https://doi.org/10.5392/JKCA.2024.24.10.321

대중 매체에서 텍스트 마이닝 분석한 한국의 저출산 요인 : 네이버 뉴스 기사 댓글 중심으로

Factors of Korean Low Birthrate by Text Mining Analysis in Mass Media : Focusing on Naver News Article Comments

남가영, 김용민
전남대학교 디지털미래융합서비스협동과정

Ga-Young Nam(alltoclaire@gmail.com), Yong-Min Kim(ymkim@chonnam.ac.kr)

요약

한국 저출산 문제는 단순한 인구감소 이상의 심각한 문제로, 경제, 사회, 복지 등 다양한 분야에 영향을 미치고 있으며, 복합적인 요인들로 인해 출산율은 낮은 수준에 머물러 있다. 기존 요인분석 연구들은 특정 기사 내용에 기반한 제한적인 범위로, 저출산 대응의 성격을 포괄적으로 바라보는 데에 제한점을 가지고 있다. 본 연구에서는 연구자의 주관적 해석이라는 기존 질적 분석에서 나타나는 단점을 보완하고, 다수 대중의 의견을 반영한 양적 분석을 통해 신뢰도와 타당성을 높이기 위해 텍스트 마이닝 방법을 이용하였으며. 2022년 3월부터 2023년 8월까지 기간 중의 네이버 뉴스 댓글을 수집하여 대중들의 다양한 의견을 반영하였다. 분석 결과, 저출산 잠재요인으로 총 9가지를 식별하였고, 기존 연구와 비교하여 유사한 토픽 6가지와 본 연구에서만 나타난 토픽 3가지를 바탕으로 저출산 문제를 분석 및 논의하였다.

■ 중심어 : | 저출산 | 댓글 | 키워드 빈도분석 | 감성 분석 | LDA 토픽 모델링 |

Abstract

South Korea's declining birthrate is more than just a population issue, it is a serious problem that affects various sectors, including the economy, society, and welfare, and is caused by a combination of factors. Existing factor analysis studies have a limited scope based on the content of specific articles, which limits their ability to provide a comprehensive view of the nature of the response to the declining birthrate. In this study, we used a text mining method to compensate for the shortcomings of existing qualitative analysis, which is subjective interpretation by researchers, and to increase reliability and validity through quantitative analysis that reflects the opinions of the majority of the public, and collected comments on Naver News from March 19, 2022 to August 24, 2023 to reflect the diverse opinions of the public. As the results of the analysis, we identified a total of nine potential factors for the declining birthrate, and analyzed and discussed the issue of the declining birthrate based on six similar topics compared to previous studies and three topics unique to this study.

■ keyword : | Low Birth Rate | Comments | Keyword Analysis | Emotional Analysis | LDA Topic Modeling |

접수일자 : 2024년 06월 25일
수정일자 : 2024년 08월 08일
심사완료일 : 2024년 08월 09일
교신저자 : 김용민, e-mail : ymkim@chonnam.ac.kr

I. 서 론

저출산은 현재 한국사회에서 가장 뜨거운 사회적 이슈이다[1]. 2022년 한국 합계출산율은 0.78로 출생통계 작성이래 역대 최저치로 지구촌에서 출산율이 가장 낮은 나라이다. 합계출산율은 한 여성이 평생 낳을 것으로 예상되는 평균 출생아 수를 나타낸 지표이다[2]. OECD(Organization for Economic Cooperation and Development)에 따르면 38개 회원국 평균 합계출산율(2021년 기준으로)은 1.58명으로 회원국 중에 합계출산율이 0명대인 국가는 한국이 유일한 상황이다[3]. 2000년대 중반 이후 저출산·고령화 문제의 심각성을 인지하고 출산을 장려하는 인구정책으로 방향을 회귀하였다[4]. 정부는 21년간 천문학적인 액수의 예산을 투입하였으며, 나름의 정책적 성과도 있었지만, 출산율은 반등하지 않고 오히려 낮아지고 있다[5]. 급격한 사회 환경의 변화가 생기며 삶에 대한 위기감과 불확실성이 높아짐에 따라 현재 국민은 교육, 보육정책 등에 대한 기대와 우려의 목소리를 동시에 높이고 있다[6]. 또한, 저출산 문제는 평균수명 증가에 따른 인구 고령화 현상과도 맞물려 여러 문제를 야기한다. 출산율 하락은 곧 노동력 감소로 이어지게 되고 생산성 저하라는 결과를 낳게 되어 경제 성장률이 더디게 되는 현상이 발생한다[7]. 결국, 저출산 현상은 경제, 사회, 복지 등의 사회 전반에 걸친 부정적인 사회문제를 초래할 것으로 전망된다[8].

출산은 개인과 개인이 만나 하게 되는 매우 미시적인 현상이며 선택의 결과이다. 경제학적 합리적 선택이론에 따르면 인간은 특정한 행위를 선택할 때 그 행위에 대한 효용과 비용의 선택에서 효용은 크게, 비용은 작은 방향으로 선택한다고 가정한다[9]. 출산은 아이를 갖고자 하는 욕구가 먼저 생기고, 아이를 임신하고 낳는 출산행위에 이르기까지 나타나는 일련의 출산 결정 과정으로 보아야 한다. 즉, 출산은 출산행위가 이루어지기 이전 의도가 먼저 생성되어야 하는데 이는 의도와 행위 간에 연결고리가 이어질 수도 있고 단절될 수도 있기 때문이고, 출산율 개선의 측면에서 볼 때 실제 출산 결정을 하게 된 요인들의 탐색뿐 아니라 출산을 고려하는데 영향을 미치는 요인들도 분석함으로써 더 효과적인 접근을 해야 한다[10].

따라서 본 연구는 저출산 현상에 대한 파급효과, 원인 규명, 정책 효과성 평가 등 기존의 저출산 관련 연구들과는 달리 특정한 관점에서 저출산 현상을 바라보기보다는 그동안 누적된 저출산 관련 논의 전체를 아우른다[11][12]. 대중들의 의견들을 통해 저출산 자체에 주목하여 한국의 저출산 문제를 종합적으로 이해하고, 요인들을 찾아내고자 한다.

이를 위해 본 연구는 국내 최대 온라인 포털사이트인 네이버 제휴 뉴스에서 최근 기사 속 댓글을 대상으로 분석하여, '저출산'이라는 사회적 문제에 대해 텍스트 마이닝이라는 분석 방법론을 적용하여 기존 연구들과의 차별성을 강조하기 위하여 빅데이터를 수집하고, 텍스트 마이닝 기법을 적용하여 분석한 토픽 결과를 바탕으로 저출산에 대한 대중들의 다각적인 시각을 포착하여 요인들을 도출해내고 분석한다. 이를 통해 저출산 문제에 대한 대중의 사회적 인식과 복잡성을 이해하고, 분석 결과와 선행연구 결과의 비교를 통해 앞으로의 저출산 문제에 맞서 나아가야 할 종합적인 방향성을 의논하고자 한다. 특히, 본 연구에서는 기사라는 대중매체가 갖는 보급성과 공공성에 주목한다.

II. 관련연구

1. 저출산 현황 연구

저출산 현상은 출산율이 낮아지는 사회 현상이며, '하위대체 출산율'로 특정 어느 지역에서 이전 세대보다 새로운 세대에서 더 적은 인구를 갖게 된다는 합계출산율을 의미한다[13]. 우리나라는 2001년 1.31명으로 진입하여 초저출산 국가가 되면서, 2001년 이후 21년 동안 연속적으로 이 범주를 벗어나지 못하고 있다.

저출산에 대한 관심이 커지기 시작한 2000년 이후부터 저출산 현상이 장기화됨에 따라 저출산에 관한 언론과 학계에 연구들 또한 오랫동안 진행되고 쌓여왔다. 그중에서 저출산에 대한 논의는 저출산 문제의 심각성 및 인구 고령화 문제와 결부되어 선거철, 그리고 저출산·고령사회 기본계획이 발표된 시점을 중심으로 집중적으로 등장하였다.

김연권은 2003년부터 2018년까지 조선일보, 한겨레, 중앙일보, 경향신문의 4개의 신문사에 제시된 저출산 관련 단어들을 중심으로 담론 분석을 하였으며, 저자는 매체의 진보 혹은 보수라는 이념적 성격에 따라 저출산 문제에 대한 서술 방식이 확연하게 다름을 지적하였다[1]. 이선정, 반현은 저출산이라는 공동의 사안에 대해 정부자료와 언론 보도에서 사용되는 프레임이 서로 어떻게 다른지 살폈다[14]. 연구 결과, 저출산 이슈에 대한 현실을 구성함에 있어 정부와 언론의 시각차는 뚜렷하게 존재하는 것으로 나타났다. 성정현, 홍석준은 엠파스 닷컴에서 2005년부터 2007년 사이에 수집된 저출산 관련 기사문을 분석하였다[15]. 이들은 출산의 주체인 여성 개개인의 선택 및 의사결정에 대한 논의보다 출산 그 자체를 국가가 도구적으로 해결하고자 하는 담론이 우세한 점을 지적하였다.

남정은, 정정희는 1985년부터 2010년 사이에 조선일보와 동아일보에 나타나는 기사문 분석을 통해 저출산의 원인으로 기술되어온 자녀에 대한 가치관 변화, 공동육아 미흡 등의 현상들이 실제로 저출산의 원인이 되는지 내용분석을 통해 확인하였다[16].

또한, 최영미, 박윤환은 저출산 원인들을 세부적으로 분해하고 공통의 특성을 뽑아내 묶어내는 유형화를 시도하였고, 총 8개의 유형이 도출되었고 이를 FGI를 통해 검증을 수행하였다[11]. 연구 결과 결혼과 출산이 자발적인 거부보다 이를 거부하게 만드는 사회 구조적인 원인을 다각적으로 파악해야 하는 것이 필요함을 시사하였다. 박언하는 서울·경기지역의 보육 및 양육 관련 전문가 9인을 대상으로 심층면접을 실시 하였는데, 연구 결과 저출산을 개선하기 위해서는 육아가 더 이상 엄마 및 가정에 한정된 문제가 아니라는 의식의 변화가 필요함을 시사하였다[17]. 정성호는 “우리나라의 저출산 현상은 아이를 안 낳는 데서 출발한 것이 아니라 아이를 낳을 수 없게 만드는 환경에서 비롯된 것으로, 이를 개선하지 않고 서는 상황이 나아지기를 기대하기 어렵다”는 주장에 매우 귀 기울일 필요가 있는 시점이라고 하였다[18].

조형숙, 조현정은 저출산 관련 연구 동향을 밝혔는데, 저출산 해결은 전 국민 의식 변화에서 찾을 수 있으며, 추후 연구에서는 다양한 계층의 소리를 듣고 사회와의 합의점이 진행돼야 한다고 지적하였다[19].

이와 같은 선행연구에서, 개별심층면담 또는 FGI를 통해 저출산의 원인과 극복안을 연구하거나 문헌 연구를 통해 비판적 독해를 하는 방식을 따랐다. 또한, 특정 신문사에 제시된 기사문을 분석하는 연구를 하거나 기사문에 대한 내용을 분석하는 방법 등을 따랐지만, 대중의 다양한 목소리와 관점을 충분히 반영하지 못하는 한계가 있다. 본 연구에서는 기존 연구에서 간과된 대중의 목소리를 포착하고, 저출산 문제에 대한 다양한 관점과 경험을 포괄적으로 이해하려 한다.

2. 텍스트 마이닝

텍스트 마이닝은 방대한 양의 비정형 데이터에서 의미 있는 정보나 주제를 확률적으로 분석하는 기술이다. 이를 통해 정보추출과 요약 작업을 자동화하여 인간의 수작업에 비해 훨씬 빠르고 효율적으로 비정형 데이터를 처리할 수 있다. 구조화되지 않은 대규모의 텍스트 집단으로부터 새로운 지식을 발견하는 과정을 거치며[20], 텍스트 마이닝 기술을 통해 거대한 텍스트 뭉치에서 의미 있는 정보들을 추출해, 연관성을 파악하며 키워드 빈도분석, 감성 분석, LDA 토픽 모델링에 이르기까지 다양한 분석기법을 적용하여 잠재된 의미를 도출하고 각각의 텍스트들이 가진 범주를 찾아내는 등의 단순한 정보 탐색 이상의 결과를 얻어낼 수 있는 것이 핵심이다[21].

텍스트 마이닝 기법은 연구자의 주관적 해석이라는 기존의 질적 분석에서 나타나는 단점을 보완하고, 통계를 바탕으로 하는 양적 분석을 통해 신뢰도와 타당도를 높인다는 점에서 주목을 받고 있다[22]. 온라인상에 있는 방대한 비정형 데이터를 실시간으로 수집하고 분석하여 고객의 감성이나 의도 등을 분석하는 데 많이 활용되고 있다. 또한, 정확한 파악이 힘든 사회 현상 및 이슈와 관련된 다양한 논의를 파악하는데 유용하게 사용되고 있다[23]. 즉 텍스트 마이닝은 다양한 분석기법을 통해 문서 속에 다양한 주제를 도출하고, 다른 축적된 문서들을 종합하여 새로운 의미를 발견한다는 점에서 비정형 데이터를 객관화하여 효율적으로 분석한다는 장점이 있다.

본 연구는 대량의 댓글을 토대로 저출산에 관한 논의

들을 비교·분석하기 위해서 텍스트마이닝 기법을 활용하여 온라인 커뮤니티에서 대중들이 많이 언급하는 저출산의 다양한 요인에 대해서 도출하고 이에 대한 대중들의 감성 및 경험에 대해 활용하여, 보다 종합적으로 대중들의 의견에 대해 분석하고자 한다.

Ⅲ. 텍스트마이닝 저출산 요인분석 방법

1. 분석대상 선정 및 자료수집

로이터 연구소의 디지털 뉴스 보고서 2018년에 따르면, 네이버는 플랫폼 점유율 65%로 국내 최대 포털사이트이다[24]. 이에 본 연구에서는 저출산의 요인을 분석하고자 국내 최대 온라인 포털사이트인 네이버 제휴 뉴스 기사 속 댓글만을 분석 자료로 활용하였다.

네이버에서 기간을 2022년 3월 19일부터 2023년 8월 24일까지로 설정하고, 네이버 제휴 기사들을 대상으로 BeautifulSoup과 Selenium을 활용하여 웹 스크래핑을 시행하여 총 2360개의 뉴스 기사를 수집하였다. 네이버 제휴 기사의 URL을 먼저 추출 및 수집한 뒤 이어서 본문에 '저출산' 및 '출산율'을 포함하는 기사 내용 아래에 달린 댓글을 추출 및 수집하였으며, 총 166,120건의 댓글들이 수집되었다[25].

2. 데이터 전처리

전처리 과정은 컴퓨터가 처리하기 쉽도록 비정형 데이터를 인식할 수 있는 인공어로 변환하는 작업이다[26]. 데이터 분석 또는 기계학습 모델에 텍스트 데이터를 입력하기 전에 꼭 필요한 중요한 단계 중 하나이므로 수집한 166,120건의 댓글들을 분석하기에 앞서 댓글 데이터를 분석에 적합한 데이터로 만들 수 있도록 전처리를 진행하였으며, [그림 1]과 같다.


수집 된 데이터 전처리 후

그림 1. 수집된 데이터 및 전처리

전처리 과정으로는 띄어쓰기를 기준으로 어절을 구분하고, 명사, 동사, 형용사 단위로 키워드를 구분하였다. 또한, 추출된 목록과 최초 수집된 목록을 비교하며 단어를 정제하는 작업을 반복적으로 수행하였다.

단어 정제작업에서는 다음과 같은 기준을 세워 진행하였다. 첫째, 추출한 텍스트는 언급된 데이터 전처리 과정을 거쳤다. 둘째, 축약어, 어근, 유의어, 지시어를 통일하여 정제하였다. KoNLP 패키지의 *extractNoun()* 함수로 명사를 추출하였다. 명사 추출 및 추출어 확인을 반복 시행하여 전처리의 완성도를 높여 연구의 신뢰도를 높이고자 하였다. 셋째, TF-IDF(Term Frequency-Inverse Document Frequency)를 사용하여 각 댓글을 독립된 문서로 간주하여 단어의 중요성을 평가하는 방법으로 신뢰도를 높이고자 하였다. TF-IDF 방법은 빈도분석보다 더욱 정교한 방법으로, 특정문서에서 등장하는 특정단어 빈도가 많으면서 동시에 그 단어가 다른 문서에서는 출현 빈도가 적을 때에 그 단어를 특정 문서의 핵심어로 간주한다. 특정단어가 많이 등장하는 어휘빈도와 다른 문서에서 등장하지 않는 문서의 역빈도 곱함을 통해 핵심어를 추출하며[27], TF-IDF 수식은 [그림 2]와 같다.

$$TFIDF(w,d) = TF(w,d) \times \log\left(\frac{N}{DF(w)}\right)$$

*$TF(w,d)$: 문서 d에 단어w가 나타난 횟수
*$DF(w)$: 단어 w가 들어가는 문서의 총수
*N: 전체문서의 총수

그림 2. TF-IDF 수식

3. 데이터 분석

본 연구에서는 키워드 빈도분석, 감성 분석 그리고 LDA 토픽 모델링을 실행하였으며, 적용한 텍스트 마이닝 접근의 분석 절차는 [그림 3]과 같다.

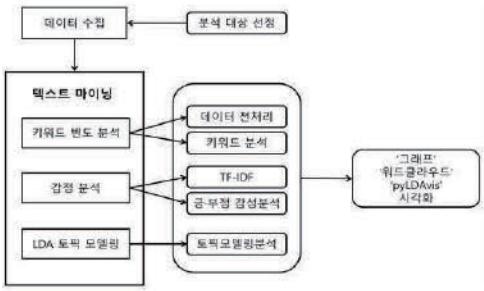

그림 3. 연구의 텍스트마이닝 분석 절차

키워드란 텍스트 자료의 중요 내용을 압축적으로 제시하는 단어 또는 문구를 말한다. 자연어 처리를 시행한 후에 단어의 등장 빈도를 분석함으로써 핵심어를 추출하는 것으로, 텍스트 마이닝의 가장 기본적인 방법이다[28]. 자주 등장하는 단어들을 파악하여 주요 주제나 관심사를 도출하여, 전반적인 흐름과 주요 이슈를 빠르게 파악할 수 있는 장점이 있어 실행하였다.

감성 분석은 텍스트에 등장하는 단어 표현 등을 통해 어떤 감성이나 태도 및 의견이 드러나는가를 알아보는 기법으로, 작성자의 감성의 종류와 정도 및 감성이 일어난 대상에 대하여 분석할 수 있다[29]. 본 연구에서는 GitHub에 공개된 데이터 셋을 다운로드하여 훈련에 활용하였다. KoNLPY의 *Okt()*를 활용해 토큰화하고, *TfidfVectorizer()*를 이용해 TF-IDF 벡터화를 수행했다. 긍정과 부정의 감성을 이진 분류하기 위해 로지스틱 회귀 알고리즘을 사용했으며, *GridSearchCV()*로 하이퍼 매개변수 C의 최적값을 찾은 후, 최적의 분석 모델을 훈련 시켰다. 분석 모델을 사용하여 댓글의 긍정, 부정을 분류하였으며, 추가로, 한 단어가 긍정/부정 문에서 활용되는 비율을 계산하고 이를 다시 TF-IDF 값에 곱하는 방식으로 단어의 중요도를 보정하였다. 감성 분석은 빈도가 높은 단어 10개 순으로 바 차트를 통해 시각화하였다.

LDA(Latent Dirichlet Allocation)는 비지도 학습 알고리즘으로 대량의 비구조적 문서에서 단어 간에 관련성에 따라 토픽별로 분류하는 확률적 토픽모델링 알고리즘이다[30]. 미리 알고 있는 주제별 단어 수 분포를 갖고 주어진 문서에서 발견된 단어 수 분포를 분석함으로써 해당 문서에 어떤 주제들과 함께 다루고 있을지 예측할 수 있다[31]. 이를 통해 특정 주제가 얼마나 자주 등장하는지, 그 주제가 다른 주제들과 어떻게 연관되는지를 분석할 수 있기에 LDA를 실행하였다.

Ⅳ. 텍스트 마이닝 저출산 요인분석

1. 키워드 빈도분석

본 연구 분석에 사용된 네이버 제휴 뉴스 기사에 달린 댓글의 특징은 다음과 같다. 총 수집된 댓글 수는 166,120건이었으며 그 중, 컴퓨터가 처리하기 쉽도록 비정형 데이터를 인식할 수 있는 인공어로 변환하는 작업을 거친 147,174개의 댓글을 분석들에 사용하였다. 빈도순으로 상위 30개의 단어는 [표 1]과 같다.

표 1. 키워드 빈도분석 상위 30개

	상위 키워드				
1	결혼(18943)	11	사회(7513)	21	자식(5748)
2	아이(16852)	12	부모(7402)	22	국민(5666)
3	사람(15166)	13	국가(6750)	23	자녀(5550)
4	문제(12182)	14	저출산(6682)	24	거지(4666)
5	생각(11564)	15	여성(6440)	25	혜택(4260)
6	여자(9893)	16	집값(6325)	26	육아(4164)
7	정책(9736)	17	교육(6141)	27	소리(4019)
8	지원(8368)	18	정부(6133)	28	이상(3949)
9	인구(7979)	19	해결(5952)	29	현실(3466)
10	남자(7745)	20	세금(5882)	30	대책(3433)

2. 감성 분석

감성 분석으로 부정과 긍정에서 나타난 단어를 단어별 출현 빈도를 [그림 4]와 같이 정리하였다.

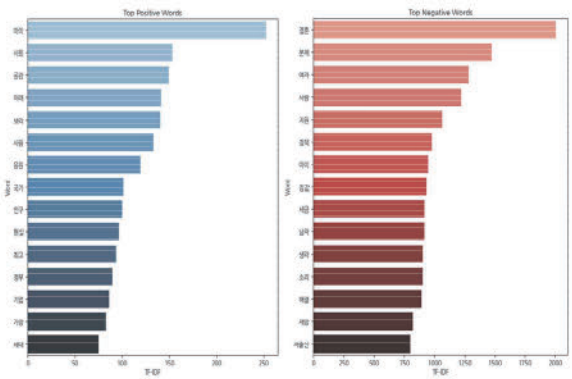

그림 4. 긍정과 부정 댓글의 단어 상위 10개

분석의 과정은 명사 단어를 추출한 후, TF-IDF 값이 높은 순으로 정리한 후에 단어별 합을 구하여, 바 차트로 시각화하였다. 긍정의 상위 10개 단어는 아이, 사회, 공감, 미래, 생각, 사람, 응원 국가, 인구, 현실, 최고, 정부, 기업, 가장, 세대. 부정의 상위 10개 단어는 결혼, 문제, 여자, 사람, 지원, 정책, 아이, 집값, 세금, 남자, 생각, 소리, 해결, 재앙, 저출산의 순서로 출력되었다.

3. LDA 토픽 모델링 분석

LDA 토픽 모델링에서는 만들어진 언어 모델을 가지고 모델 평가 및 최적화된 모델링을 위해 분류하는 토픽의 개수를 일관성 점수를 출력하여 지정해주는 것이 중요하다. 이때 토픽 일관성 점수값이 높을수록 의미론적으로 일관성이 높다고 해석된다. 토픽의 일관성 점수를 출력하여 토픽의 개수를 선정한 결과, 9개일 경우 데이터 내의 다양한 의견과 주제들을 잘 반영하고 있으며, 높은 일관성을 유지하였다. 모델해석 가능성, 유용성을 극대화하는 좋은 균형점을 제공하기에 본 논문에서는 총 9개의 토픽을 형성하였다.

토픽의 갯수를 정한 후에는 각 토픽에 핵심 단어를 선정하여 그에 맞는 토픽명을 지정하였으며, [표 2]와 같다.

표 2. LDA 토픽 구성 단어 목록과 토픽명

	토픽명	핵심 구성 단어	비중(%)
1	여론과 대중 심리	선동, 기사, 댓글, 수준, 관심, 정신, 마인드, 생각	8.0
2	사회 계층 불평등	기득권, 정치, 정치인, 부자, 서민, 거지, 노예, 이민	9.1
3	가족생활의 경제적 부담	결혼, 일자리, 육아, 도우미, 가사, 임금, 취업, 수도권	9.4
4	젠더간 갈등	군대, 국방, 페미, 남녀, 갈등, 비혼, 연애, 대책	10.9
5	세계 인구변화	인구, 중국, 일본, 한국, 감소, 소멸, 세계, 걱정	8.9
6	가족생활의 균형	근무, 집안일, 월급, 독박, 난임, 기관, 회사	6.9
7	사회적 지원	육아, 교육, 지원, 수당, 어린이집, 맞벌이, 임신, 현실, 혜택	18.0
8	주거 및 경제적 부담	집값, 아파트, 주거, 소득, 대출, 물가, 교육비, 연금	11.2
9	출산에 대한 인식변화	세대, 문화, 시대, 혼자, 자식, 인생, 인식, 환경, 본인	17.6

V. 논의

1. 분석 결과 논의

키워드 빈도분석에서 상위 30위 내에 있었던 단어 가운데 '결혼', '아이', '정책', '지원', '인구', '저출산', '교육', '정부', '세금', '자녀', '대책' 등은 사회적, 인구 문제와 관련이 있으며, '사람', '문제', '생각', '여자', '남자' 같은 단어는 개인적 관점이나 성별 문제를 반영할 수 있다. 특히 '결혼', '아이', '교육', '집값', '정책', '지원' 등은 결혼과 출산에 대한 경제적 부담감이 큰 상태임을 알 수 있다.

감성 분석을 통해 얻은 긍정과 부정의 감성에 대한 각 상위 10개의 단어에서는 저출산 문제를 바라보는 대중 시각의 복잡성과 다양성을 볼 수 있었는데 긍정적인 단어들은 대체로 희망, 지지, 연대감과 같은 긍정적인 감정이나 반응을 나타내는 반면, 부정적인 단어들은 저출산 문제의 원인이나 그에 대한 불만과 같은 우려와 부정적인 시각을 드러냈다. 긍정의 댓글 안에서는 미래 세대를 지원하고, 같이 힘을 합쳐 나아지기를 응원하는 모습을 볼 수 있으며, 부정의 댓글에서는 아이를 위한 실질적인 지원이 출산율 증가에 중요하다고 바라보고 있음을 알 수 있다.

현재 저출산 현상은 우리 사회에 있어 심각한 도전과제이다. 긍정적인 댓글들에서 드러난 바람과 기대를 통해 희망을 볼 수 있듯이, 부정적인 목소리도 중요하지만, 긍정적인 관점 또한 사회가 직면한 문제에 대한 해결책을 모색하고 발전시키는 것에 더욱 중요한 도움이 될 수 있음을 말하고자 한다.

LDA 토픽 모델링을 통해 [표 2]의 9개 토픽을 볼 수 있었다. Topic 1의 핵심 단어 구성을 통해 사회적 이슈가 미디어를 통한 여론 형성, 그리고 대중의 심리적 반응과 태도와 어떻게 상호작용하는지를 보여준다. Topic 2의 핵심 단어 구성을 통해 대중들이 겪고 있는 겪고 있는 사회 계층에 대해 불평등한 생각을 지닌 것을 보여준다. Topic 3의 핵심 단어 구성을 통해 한 가족이 출산을 결정하는 것에 있어서 사회적 지원 체계의 중요성이 부각되는 것을 볼 수 있다. Topic 4의 핵심 단어 구성을 통해 현대 사회에서 여성과 남성의 사회적 역할과 권리에 따라 다양한 반응과 논란이 있다는 것을 볼 수 있다. Topic 5의 핵심 단어 구성을 통해 세계에서 발생하는 저출산 및 인구감소 현상 또한 대중들의 관심이 있다는 것을 알 수 있다. Topic 6의 핵심 단어 구성을 통해 광범위한 맥락 속에서 이해되어야 함을 보여준다. 이는 고용 조건, 가사 노동의 분담, 출산의 생리적 요인, 그리고 이러한 문제에 대응하는 정책 및 기관

의 역할에 대해 말하는 것을 알 수 있다. 특히 출산의 생리적 요인은 부분에서는 난임과 관련한 댓글들이 많이 분포되어 있었던 것으로 보아 난임에 대한 적극적인 도움이 필요할 것으로 보인다. Topic 7의 핵심 단어 구성을 통해 시간과 자원을 많이 요구하는 육아에서 맞벌이를 하는 부부에게 어린이집과 같은 외부 돌본 시설의 중요성이 부각되는 것을 알 수 있다. Topic 8의 핵심 단어 구성을 통해 출산과 양육의 직접적인 비용을 나타낸다. 교육비 부담은 특히 한국과 같은 교육 중시 사회에서 중요한 고려사항이며, 이는 부모가 자녀 수를 결정하는 데 중대한 요인으로 작용하는 것으로 볼 수 있다. Topic 9의 핵심 단어 구성을 통해 개인의 삶과 사회적 조건 속에서의 선택 그리고 시대적 변화를 모두 고려해야 함을 알 수 있다.

2. LDA 토픽 모델링과 선행연구 비교

본 연구에서는 LDA 토픽 모델링으로 도출한 것에서 더 나아가 토픽들과 선행연구에서 FGI를 통해 도출된 주제를 비교하여 저출산 문제에 대한 깊이 있는 이해와 비교를 해보고자 한다. 비교한 내용은 [표 3]과 같다.

표 3. LDA 토픽 주제와 FGI 주제간의 비교

최영미외, 2019 (FGI, 38명)		본 연구
의지적 저출산	"개인이 스스로의 이익 추구에 따라 출산을 거부함으로써 출산 저하를 가져온 것"	Topic 9
숙명적 저출산	"타인이나 사회 공동체 이익에 스스로를 희생당하며 비출산을 받아들임으로써 출산 저하를 가져온 것"	Topic 9
사회적 저출산	"전 사회적 인식 변화로 인하여 사회질서가 심각하게 재적응해야 하는 상황에서 기존의 집단적 질서가 흔들리며 출산 저하를 가져온 것"	Topic 9
경제적 저출산	"국가경제에 대한 비관적 전망이나 결혼 생활을 유지하기 위한 경제적 비용 부담으로 인하여 출산 저하를 가져온 것"	Topic 3,8
정책적 저출산	"정부의 저출산 정책의 부재나 한계 혹은 정책 그 자체가 가져온 문제점으로 인하여 오히려 출산 저하를 가져온 것"	Topic 7
이타적 저출산	"인류와 세계의 복리증진을 위하여 출산을 하지 않는 것이 바람직하다는 인식하에 나타나는 비혼 및 비출산"	Topic 5
물리적 저출산	"출산 가능 인구의 신체적 어려움이나 물리적인 출산저하 요소들로 인하여 출산 가능성이 떨어져서 출산저하를 가져온 것"	Topic 6
차별적 저출산	"사회가 비혼에 의한 임신이나 장애인에 대한 출산을 암묵적으로 가로막음으로써 출산저하를 가져온 것"	-

먼저 유사한 주제들을 비교해보았을 때 Topic 3과 Topic 8은 경제적 저출산이라는 주제와 유사하였다. Topic 5는 이타적 저출산이라는 주제와 유사하였다. Topic 6은 물리적 저출산이라는 주제와 유사하였다. Topic 7은 정책적 저출산이라는 주제와 유사하였다. Topic 9는 의지적 저출산, 숙명적 저출산, 사회적 저출산이라는 주제들과 유사한 것을 볼 수 있다.

[표 3]의 본 연구와 기존의 연구가 서로 다른 데이터와 분석 방법을 사용했음에도 불구하고 유사한 결과를 도출했다는 것은, 이러한 주제들이 저출산 원인에 있어 핵심적이고 보편적인 요소임을 강력하게 시사할 수 있으며, 해당하는 주제들의 신뢰성을 높일 수 있다. 꾸준히 언급되고 있었던 이와 같은 주제들의 논의는 아직도 끊이지 않고 진행됨에 있어 해결되지 못한 부분으로 볼 수 있기에 더 자세히 다뤄야 한다는 것을 알 수 있다.

다음으로는 본 연구에서만 나타난 Topic 1, Topic 2, Topic 4에 대해 분석하고자 한다.

본 연구의 Topic 4는 차별적 저출산과 유사해 보이지만 선행연구에서의 차별적 저출산에서는 가족 구성원의 다양함을 인정받지 못하는 부분에 대하여 언급하였고, Topic 4에서는 젠더간 갈등에 대한 언급이 많은 것을 볼 수 있었으므로 다른 시각으로 바라보고 있음에 따라 유사하지 않은 토픽임을 알 수 있다. 이는 젠더간 갈등에 대해 어떠한 현상이 일어나고 있는지에 대해 먼저 이해를 해야 해결책을 찾을 수 있는 실마리가 보일 것으로 생각한다. 또한, 가족 내에 균형 잡힌 역할 분담을 장려하여 가족 구성원 모두가 동참하여 육아할 수 있도록 개선하는 것이 중요한 것임을 알 수 있다. Topic 1은 미디어를 통한 여론 형성에 따른 저출산에 대한 대중의 심리적 반응이 일어나고 있다는 것을 알 수 있다. 이는 새롭게 나타난 주제로 사회적 분위기에 따른 출산에 대한 인식이 바뀐다면, 여론을 통해 긍정적인 방향으로 충분히 이끄는 것을 노력하여 분위기를 만들어 출산율을 높일 수도 있음을 바라볼 수 있다. Topic 2는 현재 사회에서 발생하고 있는 계층 간 불평등 문제에 대해 다루는 것을 알 수 있다. 점점 심해지는 계층 간 경제적 격차 혹은 교육 및 취업 기회의 불균형 등에 대해 보다 평등한 정책이 절실히 필요한 순간임을 알 수 있다.

이러한 본 연구에서만 나온 새로운 주제들을 통해 최근 사회에서 저출산과 관련해서 활발히 논의되는 의견을 텍스트 마이닝 분석방법을 사용하여, 업데이트함에 있어 현 시점의 대중들이 바라보는 시각을 다시 알아볼 수 있다. 한국 사회가 직면한 현재의 저출산 문제는 단순한 현상이 아니다. LDA 토픽 모델링을 통한 대중의 댓글 분석은 복잡한 문제를 다각적으로 이해하는 데 도움이 될 수 있다. 사회, 경제, 문화 등 여러 분야에서 얽힌 문제들이 어떻게 상호작용하며 저출산 현상에 영향을 미치는지를 볼 수 있으며, 다양한 관점에서 바라볼 수 있는 기회를 제공한다. 현재 한국사회는 저출산의 명확한 원인과 해결책을 찾기 위해 모든 사회 구성원이 함께 노력하고 있다. 방대한 데이터에서 추출된 정보는 저출산 문제의 다양한 측면을 조명하고, 이에 대한 종합적인 접근 방식을 모색하는 데 기여할 수 있다. 따라서, 연구자들과 정책수립자들은 이러한 데이터를 근거로 다양한 요인을 고려한 종합적인 해결책을 모색해야 한다. 결론적으로, 텍스트 마이닝은 저출산 문제에 대한 대중의 인식을 파악하고, 이를 통해 보다 심층적이고 포괄적인 해결방안을 찾는 데에 중요한 역할을 할 수 있다.

VI. 결론

본 논문은 저출산 문제에 대한 기존의 접근 방식과 달리 대량의 데이터를 통한 실증분석을 도전하며, 데이터마이닝, 기계학습 등의 인공지능 관련의 분석방법을 적용하고자 하였다. 실증분석은 데이터에 근거하여 결론을 도출해내는 것으로 이는 연구 결과에 대한 신뢰도 또한 높였다. 이러한 결과들은 대중의 의견이 담긴 데이터가 저출산 문제에 대해서 이해하고, 이를 바탕으로 종합적인 해결책을 모색하는 데 있어 도움이 될 것으로 본다. 저출산 문제 해결에 도움이 되기 위해서 텍스트 마이닝은 대중의 인식을 파악하는 동시에, 보다 심층적이고 포괄적인 정책 방안을 도출하는 데 있어 중요한 역할을 했으며, 다면적인 전략을 수립하는 데에 중요한 기초 자료를 제공할 수 있다고 생각한다.

본 연구에서는 학문적 시사점으로는, 본 연구가 최신 대규모 온라인 데이터를 기반으로 실증분석을 함으로써 저출산 문제에 대한 대중의 사회적 인식과 복잡성을 이해하는 데 기여했다는 점에 있다. 이는 향후 저출산 연구에 있어서 인구학적 접근뿐만 아니라, 사회학적, 심리학적, 경제학적 접근을 포함한 다학제적 연구의 중요성을 강조할 수 있다. 대중의 의견을 반영한 데이터의 텍스트 마이닝을 통해, 저출산 문제에 대한 현재 대중 시각에서의 이해가 실질적인 해결책 모색에 기여할 수 있음을 보여준다. 이는 텍스트 마이닝의 분석방법이 사회문제 연구에서도 중요한 도구로 자리 잡을 수 있음을 시사한다.

실무적 시사점으로는, 연구 결과를 바탕으로 정책수립과 실행과정에 현실적인 문제에 해결점을 제공한다는 점에 있다. 단순히 인구수를 늘리는 것을 넘어서, 사회 구조와 문화적 가치관의 변화를 포함해야 한다. 출산율 증진을 위해서는 단순한 재정적 지원을 넘어서, 일과 생활의 균형, 가족 친화적인 사회 환경 조성, 여성의 경력 단절 문제 해결 등 보다 광범위한 사회적 변화를 목표로 해야 한다. 결국, 저출산 문제의 해결은 한국사회가 직면한 가장 시급한 과제 중 하나로, 정부의 강한 의지와 모든 사회 구성원의 적극적인 참여와 협력이 필요하다는 것이다. 사회적 인식의 변화 촉진을 통하여 저출산 문제에 대한 대중의 인식과 가치관의 변화를 이끌어내는 것이 중요하다. 이에 대중의 목소리가 반영된 앞선 키워드 빈도분석과 감성 분석 결과 그리고 LDA 토픽 모델링 결과를 바탕으로 결합한 것을 참고한다면, 앞으로 우리 사회의 저출산 문제를 해결하는 진행 방향에 대해 큰 도움을 줄 것이라 본다.

VII. 연구 한계 및 향후 연구 방향

본 연구는 텍스트 마이닝을 통해 대규모 데이터에서 유의미한 인사이트를 도출하였지만, 텍스트 마이닝의 다양한 모델과 알고리즘을 적용하지 못한 한계가 있다.

또한, 네이버 뉴스 기사 속 댓글만을 활용한 것에 한계가 있다. 웹사이트 경우 각 사이트가 가지고 있는 성향과 그에 따른 사용자의 성향도 다를 수 있다. 본 연구에서는 플랫폼 점유율 65%로 국내 최대 포털사이트인

네이버를 사용하였지만, 다양한 제휴 사이트 혹은 다른 포털사이트를 활용하지 못하였다. 따라서 향후 연구에서는 네이버 속 다양한 사이트들 혹은 다른 포털사이트를 대상으로 하여 분석하는 것을 통해, 보다 다양한 대중들의 시각을 얻을 수 있을 것으로 본다. 더 나아가 향후 연구에서는 본 연구의 방법론을 다른 국가의 저출산 문제에 적용하여, 국가 간 비교 연구를 수행함으로써, 저출산 문제의 보편적인 요인과 특수한 요인을 분리해 분석하는 것도 중요한 연구 방향이 될 것이다. 이를 통해 각국의 사회적, 경제적, 문화적 맥락에 맞는 맞춤형 해결책을 제시할 수 있는 기반을 마련할 수 있을 것으로 생각한다.

마지막으로 연구 결과를 바탕으로 실정책에 반영될 수 있는 방안을 모색하고, 정책 시뮬레이션 연구를 하는 것도 좋은 연구 방향이 될 수 있다. 이러한 연구 방향은 저출산에 대해 보다 심도 있는 이해를 가능하게 하고, 실질적인 해결책을 제시하는 데 기여할 것이다. 또한, 이는 학문적 연구 뿐만 아니라 실질적인 정책 수립에 있어서도 중요한 지침이 될 것이며, 저출산 문제를 해결하기 위한 국가적, 국제적 노력에 중요한 역할을 할 것이라 본다.

참 고 문 헌

[1] 김연권, "'저출산'에 대한 신문 담론 분석," 시민인문학, 제36권, pp.43-100, 2019.

[2] 통계청, *KOSIS 인구동향조사*, 2023.

[3] http://www.wlv.kr/news/articleView.html?idxno=3000872

[4] 김수정, "저출산· 고령화 시대 한국의 인구 정책에 관한 비판적 고찰," 한국도시지리학회지, 제22권, 제2호, pp.143-158, 2019.

[5] 이상협, 이철희, 홍석철, *저출산 대책의 효과성 평가*, 한국보건사회연구원, 2016.

[6] 김태준, *국민의 삶을 책임지는 교육정책과 사회정책의 균형과 통합을 위한 과제*, 한국교육개발원 경제 · 인문사회연구회, 2021.

[7] 김천권, 정진원, "한국사회의 저출산 현상, 재앙인가 기회인가?," 국가정책연구, 제33권, 제3호, pp.1-4, 2019.

[8] 이현옥, "한국여성의 출산의지 결정요인," 정책개발연구, 제11권, 제1호, pp.99-132, 2011.

[9] G. Browning, Abigail Halcli, and F. Webster, *Understanding Contemporary Society: Theories of the Present*, London: SAGE Publications, 2000.

[10] 이용복, *보육서비스가 기혼여성의 출산의도와 출산결정에 미치는 영향*, 숙명여자대학교, 박사학위논문, 2004.

[11] 최영미, 박윤환, "결혼 및 출산에 대한 인식변화 분석과 저출산 원인의 유형화," 시민인문학, 제36권, pp.101-137, 2019.

[12] 정성호, "저출산 대책의 패러다임 전환에 대한 비판적 검토," 공공사회연구, 제8권, 제2호, pp.36-64, 2018.

[13] T. J. Espenshade, J. C. Guzman, and C. F. Westoff, "The surprising global variation in replacement fertility," Population Research and Policy Review, Vol.22, No.5/6, pp.575-583, 2003.

[14] 이선정, 반현, "저출산 이슈에 대한 정부 보도자료와 언론 보도의 프레임 비교," 광고PR실학연구, 제10권, 제4호, pp.161-190, 2017.

[15] 성정현, 홍석준, "인터넷 자료를 통해 본 한국에서의 저출산 담론의 특징과 의미: 정체성의 정치학(politics of identity) 담론 극복을 위한 제언," 복지와 문화다양성연구, 제3권, pp.29-53, 2009.

[16] 남정은, 정정희, "자녀양육 양상을 통해 본 저출산 문제의 이해: 1980년대 중반~2000년대 신문기사 분석을 중심으로," 한국유아교육학회, 제33권, 제2호, pp.53-78, 2013.

[17] 박언하, "저출산사회 대응을 위한 아동수당정책에 관한 질적 연구," 인문사회 21, 제11권, 제2호, pp.901-914, 2020.

[18] 정성호, "저출산 대책의 패러다임 전환에 대한 비판적 검토," 공공사회연구, 제8권, 제2호, pp.36-64, 2018.

[19] 조형숙, 조현정, "국내 저출산 관련 연구 동향 분석: 2011~2020년 중심으로," 유아교육학논집, 제25권, 제3호, pp.253-277, 2021.

[20] 김수연, 정영미, "텍스트 마이닝 기법을 이용한 연관 용어 선정에 관한 실험적 연구," 한국정보관리학회, 제23권, 제3호, pp.147-166, 2006.

[21] 조성우, *Big Data 시대의 기술*, KT 종합기술원, 2011.
[22] Y. Zhang, X. Ran, C. Luo, Y. Gao, Y. Zhao, and Q. Shuai, "Only visible for three days: Mining microblogs to understand reasons for using the Time Limit setting on WeChat Moments," Computers in Human Behavior, Vol.134, p.107316, 2022.
[23] M. Böhmecke-Schwafert and E. G. Moreno, "Exploring blockchain-based innovations for economic and sustainable development in the global south: A mixed-method approach based on web mining and topic modeling," Technological Forecasting and Social Change, Vol.191, p.122446, 2023.
[24] https://redefault/files/digital-news-report-2018.pdf
[25] 남가영, *텍스트마이닝을 활용한 저출산 요인 분석 네이버 뉴스 기사 댓글 중심으로*, 전남대학교, 석사학위논문, 2024.
[26] 이상훈, 조장식, 강창완과 최승배, "텍스트 마이닝을 활용한 영화흥행 예측 연구," 한국데이터정보과학회지, 제26권, 제6호, pp.1259-1269, 2015.
[27] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, Vol.24, No.5, pp.513-523, 1988.
[28] 윤태일, 이수안, *파이썬으로 텍스트 분석하기*, 서울: 늘봄, 2018.
[29] H. Wang, D. Zhang, and C. Zhai, "Structural topic model for latent topical structure analysis," Computational Linguistics: Human Language Technologies, pp.1526-1535, 2011.
[30] D. M. Blei, "Probabilistic topic models," Communications of the ACM, Vol.55, No.4, pp.77-84, 2012
[31] 송민, *텍스트 마이닝*, 청람, 2017.

저 자 소 개

남 가 영(Ga-Young Nam) 정회원

- 2022년 2월 : 전남대학교 전자상거래전공 학사
- 2024년 2월 : 전남대학교 디지털미래융합서비스협동과정 석사
- 2024년 3월 ~ 현재 : 전남대학교 디지털미래융합서비스협동과정 박사과정

〈관심분야〉 : 전자상거래, 디지털융합서비스, 빅데이터 분석, 사회문제 등

김 용 민(Yong-Min Kim) 종신회원

- 2002년 8월 : 전남대학교 대학원 (이학박사)
- 2006년 3월 ~ 현재 : 전남대학교 문화콘텐츠학부 교수, 정보보호융합학과(원) 교수, 디지털미래융합서비스협동과정(원) 교수

〈관심분야〉 : 시스템및네트워크 보안, 전자상거래 보안, 융합보안, 디지털융합서비스 등