituting power structures by representing them. As a socially-situated activity of producing meaning, discourse produces and legitimizes knowledge within society. Furthermore, discourse plays a crucial role in constructing subjectivity as it defines, identifies, and creates actorhood by attributing places and positions to those who enter discourse (Angermuller, 2015). Such a viewpoint follows the works of Foucault, who refers to "discourse" as ways of constituting knowledge, together with social practices, forms of subjectivity and power relations which inhere in such knowledges and relations between them (Diamond & Quinby, 1988). Foucault focuses on questions of how some discourses have created meaning systems that dominate the way we define and organize ourselves and our social world while other discourses have been marginalized and subjugated.

Thus, discourse is ideological, reflecting inequal power relations within society, and the main objective of analyzing discourse lies in disclosing the ideological structure that lies beneath discourse, mostly presented in the form of text.

A majority of the research on discourse focuses on the significance of media text. Media in modern society not only provides knowledge and information but also forms individual's perception and practice within society. Media selects or excludes, as well as emphasizes or conceals information following certain ideological criteria. In this sense media is the site where competing discourses on important social agenda collide, gain

[3] Angermuller (2015) suggests that at least three strands could be distinguished within discourse theory: post-structuralist, normative-deliberative, and critical-realist discourse theories.

dominance, and are dispersed and consumed. Therefore, considering the significant role media is playing in modern society, it is reasonable to take media discourse as a proxy for social discourse and public opinion.

Research Analyzing Media Discourse on Aging in Korea

Among the vast range of previous research on aging, it is disappointing to find very little research analyzing the media discourse of aging in Korea. Han and Yoon (2007) take the initial step in this area. They critically review 1,725 newspaper articles from 3 major newspapers between 1997-2006 containing either "aging" or "elderly" within the text. They found an increase in the articles portraying positive images of aging in Korea by introducing new discourse such as "successful aging," which emphasizes a productive and active lifestyle for the elderly. On the one hand, they welcome the change in trend brought by successful aging discourse, altering the negative images of elderly as a dependent burden on society. On the other hand, they warn that an overemphasis on successful aging might lead to stereotypes of the elderly and unintentionally marginalize certain groups of people who cannot meet the standards, i.e., disadvantaged elderly or elderly women in low socioeconomic strata.

Kim (2017) starts by questioning how society recognizes the elderly and conducts media discourse analysis on aging. In the era dominated by neo-liberalist ideology, all human existence is driven into infinite competition, forcing people to increase their market value. The research suggests that elderly people are not exceptional in this respect. In order to be a welcoming member of society, the media imposes social discourses on elderly people to be functional as socio-economic agents within society. Independence is a proxy for a wonderful later life, subjective youth as a core value of old age underlies being economically productive throughout one's lifespan. The study concludes that such media discourse may conceal the societal responsibility of caring for the elderly and marginalizing a majority of older people. Although Kim's (2017) study was published almost a decade after Han's (2007) work, both have the same context.

Lee and Kim (2019) conducted analysis online social media posts from 2004 to 2017 related to aging and the elderly, with the purpose of identifying major themes and temporal trends of discourse on the elderly. Prior to analysis, they theoretically identified significant factors affecting the lives of older people and generated four domains, which are economy, health, relationships, and culture. Based on literature review, they constructed a noun dictionary for each domain. Web crawling and social network analysis was implemented. The analysis found that quality of life of older people is placed at the center of the discourse and has a strong association with

economic welfare and health status. Poverty, jobs, and dementia appeared as the most salient themes. Furthermore, a sense of alienation and isolation in interpersonal relationships was closely related to mental health and quality of life.

In summary, previous research was fruitful in supporting the idea that the media plays some role in framing the discourse or image of older people. However, analyzing large media texts focusing on the underlying discourse structure on aging is yet an unexplored field.

Method

Data Collection

The data analyzed in this article was collected from the newspaper digital archive, Big KINDS (Korea Integrated Newspaper Database System), provided by the Korea Press Foundation. Big KINDS is a news analysis service that provides metadata information by combining a big data analysis technique and news media. It provides over 6.5 million news contents collected from 54 media of all kinds, such as daily newspapers, financial newspapers, local newspapers, and newscasts dating from 1990. Due to copyright issue, Big KINDS provides metadata files consisting of keywords and feature-extracted words generated by morpheme analysis instead of the original text. The keywords contain all the words and phrases extracted from the original article while most of the stop words[4] are excluded. On the other hand, feature-extracted words are derived by applying the Text Rank algorithm to the keywords generated by morpheme analysis with the purpose of suggesting word groups of higher importance (Park et al., 2017).

Articles were identified with a search algorithm that includes the word "aging" in the title between 2006 and 2019, a time period during which the first, the second, and the third government "Basic Plan on Aging Society and Low Fertility" were put into action. Financial newspapers and local newspapers were excluded under concern of deviation towards either economic or local issues. Collected articles were screened manually to ensure that they fit the inclusion criteria and as a result, a corpus of 1,472 texts was collected. The number of articles varied by year, showing the ebb and flow of media interest in aging issues. As previously mentioned, Big KINDS provides metadata

[4] Stop words are words without much meaning or information such as prepositions, pronouns, etc. These words are filtered out in processing natural language data in order to save space and time.

for each article and this study used the keywords of each document for analysis. By so doing, it was possible to avoid the labor of noun extraction and stop word elimination.

Analysis Tools and Techniques

All data handling, screening, preprocessing, and analysis was performed via R version 3.6.2. The main R packages used for analysis were dplyr, ggplot2, KoNLP[5], lda, ldatuning, stringr, tidytext, tm, and topicmodels.

The analysis for this study is composed of three stages: first, tidying the text and pre-processing; second, topic modeling; and third, time-series analysis of changes in topics.

Pre-processing is composed of stripping empty spaces, eliminating numbers, punctuation, special characters, and, last but not least, extracting stop words (Baek, 2019). This study used the tm package and stringr package to accomplish basic text tidying job. While pre-processing is generally acknowledged as the most burdensome yet critical component of successful topic modeling, most of the burden can be avoided by using the pre-processed keywords provided by Big KINDS. However, because the keywords are mainly nouns extracted from the raw material based on morpheme analysis, some words and phrases may lose their original meaning being decomposed into morphemes. This study combined important word chunks and phrases such as "baby+boomers," "salary+peak," "national+pension," "social+insurance," "elder+abuse," etc.,[6] detected by running frequency to avoid losing meaning.

Text mining is basically a method to determine what a document is about by quantifying the words that make up a document. A simple way to decide how important a certain word is to the document is to look at its frequency. However, the texts at hand are from the same theme (generally the study theme) and thus share words in large frequency that could not be regarded as stop words (in the case of this study: aging, elderly, fertility, etc.). When extracting topics from the set of documents (corpus), these words could act as noise, disturbing exact analysis and interpretation in the same way

[5] The text data for this study is in Korean and thus KoNLP package, a morphological analyzer for Korean text-based research, was used.

[6] The importance of word chunks was decided upon based on theory and policy (Basic Plan on Aging Society and Low Fertility). It is agreed upon that pre-processing and interpreting textual data is a complicated job and, in many cases, requires a lot of supervised approach to use textual data effectively.

as stop words. Applying the statistic tf-idf is a useful way to solve this problem. Tf-idf[7] is intended to measure how important a word is to a document in a corpus by decreasing the weight for frequently used words while increasing the weight for those less commonly shared within the corpus but having significance within a specific document. The tf-idf score is 0 or near 0 for common words (Silge & Robinson, 2017). This study used the tidytext package to calculate the tf-idf score for each word in the corpus. The calculation showed 30 words with tf-idf value less than 0.01, and all 30 words were removed from the corpus.[8]

The second and focal part of analysis, topic modeling, is a valuable method for identifying the linguistic contexts that surround social institutions or policy domain (DiMaggio et al., 2013). There are several kinds of topic modeling techniques among which Latent Dirichlet Allocation (LDA) is especially popular. The LDA model is a probabilistic model that identifies sets of words, or "bags of words," that co-occur across documents. LDA is called a "topic model" because the identified sets of words tend to reflect underlying topics that, in combination, characterize every document in a corpus (Blei et al., 2003; Blei & Lafferty, 2006; McFarland et al., 2013; Liu et al., 2016).

One of the main challenges of topic modeling is to identify the number of topics that are latent in the corpus. There have been a number of approaches to validate the number of topics, one of which is to rely on model fitness scores of a given number of topics calculated via the FindTopicNumber() function.[9] As Baek (2019) claims, the FindTopicNumber() function in the ldatuning package is convenient to use not only because it is possible to input plural k (number of topics) simultaneously, but also because it calculates four different model fitness scores at the same time, allowing for a

[7] The statistics tf-idf is theoretically based on Zipf's law, stating that the frequency of a word is inversely proportional to its rank.

[8] The list of 30 words removed are as follows (in increasing order of tf-idf score): aging, Republic of Korea, low fertility, our country, Seoul, Bureau of Statistics, municipal government, spouse, systematic, elementary school, jobs, living costs, OECD (in Korean), OECD (in English), daycare center, persons concerned, report, continually, Gyeonggi province, youths, researcher, developed countries, families, public institutes, age group, point, the Bank of Korea, commission, elderly, fertility rate.

[9] This study chose the ldatuning package for assessing the fit statistics of different levels of topic k. It is necessary to take two steps in order to get the optimal number. First, to input the largest possible number of k with a large interval (i.e., the number of k starting from 1 till 30 at the interval of 5) and depending on the output, to narrow the number of k and the interval until concluding the optimal number.

more comprehensive determination of the number of latent topics. The four model fitness scores estimated in the ldatuning package are Griffiths2004, Deveaud2014, CaoJuan2009, and Arun2010. The four of them differ in their details but share the same objective of calculating the number of topics that group words in a most distinctive way.[10]

The final part of this study measures the load of topics for each year (as sums of those word-sets, or number and percent of documents using those words), to plot time-serial changes. To observe such time-series variation of latent topics, this study employed the same analytic model previously implemented by Roh and Yang (2019). Each document (newspaper article) was assigned to a topic based on LDA estimation, specifically the gamma value, which indicates the degree of affiliation of a document to the topics. Each document was assigned to the topic with largest gamma value. Nevertheless, documents with a largest gamma value smaller than average, were omitted from topic assignment.

Results

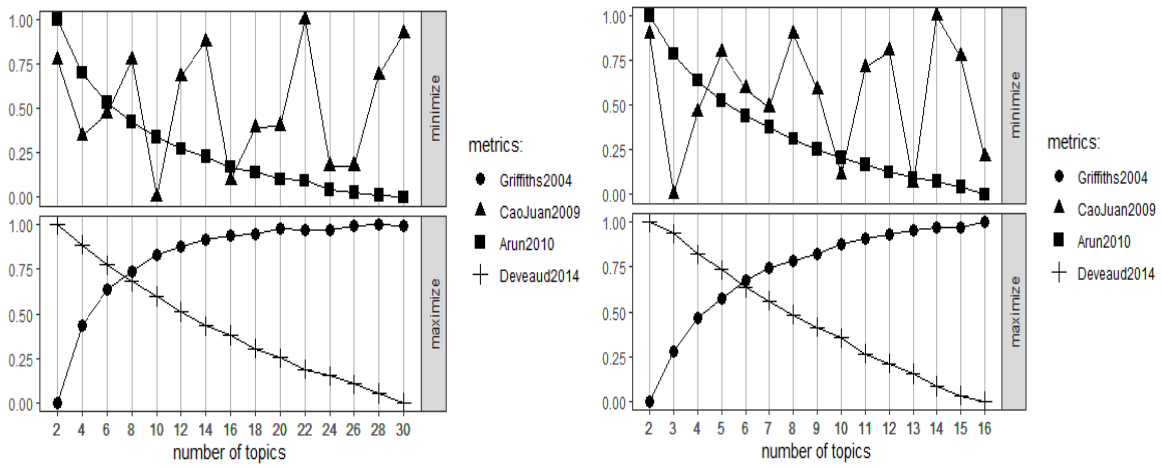
Topic Extraction and Definition

The result of model fitness estimation of the number of topics is shown in Figure 1. For the primary estimation, the number of topics input was 2 to 30 with an interval of 2 (left graph). Because the four scores indicate different directions, deciding on the appropriate number of topics is quite complicated. The CaoJuan2009 score fluctuates whereas Arun2010 decreases with the number of topics. The lowest converging scores of the two is 16. As for the Griffiths2004 and Deveaud2014, they go in totally opposite directions. The two converge somewhere around 7 or 8, but afterwards, the Deveaud 2014 becomes too low. Griffiths2004 continually increases but turns quite flat around 16. Put together, the appropriate number of k seems to be between 2 and 16. The second estimation was implemented with a narrower range of topic numbers, starting from 2 to 16 with an interval of 1. The result is the right-hand side graph of Figure 1. CaoJuan2009 and Arun2010 converge at 10 and 13. However, the scores for Deveaud2014 at 10 and 13 are too low (below 0.5). The next lowest converging point of CaoJuan2009 and Arun2010 is 7 and the scores for Griffiths2004 and Deveaud2014 are both reasonably high at k=7. Thus, the number of topics, k, was set as 7.

[10] For Griffiths2004 and Deveaud2014, the number of topics with a larger score has better fitness whereas a smaller score implies a more optimal LDA model (Baek, 2019).

Figure 1

Line Graphs Showing Four Model Fitness Scores of k Topics (First and Second Estimation)



Note. The graph on the left depicts first estimation.

The 7 topics extracted via Latent Dirichlet Allocation model are summarized in Table 1. Listed in Table 1 are the 20 keywords of each topic with the highest beta value[11] in the order from most important to least important. Different topics might share the same words since LDA is an admixture model (Sutherland et al., 2020). In order to clarify the theme of each topic, it is necessary to observe not only the keywords but also the headlines of articles most representative within the topic. Deciding the representativeness of an article within the topic can rely on the Gamma value generated from the LDA model, as is shown in Table 2.

Topic 1 includes core keywords such as elder abuse, elderly living alone, and crime towards the elderly. However, keywords show a large variation by including words such as blood donation, voters, candidates, etc. Such confusion can easily be solved by observing newspaper headlines. By taking the newspaper headlines into account, we can conclude that Topic 1 is closely related to various social issues that can arise in an aging society. Consequently, the theme for Topic 1 was determined to be *miscellaneous social issues* of an aging society.

[11] The beta value implies highest relative probability of belonging to the given topic (Sutherland et al., 2020).

Table 1

Topic Solution

Miscellaneous Social Issues		Death		Private Insurance Market		Economic growth		National Debt		Labor Market Innovation		Income Security	
Terms	*Beta*	*Terms*	*Beta*	*Terms*	*Beta*	*Terms*	*Beta*	*Terms*	*beta*	*Terms*	*Beta*	*Terms*	*Beta*
Elder abuse	0.127	Death	0.0204	Drivers	0.0153	GDP	0.0187	Healthcare cost	0.0266	Retirement age extension	0.0226	National pension system	0.0454
Elderly living alone	0.0105	Old people	0.159	Service	0.0102	Medical cost	0.0180	GDP	0.0167	Salary peak system	0.0198	Insurance contribution	0.0282
Service	0.0100	Dementia	0.0105	Insurance fee	0.0065	Growth rate	0.0123	Welfare expenditure	0.0109	Potential growth rate	0.0138	Recipient	0.0173
Blood donation	0.0082	Pneumonia	0.0079	Pension insurance	0.0054	Fiscal	0.0111	Disabled	0.0103	Employed	0.0134	Worker	0.0127
Crime	0.0054	Service	0.0063	Experts	0.0061	Potential growth rate	0.0100	Service	0.0092	Irregular worker	0.0121	Contributor	0.0127
Old woman	0.0053	Death rate	0.0058	Contributor	0.0057	Baby boomers	0.0089	Welfare policy	0.0089	Workers	0.0121	Reserve fund	0.0105
Dying alone	0.0051	Seniors	0.0056	Life insurance	0.0055	Realty asset	0.0086	Global	0.0074	Potential growth	0.0118	Retirement pension	0.0104
Pets	0.0049	Offender	0.0048	Cars	0.0047	Yearly average	0.0082	Fiscal burden	0.0070	Manufacturing industry	0.0105	Old age preparation	0.0087
Workers	0.0048	People	0.0048	Accidents	0.0047	Possibility	0.0080	Insurance contribution	0.0063	Productivity	0.0070	Retirement age extension	0.0080
Caregivers	0.0047	Cancer	0.0043	Insurance companies	0.0037	Workforce	0.0079	National debt	0.0062	Possibility	0.0063	Private pension	0.0072
Spending	0.0046	Separate families	0.0034	Program	0.0037	Minus	0.0073	Social	0.0050	Members	0.0051	Basic pension	0.0063
Textbook	0.0045	Elderly living alone	0.0034	Pension	0.0032	Household savings	0.0061	Budget deficit	0.0046	Regular workers	0.0049	Government	0.0063
Foreigner	00044	Difficulties	0.0031	Trend	0.0031	Self-employed	0.0059	Possibility	0.0045	Member country	0.0046	Salary peak system	0.0062
Social	0.0044	Activate	0.0030	National pension	0.0031	IMF	0.0053	Yearly average	0.0043	CEO	0.0045	Elderly age criterion	0.0060
Voters	0.0042	Crime	0.0028	Planner	0.0031	Working age population	0.0051	Insurance fee	0.0042	Yearly average	0.0045	Activate	0.0059
People	0.0041	Yearly average	0.0027	Small enterprise	0.0029	Economic growth rate	0.0048	Medical treatment	0.0040	Workforce	0.0043	Experts	0.0058
Program	0.0038	Suicide rate	0.0027	Realty asset	0.0028	Labor market participation	0.0047	Seriously ill	0.0040	Small business	0.0040	Elderly poverty	0.0052
Candidate	0.0038	Healthcare cost	0.0027	People	0.0028	Increase rate	0.0045	Insurance company	0.0038	Labor cost	0.0037	Income replacement rate	0.0051
Migrants	0.0038	Death cause	0.0027	Risk	0.0027	Labor market	0.0045	Cars	0.0036	Chairperson	0.0033	Specific	0.0050
Possibility	0.0037	Youth	0.0027	Companies	0.0026	Total population	0.0044	Cancer	0.0035	Public officer	0.0032	Service	0.0048

Note. The terms listed are 20 highest-ranked (beta score) terms per topic.

Topic 2 holds keywords associated with death or illness (death, dementia, pneumonia, cancer, etc.). By synthesizing the extracted keywords and newspaper titles belonging to Topic 2 we can note that Topic 2 is about death related issues or causes of death in an aging society. Some of the keywords seem to be protruding and bizarre (e.g., separate families[12]), but newspaper headlines indicate that these words are also related to death. Thus, the second topic could be labeled as issues relating to *death* of aging. The main keywords for Topic 3 are drivers, service, insurance fee, etc.

Articles belonging to Topic 3 are mostly related to private insurance market issues as shown in Table 2. Therefore, *private insurance market* could be an appropriate naming for Topic 3. The underlying theme for Topic 4 is rather obvious since many keywords point to the economic side of an aging society. The newspaper headlines also confirm that Topic 4 is related to *economic growth* in an aging society. Healthcare costs, welfare expenditure, GDP, etc. are keywords allocated to Topic 5; thus, Topic 5 can be labeled as *national debt* or burdens of an aging society. Topic 6 contains words related to labor market issues. *Labor market innovation* has been regarded as one of the key elements in tackling an aging society, as is shown by the newspaper headlines for Topic 6. Last, many keywords in Topic 7 are specifically about the national pension scheme. However, a wider range of words could be found, such as private pension, retirement pension, and retirement age extension, indicating that Topic 7 could be labeled *income security* in an aging society, a broader term than national pension.

[12] Many readers from outside Korea might be unfamiliar with the phrase separate families. Separate families mean people from North Korea who have been separated from their families and relatives in the course of the Korean War (1950-1953). The South Korean government has been trying to negotiate with the North Korean government to make opportunities for those families to meet. Nonetheless, there are still many people who haven't had the opportunity since the event largely depends on both countries' relationship.

Table 2

Newspaper Headlines of Each Topic

Topic	Gamma	Headline
1	0.999	Dark side of Aging Society: 90% of elder abuse occurs within family
	0.999	Calls for immediate policy measures: Rapid aging increases the number of older people dying alone
	0.998	Social abhorrence for older people
	0.975	Crimes against the elderly increase with rapid aging
	0.995	Aging increases need for blood donation but decreases the number of doners
2	0.999	Cancer and pneumonia records highest death cause due to aging
	0.989	Dementia enters top 10 cause of death for the first time due to aging
	0.977	Due to aging, last year's death rate was a record high
	0.975	Dementia, a national challenge in an aging society: Early checkup and prevention is most important
	0.955	Separate families (North and South) are also aging: 4900 died last year
3	0.998	The financial world is innovating: Insurance companies develop more products on later life planning
	0.990	Insurance planner becomes a popular job in aging society
	0993	The rise of sickness insurance and collapse of whole life insurance with aging
	0.987	Aging transforms whole life insurance schemes: Death benefit remodeled into pension during later life
	0.985	Need for auto insurance increases along with increase in car accidents of older people
4	0.995	The growth rate in 2041 predicted to drop to 0.6-0.8% due to serious aging
	0.993	Aging tsunami: Working age population will drop to 19% within the next two decades
	0.986	In 2026, household savings rate will become negative due to aging
	0.974	Health insurance benefit taken by older people increases to 40%
	0.952	Realty asset value declines at the fastest speed in an aging society

Table 2

Newspaper Headlines of Each Topic (Contd.)

Topic	Gamma	Headline
5	0.999	Aging snowballs government deficit: Will increase to 102% of GDP by 2050
	0.998	Due to aging, municipal governments will face budget deficits starting from next year
	0.995	Are medical costs of elderly people a time bomb in aging society?
	0.995	Average yearly increase in government budget on social welfare is 8%
	0.993	Rapid aging quintuples elderly welfare budget in 3 years
6	0.999	67% agree on retirement age extension
	0.999	[Editorial] Work for older people is the answer
	0.999	Modifying wage system is the key to achieving employment stability in an era of aging
	0.998	Low growth and financial deficit caused by aging should be tackled by labor market innovation
	0.995	57% of businesses in manufacturing industries consider extending employment of aged workers
7	0.999	[Tackling Aging Society] The trilemma of Korean National Pension: sustainability-coverage-trust
	0.996	Penniless, thus miserable later life
	0.983	Is basic pension effective prescription in an aging society?
	0.964	Aging faster than Japan: The need for adjusting pension replacement rate
	0.952	National Pension Fund will be exhausted by 2057, 3 years earlier than expected, due to aging

In the next section, the result of a time-series analysis of how these topics change by year will be explained, with the purpose of suggesting policy implications.

Topic Dynamics over Time

Table 3 shows how the topics and percentage of documents referencing these array of topics changes over time. Notably, most topic loadings increase over time. These changes could be interpreted differently based on different perspectives. First,

because the digital corpus has become larger over time as more and more documents get digitized, indicating that quantity is not a suitable means for illustrating change. From a different perspective, increase in the quantity might reflect increased media interest about aging. In any case, we should take caution in imposing too much meaning on quantity. However, Table 3 shows that, although topic loadings generally increase with time, they fluctuate from time to time. It is noticeable that topic loadings increase during the last year of a government basic plan implementation cycle, the years 2010 and 2015, probably because it is also a period of introducing and publicizing the new five-year basic plan.

Table 3

Topic Loading Changes by Year

	Miscella-neous Social Issues	*Death*	*Private Insurance*	*Economic Growth*	*National Debt*	*Labor Market Innovation*	*Income Security*	*Total*
2006	14.60	21.81	8.77	5.53	18.32	11.89	20.09	101.01
2007	3.37	10.96	9.37	8.31	15.22	9.66	3.10	59.99
2008	7.64	5.07	2.43	3.24	8.73	4.72	6.19	38.02
2009	3.51	2.18	4.85	3.30	4.44	6.20	10.53	35.01
2010	23.17	11.02	13.18	13.12	15.37	12.53	11.60	99.99
2011	8.11	11.09	9.51	10.63	16.61	21.84	16.20	93.99
2012	11.66	11.46	18.17	11.99	24.57	11.05	12.09	100.99
2013	9.03	11.45	5.63	14.26	22.64	11.66	20.33	95.00
2014	8.61	20.58	10.07	13.78	20.85	16.05	17.04	106.98
2015	15.46	16.84	19.12	15.41	38.84	21.80	19.52	146.99
2016	15.52	14.79	22.64	19.99	28.73	12.14	10.18	123.99
2017	16.13	10.20	9.53	20.32	42.77	16.36	12.68	127.99
2018	17.47	11.01	29.10	22.37	29.84	24.96	12.26	147.01
2019	13.36	23.98	16.75	28.74	46.59	46.54	18.04	194.00

Note. LDA calculates gamma values (examples shown in Table 2 are documents with prominent gamma value for a certain topic) for each document, showing the probability of finding the word-sets for each topic in a document. Each document has gamma values for all 7 topics that add up to 1. The topic loadings shown could be translated as the sum of gamma values (of each document) per topic by year, and the "total" indicates the total number of documents per year (however, because of rounding, it doesn't always exactly add up to a whole number).

Focusing on each topic, the result shows *social issues*, *death*, and *private insurance* topics received relatively minor attention from the media compared to other issues. The *social issues* topic was a major issue covered by the media during the first basic plan era and mainly received public attention during the first and last year of basic plan implementation. The *private insurance* topic shows the tendency to increase recently, during the third basic plan era. The *economic growth* topic seems to be gaining more attention recently, but if we compare the relative portion of *economic growth* to all the topics, it is noticeable that the loading is quite stable over time. Topic loading for *national debt* and *labor market innovation* are the most prominent topics; not only have they increased over time, but they have been the most frequently discussed topics during the third basic plan era. It is surprising to note that the topic *income security* has not been of much interest, relatively speaking, in the media considering the high poverty rate of older people in Korea, suggesting a momentary media interest in the income security issue.

Discussion

This study focused on the use of probabilistic topic models of newspaper articles to find the public discourse structure on aging, how it has changed over time, and what implications it has for the future of the super-aging Korean society. The findings can be summarized as follows:

First, compared to 2006, the quantity of media discourse on aging has increased sharply within recent years, showing an increased public interest on aging. Even though Korean society became more aware of population aging shortly after the year 2000 the statistics show that public interest was only minor or superficial before 2010. After transient interest in 2006 when the first five-year national plan on aging was announced and implemented, media discourse on aging decreases noticeably. The year 2010 records a turning point, as the quantity of newspaper articles on aging increases consistently afterwards, indicating that aging has become a major public issue in Korea. However, within the overall upward trend exists five-year sub-trends. Media discussion on aging tends to increase noticeably during the final year of national basic

plan implementation (i.e., 2010 and 2015), a period when the next basic plan is prepared and publicized, implying the government policy impacts media discourse.

Second, seven latent topics were derived from the LDA model, which can be identified as *social issues*, *death*, *private insurance*, *economic growth*, *national debt*, *labor market innovation*, and *income security*. The topic loadings varied by year but demonstrated a clear increase in public interest on certain topics. Specifically, the topic loadings of the year 2006 show, by and large, a balanced distribution among topics while *economic growth* gained little interest from the media. However, longitudinal changes in topic loadings show a marked increase in *economic growth* (5.53 in 2006 to 28.74 in 2019), *national debt* (18.32 in 2006 to 46.59 in 2019), and *labor market innovation* (11.89 in 2006 to 46.54 in 2019).

It is possible to infer from such results that the aging discourse shaped by the media has become more inclined to address efficiency and productivity issues. Regarding the fact that Korean society is facing some highly controversial issues, such as high poverty and the suicide rate of older people, productivity-focused discourse on aging might result in negative social effects. As Kim (2017) asserts, productivity focused discourse on aging forces older people to be productive subjects. They are recognized by society as economic participants obliged to fulfill a productive role within society. Such subjectivation connotes a decreased role of the government and society by shifting the responsibility to individuals. In this respect, a lack of material as well as non-material resources that older people in Korea are faced with is the result of maladjustment within the labor market and thus the responsibility of individuals. As media discourse on aging is shaped towards recognizing older people as independent and productive citizens focusing on providing opportunities in the labor market, social issues and problems can be marginalized. In reality, older people who successfully adapt to the changing environment and become independent subjects within the market are only a minority, while the majority of older people are in need of societal care and support. Furthermore, public expectations for independent and productive older people shaped by aging discourse can lead to an aversion towards the majority of people, who fail to satisfy this standard. Gerontophobia isn't a serious issue in Korea yet but is always possible, considering the highly competitive and harsh social

environment younger generations are facing, an environment that disrupts traditional values such as respect for elders.

One final issue to address concerns the methodological limitations of this study. This study focused on using big data analysis, specifically topic modeling via Latent Dirichlet Allocation. Although LDA is a widely used topic modeling technique, it suffers from "order effects," which occur when the order of data is shuffled, generating different topics. This limitation can be quite controversial since it can cause systematic errors and lead to misleading results such as inaccurate topic description (Agrawal et al., 2018). Experts in computer science and information technology are working to find a method that generates more stable LDA results, but it still needs to be elaborated. Nevertheless, readers must take into account that big text data analysis as a method is only at the beginning stages and more research is necessary to tackle such limitations.

References

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology, 98*, 74-88. https://doi.org/10.1016/j.infsof.2018.02.005

Angermuller, J. (2015). Discourse studies. In J. Wright, (Ed.), *Encyclopedia of the social and behavioral sciences* (2nd ed., pp.510-515). Elsevier.

Baek, Y. M. (2019). *Text-mining using R.* Hanul Academy.

Big KINDS. (2017). Korea Press Foundation. www.bigkinds.or.kr

Blei, D. M., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In A. Pohoreckhy, L. Bottou, & M. L. Littman (Eds.), *Proceedings of the International Conference on Machine Learning*. 113-120. http://doi.org/10.1145.1143844.1143859

Diamond, I., & Quinby, L. (1988). *Feminism and Foucault: Reflections on resistance*. Northeastern University Press.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), 570-606. https://doi.org/10.1016/j.poetic.2013.08.004

Han, G., & Yoon, S.(2007). Daejoong maechaeaeseo shinnonyeon damlon boonseok [Critical review of discourse on aging in Korean newspapers]. *Journal of the Korean Gerontological Society, 27*(2), 299-322. UCI: G704-000573.2007.27.2.002

Jang, M. J. (2013). Media damoonhwa damlon boonseok [Multicultural discourse analysis in media]. *Multicultural Education Studies, 6*(3), 157-179. http://dx.doi.org/10.14328/MES.2013.09.30.161

Kim, E. J. (2017). Mediaga pyobanghaneun goryeonghwa sahoieui balamjikhan noinsang [How media makes the elderly into welcoming citizens in the aged society]. *Korean Journal of Journalism and Communications, 61*(3), 157-188. http://doi.org/10.20879/kjjcs.2017.61.3.005

Lee, B. H., & Kim, G. Y. (2019). Social big dataro bon noinguanlyun damloneui teukjinggua byunhwa [The themes and trends of discourse on the elderly in Korea identified by analyzing social media big data]. *Social Welfare Policy, 46*(3), 179-201. https://doi.org/10.15855/swp.2019.46.3.179

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1608). http://doi.org/10.1186/s40064-016-3252-8

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. (2013). Differentiating language usage through topic models*. Poetics, 41*(6), 607-625. https://doi.org/10.1016/j.poetic.2013.06.004

Park, H. J., Kim, H., & Hong, Y. J. (2017). Topic modellingeul hwalyonghan haksaeng inkwonjoryeeui sahoejeok issue boonseok [A topic modeling analysis on the major social issues of the Students' Human Rights Ordinance in Korea]. *Asian Journal of Education, 18*(4), 683-711.

Roh, B. R. & Yang, K. E. (2019). Hangook sahoe jeochoolsan noneui gujowa ke byonhwane gwanhan text mining boonseok [Text mining analysis of South Korea's birth-rate decline issue in newspaper articles: Transition patterns

over 18 years]. *Korean Journal of Social Welfare*, 71(4), 154-176. https://doi.org/10.20970/kasw.2019.71.4.006

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly.

Sutherland, I. & Kiatkawsin, K. (2020). Determinants of guest experience in Airbnb: A topic modeling approach using LDA. *Sustainability*, *12*(8), 3402. https://doi.org/10.3390/su12083402

Statistics Korea. (n.d.) http://kosis.kr

Biographical Note

So Chung Lee is an assistant professor at Namseoul University in Korea. Previously, she was a researcher at Korea Institute of Health and Social Affairs, a government-funded research institute. Her main research area is aging, poverty, and welfare policy in general. She currently participates in various policy making organizations such as the Presidential Committee on Aging Society and Population Policy.

She can be reached at Namseoul University 91 Daehak-ro, Seobuk-gu, Cehonan-si, Chungcheongnam-do, South Korea or by email at snowvill@nsu.ac.kr.

Date of Submission: 2021-05-29

Date of Acceptance: 2022-01-14