Journal of Health Informatics and Statistics

Original Article

J Health Info Stat 2016;41(4):417-427 • https://doi.org/10.21032/jhis.2016.41.4.417

pISSN 2465-8014 | eISSN 2465-8022

소셜 빅데이터 기반 보건복지 정책 미래신호 예측

송태민[1], **송주영**[2]

[1]한국보건사회연구원, [2]펜실베이니아 주립대학

Future Signals of Health and Welfare Policies and Issues using Social Big Data

Tae-min Song[1], **Juyoung Song**[2]

[1]Korea Institute for Health and Social Affairs, Sejong, Korea; [2]Department of Administration of Justice, Pennsylvania State University, PA, USA

Objectives: The purpose of this study is to collect health and welfare-related documents mentioned in and collectable from online channels, analyze important health and welfare keywords through topic and sentimental analyses, detect future signals concerning major policies and issues related to health and welfare services, and propose a prediction model. **Methods:** 201,849 Health & Welfare related online documents from January 1 to March 31, 2016 from 171 Korean online channels and analyzed such documents using machine learning with random forest and Apriori algorithm association analysis. We used R software (version 3.2.1) for the association analysis data mining and visualization. **Results:** As for the prediction of future signals of health and welfare policies, policies that were important and supported by the people were welfare payment, health promotion, job, marriage/childbirth, health insurance, and healthcare industry (in this order). Specifically, as support for documents mentioning welfare payment and jobs was high, job creation through building a spontaneous welfare system is thought to be needed. Additionally, similar to the linkage analysis result of policies, as people were against documents that mentioned only {basic pension} policies, but supported documents that included {basic pension, welfare payment, job}, there is a strong demand for the establishment of a welfare system through active self-support and labor of the elderly. **Conclusions:** Social big data can be utilized in various areas. First, similar to the application in this study, future signals concerning government's policies and new technologies can be predicted in advance and prepared for. Second, they can be used as a new data collection methods that supplement limitations in survey data collection systems. Finally, a preemptive response system against risk can be established through monitoring and predicting social crisis.

Key words: Social big data, Machine learning, Future signals, Health & welfare

서 론

연구의 필요성

우리나라는 2001년부터 2014년까지 합계출산율 1.3 미만의 초저출산율이 10년 이상 지속되고 있고, 기대수명은 1970년 62.1세, 1990년 71.3세, 2013년 81.9세로 지속적으로 증가하고 있다[1]. 또한, 2000년 고령화 사회(노인인구 7%)에 진입한 이후 65세 이상 노인인구 비중은 꾸준히 증가하는 추세에 있다. 2017년에는 노인인구비율이 14%를 넘어 고령사회(aged society)에 진입하고, 2025년에는 20%로 초고령사회에 들어서게 되며 2050년에는 우리나라의 고령화율은 38.2%로 급증하여

Corresponding author: Juyoung Song
200 University Drive, Schuylkill Haven, PA 17972, U.S.A
Tel: +1-570-385-6171, E-mail: jxs6190@psu.edu

Received: September 26, 2016 **Revised:** November 27, 2016 **Accepted:** November 28, 2016

*This research was supported by the ICT R&D program of MSIP/IITP, [R7117-16-0219, Development of Predictive Analysis Technology on Socio-Economics using Self-Evolving Agent-Based Simulation embedded with Incremental Machine Learning], and Some of the articles are listed 'Song Tae Min(2016). Using Social Big Data Predictive Future Signal: With Special Reference to the Major Policy Issues of Health and Welfare, Health and Welfare Policy Forum(228), Korea Institute for Health and Social Affairs' is published.

No potential conflict of interest relevant to this article was reported.

How to cite this article:
Song T, Song J. Future Signals of Health and Welfare Policies and Issues using Social Big Data. J Health Info Stat 2016;41(4):417-427. Doi: https://doi.org/10.21032/jhis.2016.41.4.417

일본(39.6%) 다음으로 노인인구비율이 높은 수준(OECD 평균 25.8%)이 될 것으로 전망하고 있다[1]. 이와 같은 초저출산과 인구고령화로 인해 생산가능 인구는 감소하고 노인부양비가 급증하는 등 '지속가능한 성장'과 '국민행복'의 시대에 큰 걸림돌이 되고 있다. 인구감소와 고령화로 총부양률이 2016년부터 본격적으로 증가하고, 노년 부양비도 2010년 15.2%에서 2050년 71.0% 수준으로 급증할 것으로 예측하고 있다[2]. 그리고 상대적으로 높은 비정규직 비율과 정규직과의 임금격차는 잠재적 복지수요를 증가시키며, 이러한 고용구조와 소득분배구조의 악화로 인한 사회보장비는 증가할 것으로 본다. 이와 같은 보건복지 여건 및 환경변화에 따라 보건복지 수요는 증가하고 있으며 이러한 복지수요에 대응하기 위하여 정부는 2015년에 116조 원을 예산을 보건·복지·고용에 투입하였고 2016년에는 123조 원을 투입하고 있으며, 향후에도 지속적으로 증가될 것으로 판단한다. 그러나 제도의 미성숙, 제도 간의 연계성부족, 복지수요와 공급 간의 조응성 미흡 등으로 국민의 복지 만족도와 행복도는 높지 않는 편이다[1]. 따라서 복지 만족도와 국민 행복도를 높이기 위해서는 국민이 필요로 하는 욕구를 우선순위별로 파악하여 분배정의(distributive justice)에 따라 예산을 배분하고, 객관적인 보건복지 수요조사를 바탕으로 근거중심의 정책개발 및 예산배정이 필요하다. 국내의 대표적인 국민의 복지욕구 조사는 보건복지부·한국보건사회연구원에서 매년 실시하는 '보건복지정책 수요조사 및 분석'과 통계청에서 실시하는 '사회조사'가 있다. 이들 조사의 대부분은 보건복지와 관련된 일부 정책이나 전반적인 정책에 대한 만족도와 복지수준에 대한 인식 조사로서 보건복지와 관련된 다양한 정책의 수요예측은 미흡한 실정이다. 국민이 요구하는 보건복지정책 수요를 예측하기 위해서는 다양한 산업의 종사자나 일반인을 대상으로 설문조사를 실시해야 한다. 기존에 실시하던 횡단적 조사나 종단적 조사 등을 대상으로 한 연구는 정해진 변인들에 대한 개인과 집단의 관계를 보는 데는 유용하나, 사이버 상에 언급된 개인별 담론에서 논의된 관련 변인의 상호 간의 연관관계를 밝히고 원인을 파악하는 데는 한계가 있다[3]. 따라서 보건복지 정책을 성공적으로 추진하여 예상하는 성과를 얻기 위해서는 다양한 보건복지 욕구와 이해집단과의 갈등을 최소화하기 위한 정책동향 및 수요를 예측하여 적시에 대응할 수 있는 체계 구축이 필요하다. 이를 위해서는 오프라인 보건복지 정책 수요 조사와 함께 온라인에서 수집된 보건복지 정책에 대한 미래신호 탐색과 예측을 하여야 한다.

연구의 목적

본 연구는 우리나라에서 수집가능한 모든 온라인 채널에서 언급된 보건복지 관련 문서를 수집하여 주제분석과 감성분석을 통하여 보건복지 주요 키워드를 분류하고 보건복지와 관련하여 나타나는 주요 정책과 이슈에 대한 미래신호를 탐지하여 예측모형을 제시하고자 한다. 본 연구의 목적을 달성하기 위한 구체적인 내용은 다음과 같다.

첫째, 보건복지와 관련한 소셜 빅데이터를 분석하기 위해 주제분석(text mining)과 감성분석(opinion mining)을 실시한다.

둘째, 단어빈도와 문서빈도를 활용하여 보건복지 주요 정책에 대한 신호를 탐지한다.

셋째, 머신러닝 분석을 통하여 탐지된 보건복지 주요 신호에 대한 미래신호를 예측한다.

이론적 배경

2016년 보건복지 주요정책 및 수요

현 정부는 국민이 행복한 사회를 이루기 위한 사회보장 정책방향으로 '생애주기별 맞춤형 복지'를 제시하고, 이를 실현하기 위해 다양한 맞춤형 복지정책을 도입 및 확대하고 있다[4]. 지난 3년간 취약계층 보호 및 사회보장을 위한 생애주기별 맞춤형 복지의 큰 프레임을 구축하였으며, 저소득층의 자립유인 및 실질적 기초생활 지원 강화를 위해 통합급여를 생계·의료·주거·교육급여 등 맞춤형 급여로 개편하였고, 의료보장성 강화 및 노후생활을 지원하였다. 2016년도에는 국민이 체감하는 맞춤형 복지 확산을 목표로 맞춤형 복지제도 내실화(맞춤형 기초생활보장제도 정착, 4대 중증질환 등 의료보장 지속, 맞춤형 보육개편, 기초연금 및 장기요양 지원 확대), 복지사각지대 적극 해소(복지안내 강화, 정부3.0 위기가구 선제발굴, 취약계층 필수서비스 지속 확충, 노후준비 등 불안요인 해소 지원), 읍면동 중심 복지전달체계 구축(읍면동 복지허브화, 읍면동 중심 통합서비스 제공)을 중점 추진 중에 있다. 우리 경제에 혁신과 재도약을 위해 2014년 수립된 경제혁신 3개년 계획의 핵심개혁과제로 선정된 보건·의료서비스업 육성의 일환으로 바이오헬스산업을 새로운 성장동력으로 육성하는 정책을 추진 중에 있다. 2017년 바이오헬스 산업 7대 강국 도약을 목표로 2016년에는 한국의료의 세계적 브랜드화(외국인 환자 유치 촉진, 한국의료 해외진출 확대, 디지털 헬스케어 해외진출), ICT 융합 기반 의료서비스 창출(국민 체감형 원격의료 확산, 진료정보교류 활성화, 의료법 개정), 제약·의료기기 산업 미래먹거리로 육성(신약개발 등 제약산업 육성, 정밀·재생의료 산업 활성화, 첨단 의료기기 개발 지원)을 주요 추진과제로 선정하여 추진하고 있다.

우리나라의 복지수요는 다음의 네 가지 측면에서 증가할 것으로 보고 있다[1]. 첫째, 저출산·고령화, 경제성장률 하락, 높은 비정규직 및 자영업 비율, 빈곤 및 분배구조 변화에 따라 복지수요는 증가될 것으로 추론되고 있다. 둘째, 연금제도의 성숙 등 보건복지제도 성숙에 따라 공공사회복지지출 비중(SOCX 기준)은 2060년에 GDP 대비 약 29.0%로 증가하는 등 복지 수요는 증가할 것으로 예측하고 있다. 셋째,

상병수당, 아동수당 등의 새로운 보건복지제도 도입에 따라 복지수요는 증가할 것으로 보고 있다. 넷째, 사회복지서비스 영역에서 선별복지에서 보편복지로의 전환으로 복지수요는 증가할 것으로 보고 있다.

미래신호 예측

미래변화의 트랜드를 파악하고 미래의 핵심기술을 선별하기 위하여, 주요 선진국들은 주기적으로 국가의 미래트랜드를 분석하고 그 결과를 발표하고 있다[5]. 그동안 US Strategic Business Insight [6], Finland Futures Research Center [7] 등 많은 연구 그룹들은 미래트랜드를 예측하기 위한 다양한 연구가 시도되어 왔으나 대부분 전문가의 지식과 의견에 따라 미래를 전망하는 방법을 사용하여 왔다[8]. 최근 SNS를 비롯한 온라인 채널에서 생산되는 텍스트 형태의 비정형 데이터가 실제 경제 및 사회에 미치는 영향력이 매우 높아짐에 따라 소셜 빅데이터(social big data)를 활용한 미래예측 연구가 진행되고 있으나 수집기술과 분석기술의 어려움으로 활발히 확산되지 못하고 있는 실정이다.

미래의 환경변화를 감지하기 위한 다양한 연구가 시도되고 있으며, 여러 연구 중에서 가장 많은 주목을 받고 있는 것은 미래의 변화를 예감할 수 있는 약신호(weak signal)를 탐지하는 것이다[9,10]. 약신호는 '미래에 가능한 변화의 징후'[11]로 약신호는 시간이 흐르면서 강신호(strong signal)로, 강신호는 다시 트랜드(trend)나 메가트랜드(mega trend)로 발전할 수 있다. Hiltunen [12]은 약신호를 미래신호(future sign)라는 개념을 이용하여 미래신호를 신호(signal), 이슈(issue), 이해(interpretation)와 같이 3차원의 미래신호 공간으로 설명하였다.

온라인 채널에서 수집된 텍스트 형태의 문서를 분석하기 위해서는 텍스트마이닝(text mining)을 통하여 우선적으로 문서 내에서 출현하는 단어별 빈도를 산출해야 한다. 텍스트마이닝 분석을 위해서는 단어빈도(term frequency, TF)와 문서빈도(document frequency, DF) 산출해야 한다. 단어빈도의 산출은 각 문서에서 단어별 출현빈도를 산출하고, 문서별 출현빈도를 합산하여 산출할 수 있다. 문서빈도는 특정단어가 출현하는 문서의 수를 나타낸다. 텍스트마이닝에서 중요한 정보의 추출을 위해서 term frequency-inverse document frequency (TF-IDF) 방법을 사용하고 있다. TF-IDF는 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서에 얼마나 중요한 것인지를 나타내는 통계적 수치이다[5]. Spärck [13]는 희귀한 단어일수록 더 높은 가중치를 부여하기 위해서 역문서빈도(inverse document frequency, $IDF_j = \log_{10}(\frac{N}{DF_j})$)를 제안하였다. 따라서 단어빈도 분석에 희귀한 단어일수록 더 높은 가중치를 부여할 필요가 있다면 단어빈도와 역 문서빈도를 결합하여 'TF-IDF = $TF_{ij} \times IDF_j$'를 산출하여 가중치(단어의 중요도 지수)를 적용한다.

Yoon [9]은 웹 뉴스의 문서를 수집하여 텍스트마이닝 분석을 통해 생성된 단어빈도와 문서빈도를 Hiltunen [12]의 신호와 이슈로 각각 연계하였다. Yoon [9]은 단어빈도, 문서빈도, 발생빈도 증가율을 이용하여 keyword emergence map (KEM)과 keyword issue map (KIM)의 키워드 포트폴리오를 작성하고 작성된 키워드 포트폴리오을 이용하여 약신호를 선별하였다. KEM은 가시성을 보여주는 것으로 degree of visibility (DoV)를 산출하고, KIM은 확산 정도를 보여주는 것으로 degree of diffusion (DoD)를 산출할 수 있다.

$$DoV_{ij} = (\frac{TF_{ij}}{NN_j}) \times \{1 - tw \times (n - j)\}$$

$$DoD_{ij} = (\frac{DF_{ij}}{NN_j}) \times \{1 - tw \times (n - j)\}$$

여기서 NN은 전체 문서수를 의미하고, TF는 단어빈도, DF는 문서빈도, tw는 시간가중치(본 연구에서 시간가중치는 0.05를 적용), n은 전체시간구간, j는 시점을 의미한다. 시간 가중치는 현재부터 시간이 멀어질수록 영향력을 약하게 만드는 기능으로 본 연구에서의 시간가중치는 Yoon [9]이 적용한 0.05를 사용하였다.

분석 방법

분석자료 및 대상

본 연구는 국내의 온라인 뉴스 사이트, 블로그, 카페, 소셜 네트워크 서비스, 게시판 등 인터넷을 통해 수집된 소셜 빅데이터를 대상으로 하였다. 본 분석에서는 149개의 온라인 뉴스사이트, 4개의 블로그(네이버, 네이트, 다음, 티스토리), 2개의 카페(네이버, 다음), 1개의 SNS (트위터), 15개의 게시판(네이버지식인, 네이트지식, 네이트톡, 네이트판 등)의 총 171개의 온라인 채널을 통해 수집 가능한 텍스트 기반의 웹문서(버즈)를 소셜 빅데이터로 정의하였다. 보건복지 관련 토픽의 수집은 2016년 1월 1일부터 3월 31일까지 해당 채널에서 요일별, 주말, 휴일을 고려하지 않고 매 시간단위로 수집하였으며, 수집된 총 201,849건(1월: 87,567건, 2월: 65,278건, 3월: 49,004건)의 텍스트(text) 문서를 본 연구의 분석에 포함시켰다. 본 연구를 위한 소셜 빅데이터의 수집은 SKT 스마트 인사이트에서 크롤러(crawler)를 사용하였고, 토픽의 분류는 주제분석(text mining) 기법을 사용하였다. 보건복지 토픽은 모든 관련 문서를 수집하기 위해 '보건', '복지', 그리고 '보건복지'를 사용하였다.

연구방법

본 연구의 소셜 빅데이터를 분석하기 위해 Figure 1과 같은 연구방법을 사용하였다. 첫째, 수집된 보건복지 온라인 문서를 자연어처리기술을 이용하여 텍스트마이닝과 감성분석(opinion mining)을 실시하였다. 둘째, 분류된 온라인 텍스트 문서를 통계분석과 데이터마이닝 분석을 위해 숫자형태로 코딩하여 정형데이터로 변환하였다. 셋째, 보건

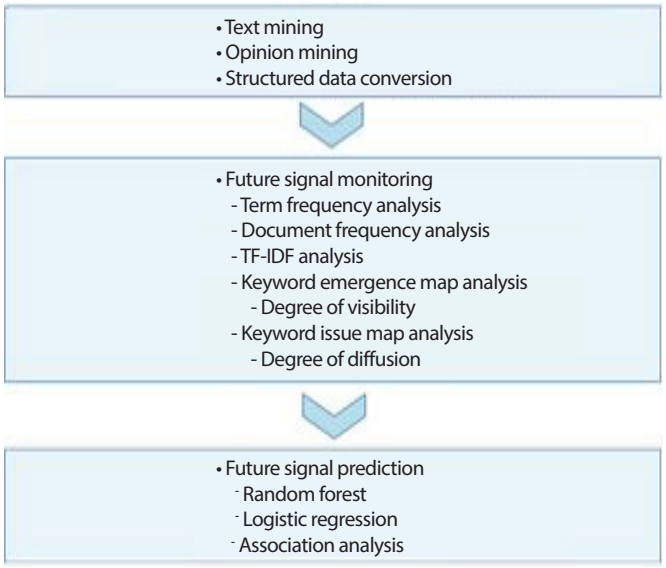

Figure 1. Flowchart of future signal monitoring and prediction.

복지 미래신호를 탐색하기 위해 단어빈도, 문서빈도, TF-IDF를 분석하고, 키워드의 중요도와 확산도를 분석하여 미래신호를 탐색하였다. 넷째, 머신러닝(machine learning) 분석기술을 이용하여 탐색된 미래신호를 중심으로 보건복지 정책의 미래신호를 예측하고 미래신호 간의 연관관계 파악하였다. 본 연구의 머신러닝에 사용된 연관분석 알고리즘으로는 선험적규칙(apriori principle)을 사용하였고, 주요 신호의 예측을 위한 분류방법으로 랜덤포레스트(random forest) 알고리즘을 사용하였다. 머신러닝 분석과 시각화는 R 3.3.1을 사용하였다. 기계학습(machine learning)에서 분류기법 중 하나인 랜덤포레스트는 Breiman [14]에 의해 제안되었다. 랜덤포레스트는 주어진 자료로 부터 여러 개의 예측모형들을 만든 후, 예측모형들을 결합하여 하나의 최종 예측모형을 만드는 기계학습을 위한 앙상블(ensemble) [15] 기법중 하나로 분류정확도가 우수하고 이상치에 둔감하며, 계산이 빠르다는 장점이 있다[16]. 연관분석은 연구자가 지정한 최소 지지도를 만족하는 빈발항목집합(frequent item set)을 생성한 후, 이들에 대한 최저지지도 기준을 마련하고 향상도가 1 이상인 것을 규칙으로 채택한다[17].

연구도구

보건복지 관련 소셜 빅데이터의 수집 및 분류는 보건복지부 홈페이지를 크롤링(crawling)하여 자연어처리와 주제분석을 과정을 거쳐 최종 정책과 이슈를 도출하여 분류하였다. 그리고 본 연구에 사용된 연구도구는 주제분석, 감성분석, 요인분석(factor analysis)의 과정을 거쳐 다음과 같이 정형화 데이터로 코드화하여 사용하였다.

Table 1. Online document of policies and issues on health and welfare

Item	Total
	n (%)
Attitude	70,640 (100.0)
Support	50,626 (71.7)
Oppose	20,014 (28.3)
Issue	19,926 (100.0)
Medical cost	754 (3.8)
Suicide	634 (3.2)
Tuition fee	459 (2.3)
Tax	5,339 (26.8)
Personal information	872 (4.4)
Real estate	483 (2.4)
Polarization	306 (1.5)
Treatment	2,637 (13.2)
Cigarette	383 (1.9)
Tax increase	8,059 (40.4)
Policy	24,059 (100.0)
National pension	738 (3.1)
Basic pension	922 (3.8)
Childcare	400 (1.7)
Marriage/childbirth	2,485 (10.3)
Family-friendly	287 (1.2)
Future generation nurturing	817 (3.4)
Grant policy	1,116 (4.6)
Healthcare privatization	662 (2.8)
Health insurance	1,062 (4.4)
Telemedicine	219 (0.9)
Advanced disease	223 (0.9)
Patient safety	212 (0.9)
Healthcare industry	581 (2.4)
Welfare payment	3,524 (14.6)
Health promotion	3,352 (13.9)
Job	7,459 (31.0)

보건복지 관련 수요

보건복지 감정은 주제분석을 통하여 총 57개(가능, 강화, 개선, 거짓말, 계획, 관심, 규제, 기부, 노력, 논란, 눈물, 다양, 도움, 도입, 마련, 무시, 문제, 반대, 발표, 방문, 부담, 부족, 비판, 사용, 소중, 시행, 신속, 신청, 실시, 실현, 어려움, 억울, 예정, 외면, 운영, 이용, 저지, 정의, 주장, 준비, 중요, 증가, 지원, 지적, 진행, 참여, 최고, 최우선, 추진, 추천, 축소, 폐지, 필요, 행복, 혜택, 확대, 확인) 키워드로 분류되었다. 따라서 본 연구의 종속변수인 보건복지 수요(찬성, 반대)의 정의는 요인분석과 감성분석의 과정을 거쳐 '계획, 예정, 추진, 강화, 실시, 운영, 지원, 확대, 개선, 도움, 관심, 다양, 중요, 참여, 필요, 진행, 노력, 확인, 사용, 가능, 이용, 발표, 혜택, 시행, 신청, 실현, 행복, 정의, 최우선, 소중, 최고'은 찬성의 감정으로, '부족, 지적, 논란, 주장, 비판, 문제, 외면, 축소, 저지, 폐지, 반대, 무시, 부담, 걱정, 거짓말, 준비, 억울, 눈물, 어려움, 규제'은 반대의 감

정으로 정의하였다.

보건복지 관련 정책

보건복지 관련 정책의 정의는 요인분석과 주제분석의 과정을 거쳐 '국민연금요인, 기초연금요인, 보육요인, 결혼출산요인, 가족친화요인, 미래세대육성요인, 무상정책요인, 의료민영화요인, 건강보험요인, 원격의료요인, 중증질환요인, 환자안전요인, 보건산업요인, 복지급여요인, 건강증진요인, 일자리요인'의 16정책으로 해당 정책이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

보건복지 관련 주요이슈

보건복지 관련 주요이슈의 정의는 주제분석의 과정을 거쳐 '의료비, 자살, 등록금, 세금, 개인정보, 부동산, 양극화, 치료, 담배, 증세'의 10개 이슈로 정의하였다. 정의된 모든 이슈는 해당 대상이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

연구 결과

보건복지 정책과 이슈의 온라인 문서 현황

보건복지 정책과 이슈의 온라인 문서 현황을 살펴보면(Table 1), 보건복지 관련 수요는 찬성의 감정을 가진 버즈는 71.7%로 나타났다. 보건복지 관련 주요 정책으로는 일자리(31.0%), 복지급여(14.6%), 건강증진(13.9%), 결혼출산(10.3%), 무상정책(4.6%), 건강보험(4.4%), 기초연금(3.8%) 등의 순으로 나타났다. 보건복지 관련 주요 이슈로는 증세(40.4%), 세금(26.8%), 치료(13.2%), 개인정보(4.4%) 등의 순으로 나타났다.

소셜 빅데이터 기반 미래신호 탐색

보건복지 관련 키워드의 단어 및 문서 빈도 분석

단어빈도, 문서빈도, 단어의 중요도 지수를 고려한 문서의 빈도의 분석을 통하여 보건복지 관련 정책과 주요이슈에 대한 인식변화를 살펴보았다(Table 2). 단어빈도에서는 일자리, 증세, 세금, 복지급여, 결혼

Table 2. Keyword analysis of health/welfare policies and issues in online channels

Ranking	Term frequency		Document frequency		Term frequency - inverse document frequency	
	Keyword	Frequency	Keyword	Frequency	Keyword	Frequency
1	Job	8,212	Tax increase	8,059	Job	6,328
2	Tax increase	8,059	Job	7,459	Tax increase	5,940
3	Tax	5,339	Tax	5,339	Welfare payment	4,955
4	Welfare payment	4,520	Welfare payment	3,524	Tax	4,890
5	Marriage/childbirth	3,419	Health promotion	3,352	Marriage/childbirth	4,267
6	Health promotion	3,352	Treatment	2,637	Health promotion	3,748
7	Treatment	2,938	Marriage/childbirth	2,485	Treatment	3,591
8	Health insurance	1,307	Grant policy	1,116	Health insurance	2,114
9	Grant policy	1,156	Health insurance	1,062	Grant policy	1,845
10	Basic pension	922	Basic pension	922	Basic pension	1,548
11	Personal information	872	Personal information	872	Personal information	1,485
12	Future generation nurturing	817	Future generation nurturing	817	Future generation nurturing	1,414
13	Medical cost	754	Medical cost	754	Medical cost	1,332
14	National pension	738	National pension	738	National pension	1,310
15	Healthcare privatization	686	Healthcare privatization	662	Healthcare industry	1,263
16	Healthcare industry	672	Suicide	634	Healthcare privatization	1,250
17	Suicide	634	Healthcare industry	581	Suicide	1,167
18	Real estate	483	Real estate	483	Real estate	946
19	Tuition fee	471	Tuition fee	459	Childcare	939
20	Childcare	460	Childcare	400	Tuition fee	933
21	Cigarette	383	Cigarette	383	Cigarette	789
22	Family-friendly	348	Polarization	306	Family-friendly	761
23	Polarization	306	Family-friendly	287	Polarization	660
24	Telemedicine	237	Advanced disease	223	Telemedicine	546
25	Advanced disease	233	Telemedicine	219	Advanced disease	535
26	Patient safety	229	Patient safety	212	Patient safety	531
Total		47,547		43,985		55,084

Table 3. DoV mean increase rate and mean term frequency (TF) for health/welfare policies and issues

Keyword	DoV			Mean increase rate	Mean term frequency
	January	February	March		
Job	0.142	0.168	0.186	0.147	2,737
Tax increase	0.158	0.237	0.063	-0.116	2,686
Tax	0.136	0.098	0.077	-0.246	1,780
Welfare payment	0.11	0.067	0.094	0.008	1,507
Marriage/childbirth	0.064	0.049	0.097	0.369	1,140
Health promotion	0.062	0.051	0.094	0.339	1,117
Treatment	0.045	0.07	0.061	0.208	979
Health insurance	0.026	0.02	0.034	0.245	436
Grant policy	0.024	0.021	0.024	0.014	385
Basic pension	0.013	0.02	0.023	0.33	307
Personal information	0.005	0.04	0.004	3.394	291
Future generation nurturing	0.01	0.005	0.04	3.498	272
Medical cost	0.015	0.01	0.022	0.415	251
National pension	0.007	0.014	0.026	0.886	246
Healthcare privatization	0.003	0.004	0.04	5.242	229
Healthcare industry	0.011	0.013	0.017	0.216	224
Suicide	0.006	0.005	0.03	2.167	211
Real estate	0.009	0.012	0.007	-0.034	161
Tuition fee	0.011	0.007	0.011	0.1	157
Childcare	0.008	0.012	0.006	0.015	153
Cigarette	0.005	0.005	0.014	0.827	128
Family-friendly	0.004	0.004	0.014	1.275	116
Polarization	0.01	0.003	0.004	-0.226	102
Telemedicine	0.005	0.005	0.004	-0.103	79
Advanced disease	0.007	0.004	0.003	-0.307	78
Patient safety	0.002	0.008	0.003	1.062	76
Median				0.23	249

출산, 건강증진, 치료 등의 순위로 나타나고 있어 정책은 일자리, 복지급여, 결혼출산이 우선이고 주요이슈는 증세, 세금, 치료가 우선인 것으로 나타났다. 문서빈도는 단어 빈도와 비슷한 추이를 나타내고 있으나 결혼출산이 단어빈도에서는 5위인 반면 문서빈도에서는 7위로 나타나 키워드의 중요성을 나타내는 단어빈도에서는 결혼출산이 중요하나 주제의 확산을 나타내는 문서빈도에서는 다소 떨어져 결혼출산 정책의 확산에 대한 노력이 필요할 것으로 본다. 중요도 지수를 고려한 단어빈도에서는 정책은 일자리, 복지급여, 결혼출산이 우선이고 주요이슈는 증세, 세금이 우선인 것으로 나타났다. 그리고 키워드의 월별 순위의 변화는 2016년 2월까지 증세, 일자리, 세금, 복지급여, 치료가 중요한 키워드로 나타나다가 3월에는 건강증진이 강조되기 시작하여 건강에 대한 관심이 확산되고 있는 것으로 나타났다.

보건복지 관련 키워드의 미래신호 탐색

미래신호 탐지방법론에 따라 분석한 결과는 Tables 3, 4와 같다. 보건복지 관련(정책, 이슈) 키워드에 대한 DoV 증가율과 평균단어 빈도를 산출한 결과 일자리와 복지급여는 높은 빈도를 보이고 있으나 DoV 증가율은 중앙값 보다 낮게 나타나 시간이 갈수록 신호가 약해지는 것으로 나타났다. 결혼출산, 건강증진은 평균단어 빈도는 높게 나타났으며, DoV 증가율은 중앙값 보다 높게 나타나 시간이 갈수록 빠르게 신호가 강해지는 것으로 나타났다.

미래신호 탐색을 위해 DoV의 평균단어빈도와 DoD의 평균문서빈도를 X축으로 설정하고 DoV와 DoD의 평균증가율을 Y축으로 설정한 후, 각 값의 중앙값을 사분면을 나누면 2사분면에 해당하는 영역의 키워드는 약신호가 되고 1사분면에 해당하는 키워드는 강신호가 된다. 빈도수 측면에서는 상위 10위에 DoV는 일자리, 증세, 세금, 복지급여, 결혼출산, 건강증진, 치료, 건강보험, 무상정책, 기초연금 순으로 포함되었고, DoD에는 증세, 일자리, 세금, 복지급여, 건강증진, 치료, 결혼출산, 무상정책, 건강보험, 기초연금의 순으로 포함되었다. DoV의 증가율의 중앙값(0.23) 보다 높은 증가율을 보이는 키워드는 결혼출산, 건강증진, 건강보험으로 나타났으며 DoD의 증가율의 중앙값(0.23) 보다 높은 증가율을 보이는 키워드는 건강증진, 건강보험으로 나타났다. 특

Table 4. DoD mean increase rate and mean document frequency (DF) for health/welfare policies and issues

Keyword	DoV			Mean increase rate	Mean document frequency
	January	February	March		
Tax increase	0.172	0.251	0.07	-0.131	2686
Job	0.138	0.163	0.186	0.159	2486
Tax	0.148	0.104	0.085	-0.238	1780
Welfare payment	0.092	0.055	0.082	0.047	1175
Health promotion	0.067	0.054	0.104	0.369	1117
Treatment	0.043	0.068	0.06	0.221	879
Marriage/Childbirth	0.054	0.042	0.069	0.224	828
Grant policy	0.026	0.021	0.026	0.015	372
Health insurance	0.022	0.017	0.032	0.309	354
Basic pension	0.014	0.021	0.026	0.337	307
Personal information	0.005	0.042	0.005	3.278	291
Future generation nurturing	0.011	0.005	0.044	3.683	272
Medical cost	0.016	0.011	0.024	0.453	251
National pension	0.008	0.015	0.028	0.902	246
Healthcare privatization	0.003	0.003	0.044	6.347	221
Suicide	0.007	0.006	0.033	2.279	211
Healthcare industry	0.01	0.013	0.015	0.229	194
Real estate	0.01	0.013	0.008	-0.04	161
Tuition fee	0.011	0.007	0.012	0.12	153
Childcare	0.008	0.012	0.005	0.015	133
Cigarette	0.006	0.005	0.015	0.875	128
Polarization	0.011	0.003	0.004	-0.203	102
Family-friendly	0.004	0.004	0.012	1.047	96
Advanced disease	0.007	0.003	0.003	-0.265	74
Telemedicine	0.005	0.005	0.004	-0.119	73
Patient safety	0.002	0.008	0.003	1.27	71
Median				0.23	249

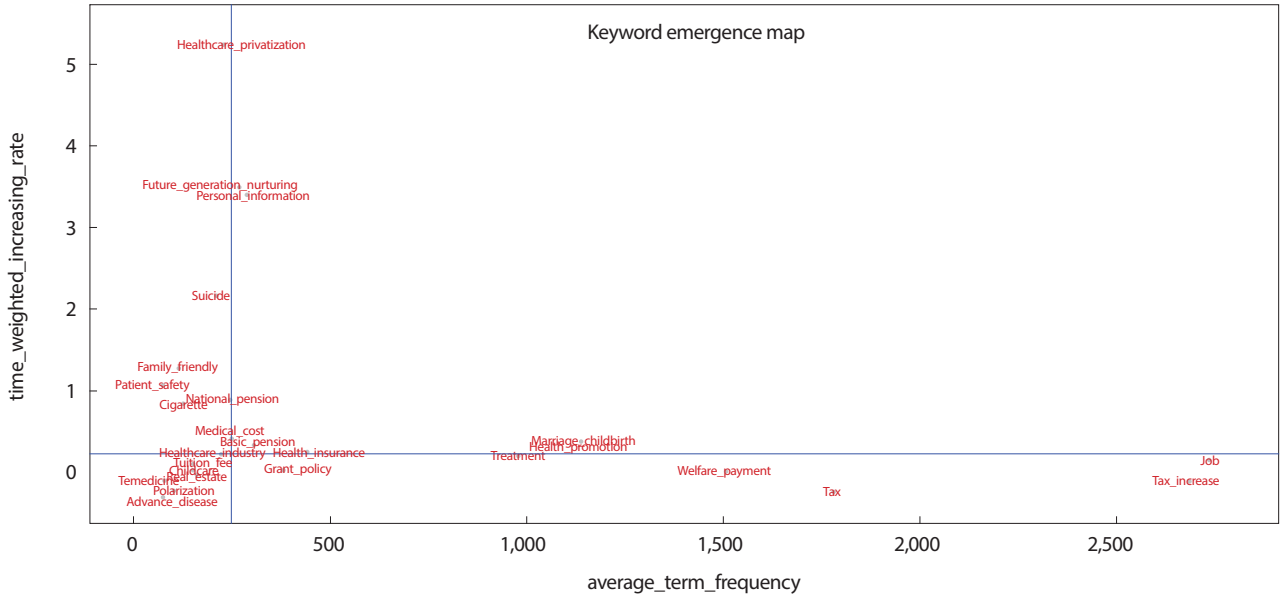

Figure 2. Keyword emergence map (KEM) of health/welfare related policies and issues.

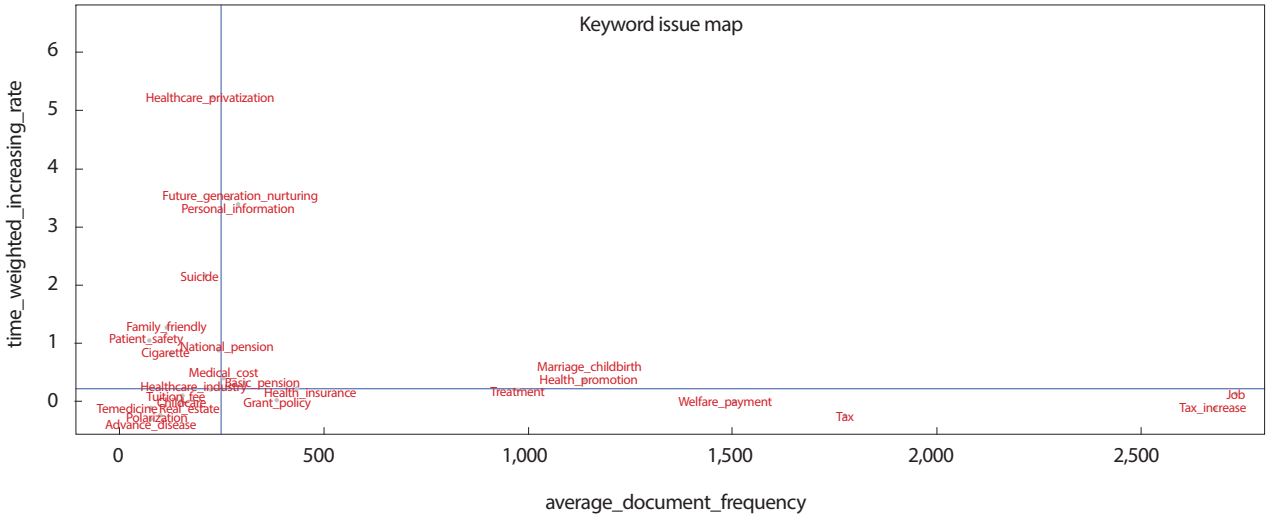

Figure 3. Keyword issue map (KIM) of health/welfare related policies and issues.

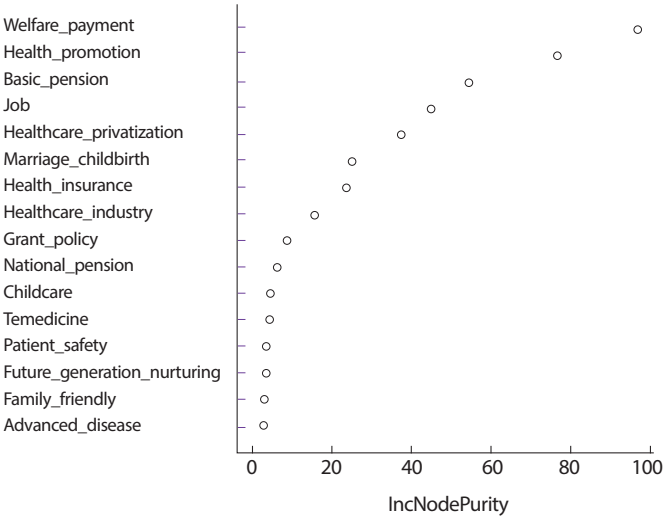

Figure 4. Importance of major health and welfare policies in the random forest model.

Table 5. Major determinants of health and welfare demand

Policy	Support[1]			
	B[2]	SE	OR	*p*-value
National pension	0.25	0.15	1.29	0.098
Basic pension	-1.48	0.10	0.23	0
Childcare	1.44	0.32	4.22	0
Marriage/childbirth	0.63	0.07	1.88	0
Family-friendly	1.12	0.30	3.06	0
Future generation nurturing	0.80	0.21	2.22	0
Grant policy	0.41	0.12	1.509	0.001
Healthcare privatization	-1.63	0.12	0.20	0
Health insurance	1.18	0.13	3.26	0
Temedicine	1.13	0.32	3.11	0
Advanced disease	1.04	0.34	2.82	0.002
Patient safety	1.30	0.35	3.68	0
Healthcare industry	1.63	0.23	5.08	0
Welfare payment	1.23	0.07	3.41	0
Health promotion	1.04	0.07	2.83	0
Job	0.50	0.05	1.65	0

SE, standard error; OR, odds ratio.
[1]Basic category is opposition.
[2]Standardized coefficients.

히 결혼출산의 DoV증가율을 중앙값 보다 높은 반면 DoD의 증가율은 중앙값보다 낮게 나타나 결혼출산 정책의 확산을 위한 방안이 필요할 것이다.

Figures 2, 3과 같이 보건복지 관련 주요 키워드는 복지급여와 일자리는 KEM에서는 강신호로 나타난 반면 KIM에서는 강하지는 않지만 잘 알려진 신호로 나타났다. KEM과 KIM에 공통적으로 나타나는 강신호(1사분면)에는 미래세대육성, 개인정보, 국민연금, 의료비, 기초연금, 건강보험, 결혼출산, 치료, 건강증진이 포함되었고, 약신호(2사분면)에는 의료민영화, 자살, 환자안전, 가족친화, 담배, 보건산업이 포함된 것으로 나타났다. KIM의 4사분면에만 나타난 강하지는 않지만 잘알려진 신호는 무상정책, 복지급여, 세금, 일자리, 증세로 나타났으며, KIM의 3사분면에만 나타난 잠재신호는 등록금, 보육, 부동산, 원격의료, 양극화, 중증질환으로 나타났다.

소셜 빅데이터 기반 미래신호 예측

랜덤포레스트 분석을 통한 주요 보건복지 정책요인 예측

본 연구의 랜덤포레스트 분석을 활용하여 보건복지 수요(태도)에 영

Table 6. Association rules of major health/welfare policies

Rule	Support	Confidence	Lift
{Healthcare privatization} = >{Opposition}	0.002845413	0.57925072	2.0444824
{Basic pension} = > {Opposition}	0.003680634	0.44444444	1.5686797
{Health insurance, advanced disease} = > {Support}	0.001670442	1	1.3953305
{Childbirth/nurture, health promotion} = > {Support}	0.001528879	1	1.3953305
{Childbirth/nurture, welfare payment} = > {Support}	0.002180068	1	1.3953305
{Healthcare industry, welfare payment} = > {Support}	0.001005096	1	1.3953305
{National pension, welfare payment} = > {Support}	0.002109287	1	1.3953305
{Health insurance, welfare payment} = > {Support}	0.002831257	1	1.3953305
{Childbirth/nurture, welfare payment, health promotion} = > {Support}	0.001090034	1	1.3953305
{Childbirth/nurture, health promotion, job} = > {Support}	0.001005096	1	1.3953305
{Childbirth/nurture, welfare payment, job} = > {Support}	0.00137316	1	1.3953305
{National pension, health insurance, welfare payment} = > {Support}	0.001061721	1	1.3953305
{National pension, welfare payment, job} = > {Support}	0.001160815	1	1.3953305
{Health insurance, welfare payment, health promotion} = > {Support}	0.001061721	1	1.3953305
{Health insurance, welfare payment, job} = > {Support}	0.001443941	1	1.3953305
{Advanced disease, welfare payment} = > {Support}	0.001613817	0.99130435	1.3831972
{Basic pension, welfare payment, job} = > {Support}	0.001359003	0.98969072	1.3809456
{Welfare payment, health promotion, job} = > {Support}	0.002548131	0.98901099	1.3799972
{Welfare payment, health promotion} = > {Support}	0.006129672	0.98858447	1.379402
{National pension, health insurance, job} = > {Support}	0.001160815	0.98795181	1.3785193

향을 주는 주요 정책요인을 살펴보면 Figure 4와 같다. 보건복지 수요(찬성, 반대)에 가장 큰 영향을 미치는(연관성이 높은) 정책요인은 '복지급여' 정책으로 나타났으며, 그 뒤를 이어 건강증진, 기초연금, 일자리, 의료민영화, 결혼출산, 건강보험, 보건산업 정책 등의 순으로 나타났다.

랜덤포레스트의 중요도로 나타난 정책요인들이 보건복지 수요에 미치는 영향을 로지스틱회귀분석을 통하여 살펴본 결과, 기초노령연금, 의료민영화는 반대의 확률이 높으며, 그외 국민연금($p<0.1$), 보육, 출산양육, 가족친화, 건강보험, 원격의료 등은 찬성의 확률이 높은 것으로 나타났다(Table 5).

연관분석을 통한 주요 보건복지 정책요인 예측

소셜 빅데이터 분석에서 연관분석은 하나의 온라인 문서에 포함된 둘이상의 단어들에 대한 상호관련성을 발견하는 것이다. 본 연구에서는 Table 6과 같이 하나의 문서에 나타난 정책요인의 수요에 대한 연관규칙을 분석하였다. {의료민영화} => {반대} 두 변인의 연관성은 지지도 0.003, 신뢰도는 0.579, 향상도는 2.044로 나타났다. 이는 온라인 문서에서 '의료민영화' 정책이 언급되면 반대할 확률이 57.9%이며, 이는 '의료민영화' 정책이 언급되지 않은 문서보다 반대할 확률이 약 2.04배 높아지는 것을 나타낸다. {건강보험요인, 중증질환요인} => {찬성}으로 세변인의 향상도는 1.40으로 온라인 문서에서 건강보험과 중증질환 정책이 언급되지 않은 문서보다 찬성할 확률이 1.40배 높은 것으로 나타났다.

고 찰

본 연구는 우리나라에서 수집가능한 모든 온라인 채널에서 언급된 보건복지 관련 문서를 수집하여 보건복지와 관련하여 나타나는 주요 정책과 이슈에 대한 미래신호를 탐지하여 예측모형을 제시하고자 하였다. 본 연구의 분석을 위하여 171개의 온라인 채널을 통해 수집된 온라인 문서를 대상으로 자연어처리기술을 이용하여 텍스트마이닝과 감성분석을 실시하였다. 보건복지 미래신호를 탐색하기 위해 단어빈도, 문서빈도, TF-IDF를 분석하고, 키워드의 중요도와 확산도를 분석하여 미래신호를 탐색하였다. 그리고 머신러닝 분석기술을 이용하여 탐색된 미래신호를 중심으로 보건복지 정책의 미래신호를 예측하고 미래신호 간의 연관관계 파악하였다.

본 연구의 보건복지의 정책과 이슈에 대한 미래신호 예측 결과를 살펴보면 다음과 같다.

첫째, 본 연구의 주제분석과 감성분석을 통한 2016년 보건복지 수요를 예측한 결과 찬성의 감정을 가진 문서는 71.7%, 반대의 감정을 가진 문서는 28.3%로 나타났다. 이는 2015년 보건복지정책 수요조사 분석[1]에서 일반국민의 전반적인 의료만족도가 만족(72.7%), 불만족(27.3%)로 나타나 비슷한 추이를 보는 것으로 나타났다. 따라서 본 연구의 첫 번째 연구목적인 주제분석과 감성분석을 통한 수요분석에 대한 타당성을 어느 정도 입증한 것으로 본다.

둘째, 본 연구의 보건복지 정책과 이슈의 미래신호 분석에서 미래세

대육성, 국민연금, 기초연금, 건강보험, 결혼출산, 건강증진, 개인정보, 의료비, 치료가 강신호로 분류되어 미래세대육성과 건강증진과 관련된 의료비와 치료 등이 강조되고 있는 것으로 나타났다. 이는 2015년 보건복지정책 수요조사 분석[1]에서 보건복지 정책의 우선순위로 의료비부담을 낮추기 위한 정책, 국민기초생활보장제도 개편, 안전한 보건의료체계, 노인들을 위한 소득보장강화, 맞춤형 보육서비스, 보건산업육성, 노인을 위한 건강증진 등의 순으로 나타나 보건복지 정책에 대한 강신호는 비슷한 추이를 보이는 것으로 나타났다. 특히, 미래세대육성과 개인정보는 강신호이면서 높은 증가율을 보이고 있어 미래세대육성 정책에 포함된 아동학대 문제의 해결과 개인정보보호와 관련된 제도개선에 대한 논의가 지속적으로 이루어져야 할 것으로 본다.

셋째, 의료민영화, 자살, 환자안전, 가족친화, 담배, 보건산업은 약신호로 분류되었다. 특히 약신호인 의료민영화와 자살은 높은 증가율을 보이고 있어 이들 키워드는 시간이 지나면 강신호로 발전할 수 있기 때문에 이에 대한 대응책이 마련되어야 할 것으로 본다. 보건복지 정책의 미래신호 예측에서 중요한 정책이면서 찬성하는 정책은 복지급여, 건강증진, 일자리 결혼출산, 건강보험, 보건산업 등의 순으로 나타났다. 이는 2015년 보건복지정책 수요조사 분석[1]에서 2016년 복지분야 중점정책에 대해 일자리 창출을 통한 탈빈곤 정책이 가장 높은 응답비율을 나타나 일자리창출에 대한 신호가 중요한 것으로 나타났다. 특히 복지급여와 일자리가 동시에 언급된 문서의 찬성이 매우 높은 것으로 나타나 능동적 복지체계 구축을 통한 일자리 창출이 필요할 것으로 본다. 이는 정책의 연관분석 결과와 같이 {기초연금} 정책만 언급된 문서는 반대하는 것으로 나타났으나 {기초연금, 복지급여, 일자리} 정책이 동시에 언급된 문서는 찬성하는 것으로 나타나 노인의 능동적 자활과 근로를 통한 복지체계의 구축에 대한 국민의 요구가 높은 것으로 본다. 넷째, {건강보험, 중증질환}이 동시에 언급된 문서의 찬성이 높은 것으로 나타나 건강보험 혜택 확대로 4대중증질환의 보장성 강화가 국민의 의료비 부담을 줄임으로써 정부의 정책에 대한 좋은 평가결과로 나타난 것으로 본다.

위의 연구결과를 바탕으로 정책을 제언하면 다음과 같다.

첫째, 생애주기별 맞춤형 복지정책을 위해 분야별, 대상자별로 다양한 보건복지 욕구를 적시에 파악하여 이들의 욕구를 충족시킬 수 있어야 한다. 둘째, 보건복지 정책 수행 과정 중 발생할 수 있는 문제점이나 한계점을 파악하여 적절한 대책을 마련하기 위한 대응 체계 구축이 필요하다. 셋째, 보건복지 정책의 효과적인 수행을 위해 보건복지 정책 수요예측 및 동향파악을 위한 적시대응 체계를 구축할 필요가 있다.

본 연구의 제한점은 다음과 같다. 첫째, 본 연구는 2016년 1월부터 3월까지 3개월간 제한된 소셜 빅데이터를 수집하여 분석함으로써 보건복지 정책의 미래신호 예측에 한계가 있을 수 있다. 따라서 실질적인 보건복지 정책과 이슈의 미래신호를 예측하기 위해서는 연도별 시계열 자료를 수집하여 분석한 후 결과를 도출해야 할 것으로 본다. 둘째, 본 연구는 개개인의 특성을 가지고 분석한 것이 아니고 그 구성원에 속한 전체 집단의 자료를 대상으로 분석하였기 때문에 이를 개인에게 적용하였을 경우 생태학적 오류(ecological fallacy)가 발생할 수 있다[3]. 또한 본 연구에서 정의된 보건복지 관련 요인은 문서내에서 발생된 단어의 빈도로 정의되었기 때문에 기존 조사 등을 통한 이론적 모형에서의 의미와 다를 수 있다.

결 론

본 연구 결과와 같이 소셜 빅데이터의 분석은 다양한 분야에 활용할 수 있다[18].

첫째, 조사를 통한 기존의 정보수집 체계의 한계를 보완할 수 있는 새로운 자료수집 방법으로 활용할 수 있다. 통일에 대한 국민 인식 조사, 정부의 금연정책(가격정책·비가격정책 등) 실시 이후 흡연 실태 조사, 스마트폰 및 인터넷 중독 실태 조사 등 여러 분야의 조사에 활용할 수 있다.

둘째, 보건복지정책 수요를 예측할 수 있다(저출산정책 수요 예측 등). 새정부 출범 이후 건강보험보장성 강화에 대한 국민의 요구가 커지고 인구고령화와 저출산이 사회적 문제로 대두됨에 따라 대상자별·분야별로 다양한 보건복지정책이 요구되고 있다. 이러한 변화에 대응하기 위해 오프라인 보건복지 욕구 조사와 더불어 소셜미디어에 남긴 다양한 정책 의제를 분석하여 수요를 파악해야 한다.

셋째, 사회적 위기상황에 대한 모니터링과 예측으로 위험에 대한 사전 대응체계를 구축할 수 있다. 예를 들면 청소년 자살과 사이버폭력 대응체계 구축, 질병에 대한 위험 예측, 식품안전 모니터링 등에 활용할 수 있다.

넷째, 새로운 기술에 미래신호를 사전에 예측하여 대비할 수 있다. 빅데이터, 사물인터넷, 머신러닝(인공지능)과 같은 새로운 기술에 대해 수요자와 공급자가 요구하는 기술 동향 등에 대한 미래신호를 탐색하여 예측할 수 있다.

끝으로 정부와 공공기관이 보유·관리하고 있는 빅데이터는 통합방안보다는 각각의 빅데이터의 집단별 특성을 분석하여 위험(또는 수요) 집단 간 연계를 통한 예측(위험 예측 또는 질병 예측 등) 서비스를 제공하여야 할 것이다. 즉 빅데이터 분석을 통한 개인별 맞춤형 서비스는 프라이버시를 침해할 수 있기 때문에 위험 집단별 맞춤형 서비스를 제공하여야 한다[19]. 또한 빅데이터를 분석하여 인과성을 발견하고 미래를 예측하기 위해서는 데이터 사이언티스트 양성을 위한 정부 차원의 노력이 필요하다.

REFERENCES

1. Kim MG, Yeo YJ, Kim SA, Kim JH, Choi MJ. 2015 Health and welfare policy demand survey and analysis. Sejong: Korea Institute for Health and Social Affairs; 2015 (Korean).
2. Statistics Korea. Availablle at http://www.kosis.kr
3. Song TM, Song J, An JY, Hayman LL, Woo JM. Psychological and social factors affecting internet searches on suicide in Korea: a big data analysis of google search trends. Yonsei Med J 2014;55(1):254-263 (Korean).
4. Ministry of Education, Ministry of Labor, Health and Welfare, Women and Family Affairs. 2016 National happiness sector work plan. Sejong: Ministry of Education; 2016 (Korean).
5. Jeong G. A study of future prediction method using text mining and network analysis. Seoul: Korea Institute of Science & Technology Evaluation and Planning; 2014 (Korean).
6. The Department for Business, Innovation & Skills (BIS). Horizon scanning centre. Available at http://www.bis.gov.uk/foresight/our-work/horizon-scanning-centre
7. TrendWiki. TrendWiki homepage. Available at http://www.trendwiki.fi/en/
8. Yoo SH, Park HW, Kim KH. A study on exploring weak signals of technology innovation using informetrics. J Technol Innov 2009;17(2):109-130 (Korean).
9. Yoon J. Detecting weak signals for long-term business opportunities using text mining of web news. J Expert Syst Appl 2012;39(16):12543-12550.
10. Park C, Kim H. A study of development direction of new industries through the internet of things-detecting future signals using text mining. Ulsan: Korea Energy Economics Institute; 2015 (Korean).
11. Ansoff HI. Managing strategic surprise by response to weak signals. Californian Manag Rev 1975;18(2):21-33.
12. Hiltunen E. The future sign and its three dimensions. Futures 2008;40:247-260.
13. Spärck JK. A statistical interpretation of term specificity and its application in retrieval. J Document 1972;28:11-21. Doi:10.1108/eb026526.
14. Breiman L. Random forest. Machine learning 2001;45(1):5-32.
15. Breiman L. Bagging predictors. Machine Learning 1996;26:123-140.
16. Jin JH, Oh MA. Data analysis of hospitalization of patients with automobile insurance and health insurance: a report on the Patient Survey. J Korea Data Analysis Soc 2013;15(5B):2457-2469 (Korean).
17. Park HC. Proposition of causal association rule thresholds. J Korean Data Inf Sci Soc 2013;24(6):1189-1197.
18. Song TM, Song J. Social big data research methodology with R. Seoul: Hannarae Academy; 2016 (Korean).
19. Song Tae Min, Juyoung Song. Cracking the big data analysis. Seoul: Hannarae Academy; 2015 (Korean).