Osong Public Health and Research Perspectives

Journal homepage: <http://www.kcdcp.hr.org>

Original Article

Analysis of Women's Health Online News Articles Using Topic Modeling

Kyoung Won Cho ^a, Shine Young Kim ^b, Young Woon Woo ^{c,*}

^a Department of Healthcare Administration, Kosin University, Busan, Korea

^b Laboratory Medicine, Misoan Clinic, Busan, Korea

^c Division of Creative Software Engineering, Dong-eui University, Busan, Korea

ABSTRACT

Article history:

Received: April 9, 2019

Revised: May 11, 2019

Accepted: May 13, 2019

Keywords:

data mining,
newspaper article,
women's health

Objectives: This research aimed to understand the popularity of topics in the field of women's health through analysis of online news articles which were chronologically classified and examined to determine how women's health and diseases had changed over time.

Methods: Women's health and disease news articles were collated from a popular news website between 1993 to 2015 and preprocessed using gynecological medical terminology, Korean words and nouns (excluding general nouns not related to women's healthcare topics). The resultant articles (N = 7,710) were analyzed using the Latent Dirichlet Allocation algorithm and major topics were extracted. Topic trends were analyzed by year and period for women's health.

Results: It was observed that most of the women's health articles were focused on "Healthcare", and 9 other topics were identified that represented a relatively small proportion in 1993-2000. In 2001-2005, most of the articles were focused on "Medical Services" and "Dietary Supplements" with some specific topics that peaked people's interest, as compared to those focused on "Healthcare" in the 1990s. It was also observed that differences in the proportion of each topic was small after 2011.

Conclusion: Changes in topics related to women's disease were not clearly distinguished in the 1990s but this changed from 2001 where articles related to "women disease" appeared as articles on the topics of various diseases.

<https://doi.org/10.24171/j.phrp.2019.10.3.07>
pISSN 2210-9099 eISSN 2233-6052

©2019 Korea Centers for Disease Control and Prevention. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Introduction

In Korea, women's health has been discussed for a long time regarding on population control, enhancement of pregnancy and childbirth, and qualitative development of populations only (education, crime, nutrition, race, social class, wealth, wellbeing). For this reason, matters involving women's health have focused primarily on pregnancy and childbirth of fertile women, and other related issues have been under the radar of public healthcare services and the medical field [1]. However, women have other health problems, such as cancers specific to

women and infection caused by sexually transmitted diseases. Changes in lifestyle and environment have a greater negative effect on women's health than men's health. Demographics and physical characteristics, education, economics, labor, culture, social environments, and women's role in the family all factor [2].

There have been studies on health-related articles that analyzed healthcare problems, but most studies focused on only 1 health-related issue. For example, in 2008 Jung analyzed the news framing journalism perspectives of acquired immunodeficiency syndrome and human immunodeficiency virus in the 1990s [3]. The concept of framing is related to the

*Corresponding author: Young Woon Woo
Division of Creative Software Engineering, Dong-eui University, 176 Eomgwang-ro, Busanjin-gu, Busan, 47340 Korea
E-mail: ywoo@deu.ac.kr

agenda-setting tradition but expands the research by focusing on the essence of the issues at hand rather than on a particular topic. The basis of framing theory is that the media focuses attention on certain events and then places them within a field of meaning. In 2001, Andsager reported on making sense of breast cancer and breast implants [4]. In 2013, depression and mental health coverage in the media were analyzed [5] and factors involved in the stigma of suicide prevention were studied [6].

There have been analysis of women's overall health in the media (pregnancy, childbirth, and infertility) by the National Survey on Fertility, Family Health and Welfare which publish periodically. In 2014, Kim and Kim [7] analyzed infertility-related reports from 1962 to 2013, and consequently this may have led to diverse studies on women's health policies have been conducted.

However, these studies focused on individual health issues influencing women's health. Therefore, integral health issues that specifically relate to women were difficult to identify. Thus, this study collected and analyzed articles on women's health in general and assessed how the media reported on major health issues in Korea to determine importance (objective methodology was used).

A web scraping method has recently been developed and data mining is actively used to collect and analyze large-scale web documents automatically [8]. This is enabled by many open-source libraries available for data mining where it is possible to analyze big data with only a small amount of coding necessary.

Text mining is not only a good method for identifying the structure of a text and extracting concepts, but also useful for visualization [9]. Text mining is being employed in the analysis of trends of journals [10], social network services such as Twitter and blogs [11,12], customer online reviews [13,14], and the discourse of big data in news outlets [15].

Topic modeling is a type of big data analysis methodology for discovering abstract topics that repeatedly occur in a collection of documents. When an author has a specific keyword in mind, this keyword is repeated throughout the article. The collection of keywords is modeled as a finite mixture over an underlying set of topic probabilities, and then provides a latent topic in a specific document [16].

Research using topic modeling of news articles is becoming active in diverse fields. In 2007 Falinouss reported that stock market prediction's could be made using data mining techniques [17]. Textual documents and time series can be mined concurrently to predict the movements of stock prices based on the contents of news articles. The relationship between the contents of the news stories and trends on stock prices are learned through the Support Vector Machine. The accuracy of the prediction model is 83%, which means the

model has increased its accuracy by 30% [17].

In 2013, DiMaggio used LDA to analyze how 1 policy domain (government assistance to artists and arts organizations) was framed in almost 8,000 articles. The authors illustrated the strengths of topic modeling to analyze large text corpora, discussed the correct choice of models and interpretation of model results, described the means of validating topic-model solutions, and demonstrated the use of topic models in combination with other statistical tools to estimate differences between newspapers in the prevalence of different frames [18].

Studies involved in the analysis of news articles on women's health frequently focus on 1 health issues that influence women, which means that there are limitations in understanding the general topics of women's health-related social issues. Therefore, in the present study, we have systematically collected and analyzed articles that discuss women's health-related topics. In order to maintain objectivity and draw meaningful results, topic modeling methodology was introduced to perform a chronological examination of how social issues relate to women's health and diseases, and have changed with changes in the social environment. The outcome of the present study may be used as basic data for effective introduction of women's health policies and financial management.

Materials and Methods

1. Overall process

Figure 1 shows the overall process for the analysis of women's health-related online news articles. In Step 1, the online news articles were collected from 1 selected news website (Table 1), and the collected news articles were saved as text files in comma separated values file format. In Step 2, the saved data were preprocessed (First, special terms related to gynecology were appended to dictionary using gynecological medical terminology, in order to extract special terms as well as general nouns. Second, general nouns not related to women's healthcare topics, were removed.) so that the analysis of the results in Step 3 could be accurately derived. In Step 3, the preprocessed data were analyzed using the LDA algorithm [19,20]. In Step 4, the major topics of health-related online news were extracted from the LDA analysis. In Step 5, topic trends by year and period were analyzed using the extracted topics, and interrelationships of the extracted topics were also analyzed.

2. Data collection

As shown in Table 1, Chosunilbo (<http://www.chosun.com/>) was ranked in the top 3 for online news websites by 3 website

Table 1. Representative online news sites by website rank sites.

Website rank site	Ranking #1	Ranking #2	Ranking #3
www.rankey.com	www.chosun.com	www.donga.com	www.mk.co.kr
www.alex.com	www.donga.com	www.chosun.com	joongang.joins.com
www.sinnungudok.com	joongang.joins.com	www.chosun.com	www.mk.co.kr

Figure 1. Overall process of topic trends analysis. LDA = latent Dirichlet allocation.

ranking sites, implying that the data gathered from this online news site would be reliable. This was the third most popular online newspaper read by Koreans in 2015.

To identify women's health-related topics, 3 search terms were used to collect online news articles in a test search; "women health", "women disease" and "women illness". Analysis revealed that the search terms "women disease" and "women illness" was not satisfactory because more advertising articles were retrieved than news articles. Therefore, we used "women health" to collect women's health-related news articles. Article titles and full-text were extracted after removing irrelevant text such as unique phrases and text for hyperlinking to other articles.

As shown in Table 2, the number of total collected articles was 7,710. The articles had been published over a 23-year period between 1993 to 2015, with 418 articles collected from 1993 to 2000, 643 from 2001 to 2005, 2,437 from 2006 to 2010,

Table 2. Number of news article by periods.

Period	Number of articles
1993-2000	418
2001-2005	643
2006-2010	2,437
2011-2015	4,212

and 4,212 from 2011 to 2015.

3. Data preprocess

The data were preprocessed in 4 steps. In Step 1, a basic dictionary for extracting Korean words was selected. In this study, the Hangul morpheme dictionary (NIADic) provided by the National Information Society Agency (<https://www.nia.or.kr>) for morphological analysis was used. This dictionary has more than 900,000 built-in words, which is the largest among domestic dictionaries and has an advantage that it can be used directly in the R, a computer programming language. In Step 2, a dictionary of gynecological medical terminology was used in order to increase the probability of extracting terminology related to women's health, because terminology such as various names for diseases related to women's health cannot be extracted with only a basic dictionary. In Step 3, only Korean nouns from article sentences were extracted (using the SimplePos22 morphological analysis function provided in the KoNLP library [21]) and stored for each article. In Step 4, general nouns (e.g., "most", "this year", "last year", "women", "illness", "case", "person", and "degree") were deleted that were not related to women's healthcare topics, and only the remaining words from the Step 4 were stored for each article to be part of the final data set.

4. LDA analysis, topic extraction, and topic trend analysis

In order to perform the LDA analysis, the number of topics must first be determined. In general, it is important to determine a reasonable number of topics in topic modeling

using LDA. In novels without a clear unit of source data, the decision what should be done as a document unit is one of the most important decisions [22]. However, in newspaper articles, each article is usually used as a document, and the level of difficulty involved in determining the number of topics depends mainly on the interpretability of the topics [23]. In the present study, the number of topics was extracted and analyzed from 5 to 30 to determine the appropriate number. As a result, the number of topics was set at 10, taking into account interpretability and meaningfulness. After determining the number of topics, the top 20 most frequent words were extracted for each topic using the collapsed Gibbs sampling technique for the LDA analysis algorithm [18,24]. In particular, in the present study, LDA analysis was performed for each period before 2000, 2001–2005, 2006–2010, and since 2011, and topics of each period were obtained. Through this analysis, (1) the representative topics of women's health-related issues by period and year and (2) how the topics have changed were identified.

Results

1. Period from 1993–2000

The period 1993–2000 had 10 representative topics on women's health were identified: “healthcare”, “health consultation”, “pregnancy and childbirth”, “AIDS”, “urinary health”, “mortality statistics”, “foot health”, “women's life”, “new technology”, and “mental health”.

Table 3 shows the results of topic modeling from 1993–2000. The first topic of “healthcare” contained the following words: “treatment”, “patient”, “abnormal”, “exercise”, “symptom”, “professor”, “effect”, “healthcare”, “breast cancer”, “cause”, “hospital”, “skin”, “body”, “heart”, “hormone” “cholesterol”, “alcohol”, “man”.

The second topic “mortality statistics” contained the following words: “disease”, “female”, “male”, “mean”, “death”, “last year”, “USA”, “cause of death”, “smoking”, “death” “lung cancer”, “life expectancy”, “population, man”, “mortality”, “tuberculosis”, “world”, “tobacco”, “developed country”, “respiratory”.

An examination of the proportions of the topics revealed that most of the articles were focused on the topic of “healthcare”. The proportions of articles focused on the other 9 topics were relatively small. Topic trends for 1993–2000 are shown in Figure 2.

2. Period from 2001–2005

The period 2001–2005 had 10 representative topics on women's health were identified: “cerebrovascular disease”, “arthropathy”, “skin health”, “medical service”, “kidney disease”, “dietary supplement”, “thyroid disease”, “pregnancy and childbirth”, “lifestyle disease and prevention”, “urinary health”.

Table 4 shows the results of topic modeling from 2001–2005. The first topic of “medical service” contained the following words: “patient”, “treatment”, “abnormal”, “surgery”, “professor”, “hospital”, “pain”, “medicine”, “symptom”,

Figure 2. Topic trends from 1993 to 2000.

Table 3. Analysis results of topic modeling from 1993 to 2000.

	Topics (1993–2000)	Words	Proportion
T1	Healthcare	treatment, patient, abnormal, exercise, symptom, professor, effect, healthcare, breast cancer, cause, hospital, skin, body, heart, hormone, cholesterol, alcohol, man, medicine, doctor	0.874
T2	Health consultation	stalking, work, information, afternoon, phone, college, Seoul, arrangement, operation, planning, morning, consultation, president, relation, workplace, professional, domestic, enlarge, door, house	0.014
T3	Pregnancy & childbirth	uterus, ovary, sperm, pregnancy, baby, infertility, hospital, fetus, procedure, domestic, obstetrics, secretion, success, marriage, end, development, inside, penis, advanced country, use	0.012
T4	AIDS	cesarean section, AIDS, maternity, infection, improvement, Korea, mean, abnormal, surgery, middle, virus, result, world, around, investigation, professor, announced, complication, childbirth, blood	0.016
T5	Urinary health	prostate, urinary incontinence, urine, urinary bladder, erectile dysfunction, urethra, male, penis, urology, sperm, complication, constipation, type, out, treatment, maternity, medical, deficiency, self, chronic	0.010
T6	Mortality statistics	disease, female, male, mean, death, last year, USA, cause of death, smoking, death, lung cancer, life expectancy, population, man, mortality, tuberculosis, world, tobacco, developed country, respiratory	0.039
T7	Foot health	foot, toe, athlete's foot, soles, shape, healthy, fatal, bone, possible, sweat, symptom, method, abnormal, treatment, patient, exercise, symptom, professor, male, USA	0.005
T8	Women's life	wine, book, author, novel, love, story, Human, man, self, society, people, life, woman, home, real estate, Japan, world, introduction, answer, idea	0.019
T9	New technology	gene, vaccine, robot, human, computer, universe, outlook, development, technology, progress, era, step, use, discovery, material, expectation, plan, treatment, success, AIDS	0.001
T10	Mental health	amnesia, grandmother, husband, memory, England, lady, children, stress, serious, thought, anxious, housewife, death, activity, night, USA, talking, heart disease, maternity, outcome	0.010

“problem”, “doctor”, “USA”, “test”, “cause”, “health”, “result”, “self”, “director”, “depression”, “disease”.

The second topic of “dietary supplement” contained the following words: “product”, “ingredient”, “vitamin”, “water”, “taste”, “effect”, “food”, “protein”, “market”, “last year”, “popularity”, “prevention”, “containing”, “advertising”, “diverse”, “function”, “nutrition”, “action”, “skin”, “use”.

An examination of the proportions of the topics revealed that most of the articles were focused on the topics of “medical service” and “dietary supplement”, and some specific topics were characterized as being of high interests (e.g., “healthcare” in the 1990s). Topic trends for 2001–2005 are shown in Figure 3.

3. Period from 2006–2010

The period 2006–2010 had 10 representative topics on women's health were identified: “climacterium health”, “pregnancy and childbirth”, “women disease”, “beauty treatment”, “medical service”, “skin health”, “lifestyle disease”, “hair loss”, “joint disease”, “skin care”.

Table 5 shows the results of topic modeling from 2006–2010. The first topic of “skin health” contained the following words: “skin”, “water”, “atopy”, “cosmetic”, “odor”, “germs”, “sweat”, “foot”, “product”, “eye”, “temperature”, “keratin”, “cold”, “hand”, “clean”, “use”, “allergy”, “using”, “moisture”, “summer”.

The second topic of “hair loss” contained the following words: “hair loss”, “scalp”, “hair”, “hairs”, “head”, “stress”, “treatment”, “male hormone”, “male”, “blood circulation”, “site”, “hormone”, “gene”, “product”, “nutrition”, “progress”, “massage”, “effect”, “facilitation”, “genetic”.

The third topic of “medical service” contained the following words: “patient”, “game”, “USA”, “surgery”, “voice”, “child”, “Korea”, “professor”, “domestic”, “goods”, “hospital”, “husband”, “treatment”, “self”, “Seoul”, “world”, “service”, “wife”, “thoughts”, “children”.

The fourth topic of “beauty treatment” contained the following words: “vitamin”, “food”, “exercise”, “body”, “intake”, “fat”, “diet”, “water”, “calcium”, “taste”, “ingredient”, “nutrient”, “grocery”, “effect”, “protein”, “milk”, “help”, “health”, “stress”,

Table 4. Analysis results of topic modeling from 2001 to 2005.

	Topics (2001-2005)	Words	Proportion
T1	Cerebrovascular disease	dementia, Alzheimer, memory, brain, research, stroke, personality, hyperlipidemia, name, medical school, generation, field, Aspirin, protein, obesity, prevention, like, face, plan, body	0.017
T2	Arthropathy	joint, arthritis, bone, degenerative, cartilage, exercise, knee, shoulder, finger, side effect, body, inflammation, shock, posture, start, morning, effective, tissue, medication, lifelong	0.006
T3	Skin health	skin, atopy, dermatitis, hair loss, ultraviolet, bath, scalp, hair, allergy, water, wrinkle, head, procedure, face, moisture, site, effect, removal, product, help	0.023
T4	Medical service	Patient, treatment, abnormal, surgery, professor, Hospital, Pain, Medicine, Symptom, Problem, Doctor, USA, Test, cause, health, result, self, director, depression, disease	0.526
T5	Kidney disease	renal failure, glomerulus, dialysis, urine, acute, statistics, University hospital, creation, use, progress, prostate, contents, patient, treatment, abnormal, professor, effect, symptom, medicine, exercise	0.002
T6	Dietary supplement	product, ingredient, vitamin, water, taste, effect, food, protein, market, last year, popularity, prevention, containing, advertising, diverse, function, nutrition, action, skin, use	0.352
T7	Thyroid disease	thyroid, function, iodine, thyroid cancer, calcium, concentration, medication, therapy, rest, diagnosis, sweat, regular, containing, fetus, statistics, ingredient, alcohol, wind, patient, treatment	0.004
T8	Pregnancy & childbirth	pregnancy, uterus, infertility, contraception, leukemia, maternity, fetus, sperm, gynecology, childbirth, baby, menstruation, discharge, bleeding, marriage, menstruation, intercourse, experience, child, acute	0.018
T9	Lifestyle disease & prevention	abnormal, exercise, patient, hypertension, body, health, treatment, symptom, hormone, effect, diabetes, medicine, cause, food, professor, stress, blood pressure, risk, need, problem	0.032
T10	Urinary health	sexual behavior, male, prostate cancer, erectile dysfunction, prostate, sexual, erection, penis, intercourse, disability, urine, male hormone, urology, sexual function, sexually transmitted disease, sexual, male, psychological, mental, factor	0.020

Figure 3. Topic trends from 2001 to 2005.

Table 5. Analysis results of topic modeling from 2006 to 2010.

	Topics (2006-2010)	Words	Proportion
T1	Climacterium health	depression, male, patient, climacterium, hormone, symptom, erectile dysfunction, sexual behavior, urine, incontinence, suicide, male hormone, sex, treatment, medicine, husband, stress, urology, menopause, problem	0.022
T2	Pregnancy & childbirth	pregnancy, abortion, maternal, maternity, fetus, baby, obstetrics, child, infertility, marriage, couple, uterus, doctor, procedure, problem, society, mother, government, fact, possibility	0.014
T3	Women disease	uterus, surgery, thyroid, uterine myoma, pain, Menstrual cramps, headache, anus, thyroid cancer, breast cancer, menstruation, tooth, treatment, gum, pregnancy, migraine, breast, exam, obstetrics, patient	0.039
T4	Beauty treatment	vitamin, food, exercise, body, intake, fat, diet, water, calcium, taste, ingredient, nutrient, grocery, effect, protein, milk, help, health, stress, fruit	0.113
T5	Medical service	patient, game, USA, surgery, voice, child, Korea, professor, domestic, goods, hospital, husband, treatment, self, Seoul, world, service, wife, thoughts, children	0.127
T6	Skin health	skin, water, atopy, cosmetic, odor, germ, sweat, foot, product, eye, temperature, keratin, cold, hand, clean, use, allergy, using, moisture, summer	0.198
T7	Lifestyle disease	test, study, patient, risk, hypertension, result, diabetes, cholesterol, research team, hepatitis, male, brother, heart disease, death, swine flu, USA, stroke, abnormal, professor, colon cancer	0.110
T8	Hair loss	hair loss, scalp, hair, hairs, head, stress, treatment, male hormone, male, blood circulation, site, hormone, gene, product, nutrition, progress, massage, effect, facilitation, genetic	0.184
T9	Joint disease	joint, pain, knee, arthritis, back, spine, leg, surgery, muscle, posture, shoulder, lower limb varicose veins, exercise, foot, bone, symptom, body, neck, veins, pelvis	0.088
T10	Skin care	skin, acne, stain, ultraviolet, laser, procedure, pigment, face, wrinkle, pore, scar, effect, elasticity, product, cosmetics, hair, treatment, eye, keratin, aging	0.104

Figure 4. Topic trends from 2006 to 2010.

Table 6. Analysis results of topic modeling from 2011 to 2015.

	Topics (2011-2015)	Words	Proportion
T1	Women's cancers	patient, professor, surgery, examination, cervical cancer, breast cancer, vaccine, screening, health screening, research, gene, colon cancer, abnormal, domestic, USA, breast, mortality, mean, disease, suicide	0.036
T2	Skin health	skin, vaginitis, germ, eye, water, cold, smell, allergy, atopy, use, virus, sweat, tooth, summer, summer season, symptom, foot, product, temperature, infection	0.100
T3	Gynecology	uterus, pregnancy, uterine leiomyoma, urinary incontinence, urine, bladder, childbirth, gynecology, cystitis, surgery, menstrual cramps, menstruation, ovary, examination, fetus, infertility, test, cervical cancer, marriage, symptom	0.095
T4	Medical service	husband, government, child, field, house, children, couple, parent, USA, world, self, service, family, wife, society, people, professor, thought, program, mind	0.075
T5	Lifestyle disease	diabetes, osteoporosis, hypertension, exercise, risk, cholesterol, obesity, blood vessels, blood pressure, metabolic syndrome, abnormal, stroke, gout, bone, study, fat, intake, smoking, weight, fracture	0.130
T6	Dietary supplement	vitamin, food, intake, calcium, grocery, ingredient, protein, nutrient, milk, fruit, taste, health, water, vegetable, help, efficacy, garlic, cholesterol, diet, containing	0.132
T7	Joint disease	Joint, pain, knee, back, arthritis, foot, spine, leg, exercise, muscle, shoulder, bone, cartilage, posture, disk, shoe, surgery, ligament, varicose vein, high heel	0.119
T8	Infectious disease	MERSC, patient, infection, virus, hospital, death, large, domestic, patient, confirmation, work, occurrence, suspicion, male, case, possibility, symptom, afternoon	0.105
T9	Skin care	skin, hair loss, procedure, laser, scalp, acne, wrinkle, stain, ultraviolet, hair, face, cosmetic, effect, product, ingredient, moisture, elasticity, area, scar, treatment	0.087
T10	Climacterium	hormone, climacterium, voice, symptom, depression, surgery, erectile dysfunction, chest, headache, stress, pain, treatment, thyroid, female hormone, menopause, abnormal, brain, dementia, syndrome, disorder	0.122

“fruit”.

The fifth topic of “lifestyle disease” contained the following words: “test”, “study”, “patient”, “risk”, “hypertension”, “result”, “diabetes”, “cholesterol”, “research team”, “hepatitis”, “male”, “brother”, “heart disease”, “death”, “swine flu”, “USA”, “stroke”, “abnormal”, “professor”, “colon cancer”.

An examination of the proportions of the topics revealed that the proportion of articles on the topic of “skin health” had been steadily decreasing from 2006-2010. This decrease was also observed in several other topics, such as “beauty treatment”, “hair loss”, and “skin care”, as various topics in the field of “skin health” were subdivided. In addition, articles on the topic of “joint disease” continued to increase for 5 years. This increase reflects social issues with regard to the aging population. Topic trends for 2006-2010 are shown in Figure 4.

4. Period from 2011-2015

The period 2011-2015 had 10 representative topics on women's health were identified: “women's cancers”, “skin health”, “gynecology”, “medical service”, “lifestyle disease”, “dietary supplement”, “joint disease”, “infectious disease”, “skin care”, “climacterium”.

Table 6 shows the results of topic modeling from 2011-2015.

The first topic of “dietary supplement” contained the following words: “vitamin”, “food”, “intake”, “calcium”, “grocery”, “ingredient”, “protein”, “nutrient”, “milk”, “fruit”, “taste”, “health”, “water”, “vegetable”, “help”, “efficacy”, “garlic”, “cholesterol”, “diet”, “containing”.

The second topic of “lifestyle disease” contained the following words: “diabetes”, “osteoporosis”, “hypertension”, “exercise”, “risk”, “cholesterol”, “obesity”, “blood vessels”, “blood pressure”, “metabolic syndrome”, “abnormal”, “stroke”, “gout”, “bone”, “study”, “fat”, “intake”, “smoking”, “weight”, and “fracture”.

The third topic of “climacterium” contained the following words: “hormone”, “climacterium”, “voice”, “symptom”, “depression”, “surgery”, “erectile dysfunction”, “chest”, “headache”, “stress”, “pain”, “treatment”, “thyroid”, “female hormone”, “menopause”, “abnormal”, “brain”, “dementia”, “syndrome”, “disorder”.

The fourth topic of “joint disease” contained the following words: “joint”, “pain”, “knee”, “back”, “arthritis”, “foot”, “spine”, “leg”, “exercise”, “muscle”, “shoulder”, “bone”, “cartilage”, “posture”, “disk”, “shoe”, “surgery”, “ligament”, “varicose vein”,

Figure 5. Topic trends from 2011 to 2015.

“high heel”.

The fifth topic “infectious disease” contained the following words: “MERS (Middle East Respiratory Syndrome)”, “patient”, “infection”, “virus”, “hospital”, “death”, “large”, “domestic”, “patient”, “confirmation”, “work”, “occurrence”, “suspicion”, “male”, “case”, “possibility”, “symptom”, “afternoon”.

An examination of the proportions of the topics revealed that differences in proportion among the topics were not large after 2011. “Dietary supplement” consistently showed high interest, and the articles related to MERS were highly reported on because of the MERS incident that first occurred in Korea in 2015. Before then, even though some articles related to infectious diseases were retrieved, MERS was not classified as a topic because its numbers were insignificant compared with other topics, but related words were grouped into 1 topic due to the surge in MERS articles. Topic trends for 2011–2015 are shown in Figure 5.

Discussion

LDA analysis showed that the period 1993–2000 had 10 representative topics on women’s health that were identified: “healthcare”, “health consultation”, “pregnancy and childbirth”, “AIDS”, “urinary health”, “mortality statistics”, “foot health”, “women’s life”, “new technology”, and “mental health”. The topics of “cerebrovascular disease”, “skin health”, “kidney

disease”, “dietary supplement”, “thyroid disease”, and “lifestyle disease and prevention” were the newly emerging topics related to women’s health since 2000. Although “pregnancy and childbirth” and “urinary health” were extracted as the same representative topics as before, the previous topic “health consultation” was expanded to “medical service” and the previous topic “foot health” to “arthropathy”.

Examining the main topics in 2006–2010, “hair loss”, “skin care”, and “beauty treatment” were new topics. This indicated that this period was when interest in healthcare and beauty increased which may coincide with an overall improvement in the standard of living. “Pregnancy and childbirth”, “joint disease”, “skin health”, “dietary supplement”, and “lifestyle disease” were extracted as representative topics as before 2006. However, the previous topics of “cerebrovascular disease” and “thyroid disease” were expanded to “climacterium health” and “women disease”. In particular, in the topic related to “lifestyle disease”, several words were related to the prevention of infection against swine flu and were associated with an epochal event.

Examining the main topics of 2011–2015, “infectious disease” was observed to be a new topic. This result also indicated that the social and personal interest in the MERS infection peaked interest for a while. The previous words “skin health”, “gynecology”, “medical service”, “lifestyle disease”, “joint disease”, “skin care”, and “climacterium” were extracted as representative topics as before. It was confirmed that the

Table 7. Change of topic trends by period.

Period	1993-2000	2001-2005	2006-2010	2011-2015
Topic trends (Proportion)	Healthcare (0.874)			
	Health consultation (0.014) →	Medical service (0.526) →	Medical service (0.127) →	Medical service (0.075)
	Pregnancy & childbirth (0.012) →	Pregnancy & childbirth (0.018) →	Pregnancy & childbirth (0.014) →	Gynecology (0.095)
	AIDS (0.016)			
	Urinary health (0.010) →	Urinary health (0.020)		
	Mortality statistics (0.039)			
	Foot health (0.005) →	Arthropathy (0.006) →	Joint disease (0.088) →	Joint disease (0.119)
	Women's life (0.019)			
	New technology (0.001)			
	Mental health (0.010)			
		Cerebrovascular disease (0.017) →	Climacterium health (0.022) →	Climacterium (0.122)
		Skin health (0.023) →	Skin health (0.198) →	Skin health (0.100)
		Kidney disease (0.002)		
		Dietary supplement (0.352) →	Beauty treatment (0.113) →	Dietary supplement (0.132)
		Thyroid disease (0.004) →	Women disease (0.039) →	Women's cancers (0.036)
		Lifestyle disease & prevention (0.032) →	Lifestyle disease (0.110) →	Lifestyle disease (0.130)
				Infectious disease (0.105)
			Hair loss (0.184)	
			Skin care (0.104) →	Skin care (0.087)

previous topic “women disease” tended to become specialized as “women’s cancers”, and the previous topic “hair loss” and “skin care” were integrated into “skin care”.

The characteristics of the 1990s were that healthcare was mainstream, and there were fewer specialized articles on topic areas. However, over time, articles on various topics

were distributed in similar proportions, and the topics were expanded to medically specialized topics.

In the early 2000s, the topic related to beauty was “dietary supplement”. By the late 2000s, however, beauty-related topics began to be subdivided into “skin health”, “hair loss”, and “beauty treatment”. In addition, although “skin health”

was classified as an independent topic in the early 2000s, the number of articles related to “skin health”, which had a small proportion, increased in the late 2000s. This study observed significant changes in the topic of healthcare as shown in Table 7.

In the 1990s, readers were using keywords to search for articles on hospitals or medical counseling possibly to receive medical services. We observed that the healthcare service was represented by keywords for hospitals and support groups in the early 2000s, so it could be postulated that the proportion of families in the support group may have increased in the late 2000s. The expansion of access to the internet since the 2000s may account for the medical service expansion within society and government by 2010; e-health was available and people began using online communities to obtain health-related information.

This study observed changes in the topic of “mental health”. When words presented in the articles related to “mental health” in the 1990s were examined, it was observed that “grandmother” was highly ranked. In addition, words such as “stress”, “thought”, and “talking” were presented together. The word “climacterium” which refers to the menopause, has only recently been discussed in relation to women’s health, and did not appear in the articles until the early 2000s. Considering that the birth rate of women in their 20s has fallen from 88% to 29% over the last 30 years, most of the women currently in their 50s became grandmother in the 1990s and are recent climacterium women. According to the Korean Society of Menopause, about 89% of Korean women in their 50s suffer from climacterium symptoms. In the 1990s women in their 50s would have had the same symptoms however, recently it has been termed the climacterium period. More social consideration and support is available for women. Women’s climacterium symptoms often include difficulty sleeping, stress, and tension. The fact that the words “night”, “stress”, and “anxiety” were presented together in the topic that included “grandmother” for the 1990s supports this result.

With regard to depression, it has been reported that women are twice as likely to develop depression as men, regardless of culture and have higher incidence rates in middle age. It has been shown that depression is derived by the interaction of personal characteristics that are susceptible to depression and negative stress [25]. This is similar to the result of the words related to depression in this study.

Conclusions

Changes in topics related to women’s disease were not clearly distinguished in the 1990s, but in the early 2000s, “thyroid disease” emerged. In the late 2000s, there were many articles on the topic of “women disease” using words such as “uterus”,

“thyroid”, and “breast”. Since the 2010s, articles related to “women disease” changed to articles on the topic of women’s cancers” including “cervical cancer”, “breast cancer”, and “colon cancer”. In the late 2000s, the keyword was “muscle”, in the early 2010s the keywords were “muscle”, “protein”, “calcium”, and “osteoporosis”.

Future studies should analyze the diagnosis rate of diseases that occur in women by using big data from the National Health Insurance Service and the Health Insurance Review & Assessment Service.

Acknowledgments

This work was supported by Kosin University, Republic of Korea (No.: 2018000367).

References

- [1] Jo I. Women’s health issues from the 4th Korean national health and nutrition examination survey (KNHANES 2007), focusing on quality of life, smoking, drinking, nutrition and exercise. *Korean J Womens Health* 2009;10(1):115–52.
- [2] Ahn M. Life-cycle specific comprehensive women’s health and maternal child health. *J Korean Soc Matern Child Health* 2014;18(1):1–12.
- [3] Jung E. AIDS News Framing Analysis: Focusing on Critical Health Journalism Perspectives. *Korean J Journal Commun Stud* 2008;52(4):223–49.
- [4] Andsager JL, Powers A. Framing women’s health with a sense-making approach: Magazine coverage of breast cancer and implants. *Health Commun* 2001;13(2):163–85.
- [5] Roh S, Yoon Y. Analyzing Online News Media Coverage of Depression. *Korean J Commun Inf* 2013;61:5–27.
- [6] Lee H, Ahn S. Analysis on Stigma factors of Suicide Prevention News. *Korean J Journal Commun Stud* 2013;57(4):27–47.
- [7] Kim N, Kim Y. Women’s Overall Health Status: Life Expectancy, Self-Rated Health Status, Activity Limitations. *Health Welf Policy Forum* 2014;210:5–16.
- [8] Mitchell R. Web Scraping with Python: Collecting Data from the Modern Web. O’Reilly Media Inc; 2015.
- [9] Paranyushkin D. Visualization of text’s polysingularity using network analysis. *Prototype Lett* 2011;2(3):256–78.
- [10] Cho Y, Fu P, Wu C. Popular Research Topics in Marketing Journals, 1995–2014. *J Interact Mark* 2017;40:52–72.
- [11] Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: twitter and antibiotics. *Am J Infect Control* 2010;38(3):182–8.
- [12] Michelson M, Macskassy SA. Discovering users’ topics of interest on twitter: a first look. *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*. p.73–80.
- [13] Chen R, Xu W. The determinants of online customer ratings: a combined domain ontology and topic text analytics approach. *Electron Commer Res* 2017;17(1):31–50.
- [14] Qiao Z, Zhang X, Zhou M, et al. A Domain Oriented LDA Model for Mining Product Defects from Online Customer Reviews. 2017.
- [15] Flaounas I, Sudhahar S, Lansdall-Welfare T, et al [Internet]. Big Data Analysis of News and Social Media Content. Available from: <https://www.slideshare.net/FreeNews4All/big-data-analysis-of-news-and-social-media-content>.
- [16] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res* 2003;3(Jan):993–1022.
- [17] Falinouss P [Internet]. Stock trend prediction using news articles: A text mining approach. 2007. Available from: <http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1019373&dsid=9408>.
- [18] DiMaggio P, Nag M, Blei D. Exploiting affinities between topic modeling

- and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics* 2013;41(6):570-606.
- [19] Silge J, Robinson D. *Text Mining with R: A tidy approach*. O'Reilly Media Inc.; 2017.
- [20] Kim J, Baek S. Analysis of Issues on the College and University Structural Reform Evaluation Using Text Big Data Analytics. *Asian J Educ* 2016;17(3):409-36.
- [21] Jeon H [Internet]. Introduction to KoNLP API. 2017 [cited 2017 Sep 18]. Available from: <https://cran.r-project.org/web/packages/KoNLP/vignettes/KoNLP-API.html>.
- [22] Jockers ML. *Text analysis with R for students of literature*. Springer; 2014.
- [23] Grün B, Hornik K. Topicmodels: An R package for fitting topic models. *J Stat Softw* 2011;40(13):1-30.
- [24] Shiryayev AP, Dorofeev AV, Fedorov AR, et al. LDA models for finding trends in technical knowledge domain. *Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 2017; IEEE Conference of Russian. p.551-4.
- [25] Benazon NR, Coyne JC. Living with a depressed spouse. *J Fam Psychol* 2000;14(1):71-9.