| 이슈페이퍼 2019-08 |

빅데이터 분석을 통한 4차 산업혁명시대 육아정책의 방향과 과제

박창현

Korea Institute of Child Care and Education

빅데이터 분석을 통한 4차 산업혁명시대 육아정책의 방향과 과제*

박창현 부연구위원

♦♦♦ 요 약 ♦♦♦

- 4차 산업혁명 시대의 출산 및 양육에 대한 소셜 빅데이터 분석을 통하여 관련 이슈를 알아보고, 저출산 정책 해결에 대한 예측 모형을 구축하여 미래 육아정책의 방향성을 가늠할 필요가 있음.
- 본 연구에서는 '4차 산업혁명', '저출산', '유아교육', '보육', '육아' 키워드 추출 후, 5개 추출 키워드를 이용하여 인터넷 검색 포털 사이트 네이버(Naver)의 블로그, 카페, 뉴스 글 등에서 10년 간(2008-2017) 업로드된 10,675,342건을 수집하였음. '교육(education)', '젠더(gender)', '주거(housing)', '일자리(job)', '노동(labor)'을 추가적으로 추출 후, 해당 주제 키워드에 속한 데이터 세트를 이용하여 소셜 미디어 대상 빅데이터 분석 실시하였음.
- 소셜 빅데이터 분석 결과, 주거 변인은 '출산/보육' 및 일-가정 양립 문제와, 교육 변인은 2017년 이후 저출산 및 사회복지적 측면과 관련성이 높아짐. 일자리 변인은 2015년 이후 고용불안과, 젠더 변인은 육아휴직, 양성평등과 관련성이 높아짐.
- 머신러닝을 통한 주요 변수를 추출하고, 예측모형을 구축하였음. 예측모형에 조출산율 및 합계출산율을 반응변수로 설정하고, 연간 고용률과 가구소득을 설명변수에 추가하는 경우, '주거×교육'과 연간고용률만이 모형상의 유의성을 보였음. '주거×교육'과 연간 고용률을 기반으로 포아송 자동 회귀모형(Poisson Autoregression model)을 활용하여 예측모형을 구축한 결과, 조출산율은 다소 반등한 후 2019년~2020년에 낮은 수준에서 정체될 것으로 예측되었음.
- 이상의 주거와 교육의 동조화 현상을 4차 산업혁명시대와의 관계성을 고려할 때 주거-교육의 공공성을 확보한 소득 및 계층 격차, 주거/교육/일자리/노동/젠더의 측면에서 발생하는 위기요인을 해결, 국민의 삶의 질 제고 등이 필요함.

* 본 원고는 「박창현·김나영·이유진(2018). 4차 산업혁명 시대 육아정책의 이슈와 과제. 육아정책연구소.」의 내용을 토대로 구성함.

1 서론

가. 연구의 필요성 및 목적

- 4차 산업혁명 시대의 급변하는 환경과 함께 초저출산 기저의 우리 사회의 출산과 양육에 대한 새로운 정책적 접근 및 연구 방향의 필요성이 대두됨.
- 본 연구에서는 빅데이터를 활용한 분석을 통하여 기존 방법론의 한계를 보완하고, SNS 참여자들의 인지 구조 속에서 저출산과 양육에 관한 관련 요인들이 어떻게 상호작용하고 연결되는지를 알아보고자 함.
- 본 연구의 목적은 4차 산업혁명 시대의 출산 및 양육에 대한 빅데이터 분석을 통하여 관련 이슈 및 과제를 발굴하고, 저출산 정책 해결에 대한 예측 모형을 구축하여 미래 육아정책을 선도하고 사회적 인식에 기여하는 데 있음.

나. 연구방법

- 소셜 빅데이터 분석
  - 온라인상에 업로드된 4차 산업혁명, 저출산, 출산, 양육, 유아교육, 보육, 육아 관련 비정형 데이터 수집 후 추이 분석, 키워드 분석, 관계망 분석 등을 실시함
- 빅데이터 분석
  - 예측 모형 도출 및 머신 러닝을 통한 주요 요인을 정량화하였음.

2 4차 산업혁명 시대 육아정책에 관한 빅데이터 분석(2008-2017)

가. 소셜 빅데이터 분석 결과

- 수집 데이터 개요
  - ◆ 소셜 데이터 수집 키워드 및 채널

- 국내·외 선행연구 수집 및 검토를 통하여 '4차 산업혁명', '저출산', '유아교육', '보육', '육아' 키워드 추출 후, 5개 추출 키워드를 이용하여 인터넷 검색 포털 사이트 네이버(Naver)의 블로그, 카페, 뉴스 글 등에서 10년 간 업로드된 10,675,342건을 수집함.
- 수집 채널은 국내 이용자 수가 가장 많은 인터넷 검색 포털 사이트 네이버 블로그와 카페, 뉴스 글이었음.

◆ 화제어 추출 알고리즘

- 일차적으로 수집된 데이터에서 데이터 정제 과정(기계적 원문 검토)을 거쳐 원문이 존재하지 않는 987,471건의 노이즈 데이터를 제거한 후, 화제어 1,000위까지 추출한 뒤 형태소 분석과 카테고리 분류 작업을 진행함. 화제어 추출은 Latent Dirichlet Allocation 알고리즘과 TF-IDF(Term Freuquency-Inverse Document Frequency)를 적용하였음.

◆ 데이터 추출 및 정제 과정

- 1차 데이터 정제 과정을 거친 데이터를 대상으로 추출 키워드 포함 여부를 판별 후, 5개 키워드와 연관성을 보이는 데이터를 2차적으로 추출 후 최종 분석에 투입

[그림 1] 데이터 추출 및 정제 과정 개요

분석 주제 키워드 추출 및 카테고리 분류

- 출산 및 양육 현상의 설명력 향상을 위하여, (저)출산 및 양육과 높은 연관을 보이는

5가지 주제 키워드인 ‘교육환경’, ‘젠더문화’, ‘주거문제’, ‘고용문제’, ‘노동환경’을 추가적으로 추출 후, 해당 주제 키워드에 속한 데이터 세트를 이용하여 소셜 미디어 대상 빅데이터 분석 실시

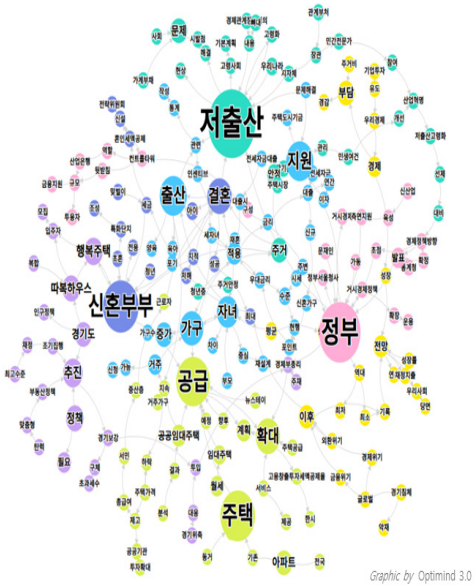
소셜 빅데이터 분석

◆ 주거(housing)

- 화제어 추이 분석 결과, 2016년 이후에는 ‘저출산’, ‘정부’, ‘정책’, ‘결혼’, ‘출산’, ‘일자리’, ‘생활’, ‘서민’의 중요도가 높게 나타났으며 ‘출산/보육’ 카테고리로 분류된 화제어가 20.8%에 달하여 주거 문제가 일-가정 양립 문제와 연결되어 있음을 알 수 있음.
- 2018년~2017년(10년간) 간의 키워드를 대상으로 한 워드 클라우드 분석 결과, ‘저출산’, ‘정부’, ‘공급’, ‘주택’, ‘신혼부부’, ‘출산’이 주거문제와 관련을 보이는 것으로 나타났으며, 관계망 분석을 통해서도 ‘저출산’, ‘정부’, ‘신혼부부’, ‘공급’,‘주택’이 가장 큰 비중을 차지함.

[그림 2] ‘주거’ 키워드 워드 클라우드 (2008년~2017년)

[그림 3] ‘주거’ 키워드 관계망 분석 (2008년~2017년)

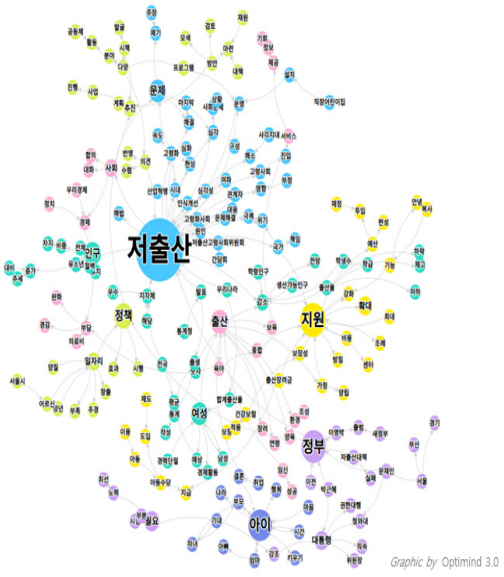


◆ 교육(education)

- 화제어 추이 분석 결과, '교육'은 화제어 순위 및 빈도가 점차 낮아지는 추세를 보인 반면, '저출산' 및 정부/정책 관련 화제어('사회', '지원', '정부', '복지', '정책', '대통령')가 2017년에 높은 순위로 나타나 최근 교육 환경이 저출산 문제 및 정부의 정책 및 사회 복지적 측면과 관련을 보이고 있음을 나타내며, 화제어의 카테고리별 분류 결과, 가구/가족(48.6%), 정부/정책(22.1%), 출산/보육(14.6%) 순으로 분류됨.
- 워드 클라우드 및 관계망 분석 결과, 화제어 추이 분석과 유사하게 '교육'보다는 '저출산', '지원', '정부', '정책', '아이', '출산'이 높은 빈도로 나타났으며, 관계망 분석에서도 '저출산'이 큰 비중을 차지함.

[그림 4] '교육' 키워드 워드 클라우드 (2008년~2017년)

[그림 5] '교육' 키워드 관계망 분석 (2008년~2017년)



◆ 일자리(job)

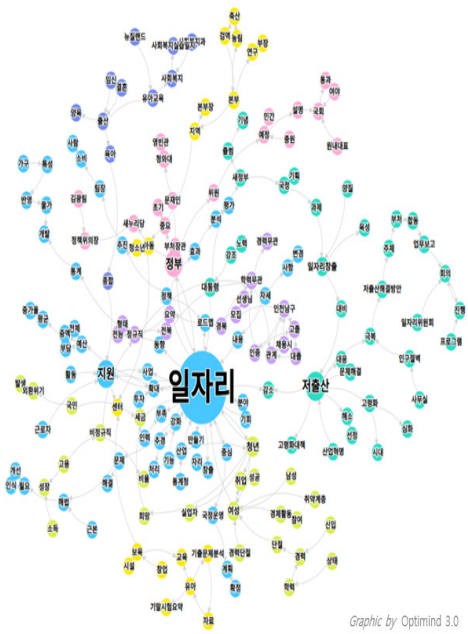
- 화제어 추이 분석 결과, 2010년 이전까지는 직업 관련, 고용/노동, 취업 형태 등의 화제어가 높은 순위를 차지한 반면, 2015년 이후로는 '저출산'의 빈도가 증가

하였는데, 이는 취업자 및 재직자들의 고용 불안이 저출산으로 이어지고 있는 것으로 보임.

- 화제어를 카테고리별로 분류한 결과에서도 출산/보육 분류는 전체 화제어의 28.0%로 가장 많은 비중을 차지하였음. 정부/정책(26.0%)과 고용/노동(19.8%) 분류보다 출산/보육 비중이 더 높게 나타났는데, 이러한 결과는 고용 환경이 출산, 보육, 육아와 직접적으로 연결되어 있다는 것을 의미함.
- '4차 산업혁명'은 2017년에 높은 순위로 나타났는데 이는 정부의 정책 기조 효과와 더불어 고용창출에 대한 국민의 관심과 기대감으로 인한 것으로 보임.
- 워드 클라우드 분석 결과, '일자리', '저출산','지원', '정부', '취업', '일자리 창출', '청년', '여성'이 높은 빈도로 나타났으며, 관계망 분석에서도 '일자리'가 가장 큰 비중을 차지함.

[그림 6] '일자리' 키워드 워드클라우드 (2008년~2017년)

[그림 7] '일자리' 키워드 관계망 분석 (2008년~2017년)

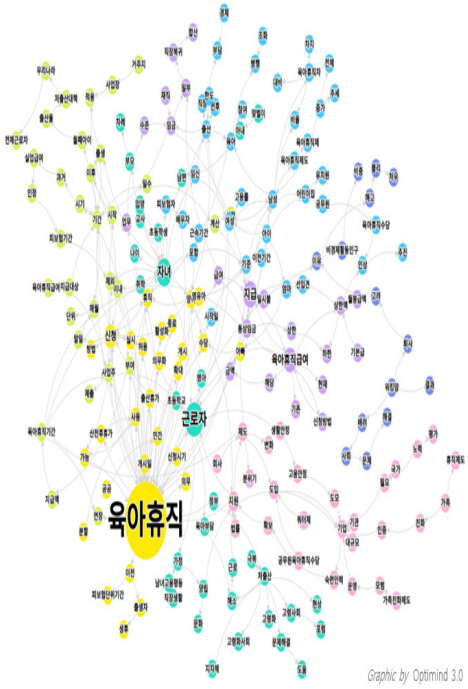


◆ 노동(labor)

- 화제어 추이 분석 및 워드 클라우드 분석 결과 지난 10년간 '육아휴직'이 높은 빈도로 나타났으며, 화제어를 카테고리별로 분류하였을 때에도 '휴가/휴직'이 43.2%로 가장 높은 비율을 차지함. 2016년 이후에는 '남성', '아빠', '자녀', '일-가정 양립' 등의 화제어가 높은 순위를 차지함으로써 양육에 대한 남성의 관심 및 참여, 일과 삶의 균형 등을 중시하기 시작하였음을 알 수 있음.
- 워드 클라우드 및 관계망 분석 결과, '육아휴직', '근로자', '자녀', '육아휴직', '급여'가 높은 빈도로 나타남으로써 양육자인 노동자의 육아휴직이 출산 및 양육과 밀접한 연관을 맺고 있음을 나타냄.

[그림 8] '노동' 키워드 워드 클라우드 (2008년~2017년)

[그림 9] '노동' 키워드 관계망 분석 (2008년~2017년)

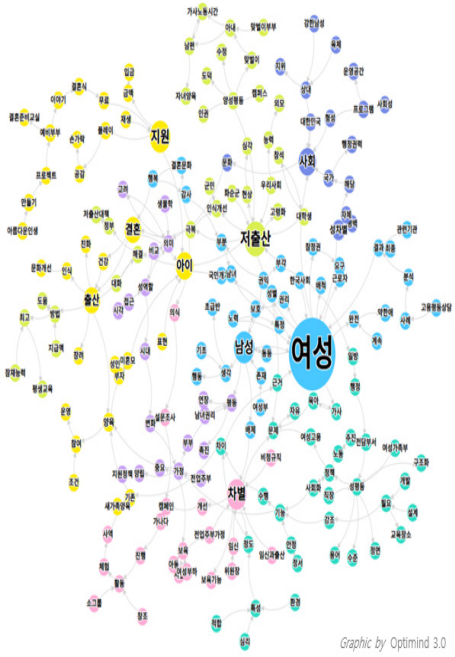


◆ 젠더(gender)

- 화제어 추이 분석 결과, 최근 10년간 '성차별', '성평등', '양육' 등의 화제어가 지속적으로 상위권에 위치하였으며, 2016년 이후부터는 '저출산'과 '출산'이 상위 순위를 기록함. 화제어 카테고리 분류 결과에서도 '양성평등'이 26.6%로 가장 높은 비중을 보이며 '출산/양육' 또한 22.6%로 유사한 수준의 비중을 차지함으로써 성 불평등에 대한 인식 및 그에 대한 불이익이 출산으로 연결되고 있는 것으로 보임.
- 워드 클라우드 및 관계망 분석 결과, '여성', '저출산', '출산', '남성', '결혼', '(성)차별'이 높은 빈도로 나타났으며, '여성'은 '육아', '가사', '차별'과 연결되어 있어 여성의 육아와 가사에 대한 부담, 노동환경에서의 차별이 출산에 영향을 미치고 있음을 보임.

[그림 10] '젠더' 키워드 워드 클라우드 (2008년~2017년)

[그림 11] '젠더' 키워드 관계망 분석 (2008년~2017년)



나. 4차 산업혁명 시대 저출산 관련 빅데이터 예측 모형 분석 결과

□ 머신러닝을 통한 주요 변수 검증 및 추출

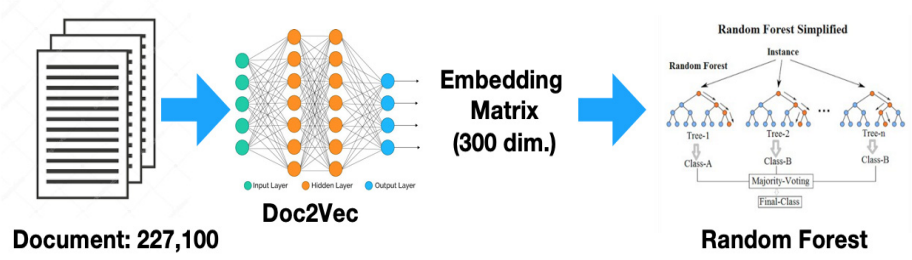
◆ 형태소 분석 및 카테고리 분류를 통하여 추출한 데이터 세트에 대하여 머신 러닝 기법을 활용하여 변수 도출

- 데이터 클리닝 작업을 통해 총 960여만 개의 데이터 중 5개 주제와 연관성을 가질 수 있는 227,100개의 문서를 일차적으로 추출하고, 각 주제의 특성을 전형적으로 잘 보여줄 수 있는 7,651개의 문서를 수기로 분류하여 기계학습 데이터(training data)로 활용함. 해당 문서의 학습 결과를 총 227,100개의 문서에 적용하여 기간별로 변인을 분류하고 적합도를 예측하였음.

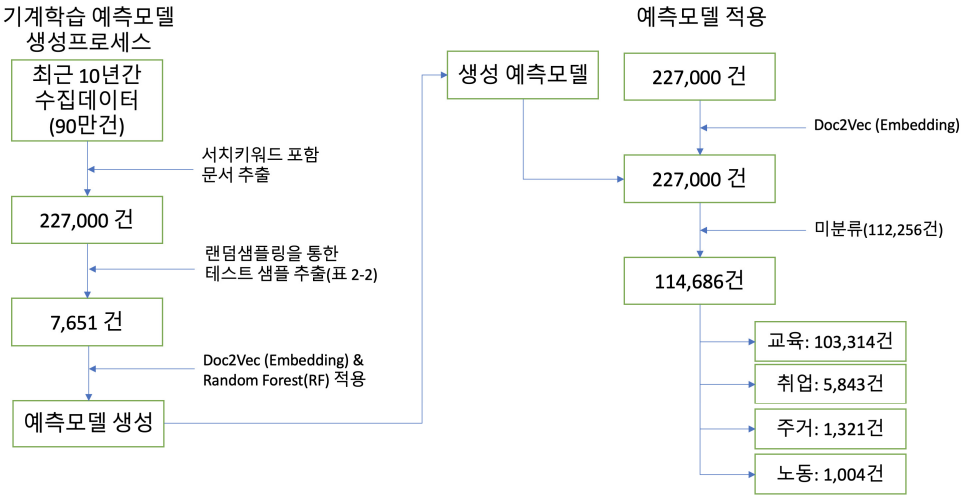
◆ 문서를 데이터로 활용하기 위하여 Doc2Vec 알고리즘 (Le & Mikolov, 2014)을 이용하였는데, Doc2Vec은 하나의 문서를 k차원의 실수 벡터(embedding vector)로 매핑하여 표현하는 대표적인 딥러닝 알고리즘임. 각각의 문서는 Doc2Vec를 통해 300개의 차원(k=300)를 가지는 실수값으로 변환되었음.

- 본 분석에서는 227,100건의 문서를 임베딩화하였기에 227,100개의 행과 300개의 열을 가지는 행렬(embedding matrix)을 Doc2Vec 알고리즘의 아웃풋으로 가지게 되었으며, 추출된 행렬은 저출산 현상 카테고리 분류를 위한 머신러닝 모델(Random Forest)의 훈련데이터로 이용되었음.

[그림 12] 적합도 예측 모형



[그림 13] 데이터 흐름 프로세스



- 기계학습 예측모델 생성을 위하여 수집데이터 90만건 중 서치키워드를 포함한 문서 227,000건을 추출하였고, 추출된 22만여건의 문서를 랜덤 샘플링하여 5가지 주요 주제별로 분류하였음.
- 분류한 데이터를 예측모델 적합을 위한 훈련데이터(training data)로 사용하여 Random Forest 예측 모델을 생성하였으며, 생성한 예측모델을 서치키워드를 포함한 문서 22만여 건에 적용하여 전체 데이터를 분류하였음.
  - 머신러닝 적합 결과, 전체 문서의 절반 정도인 112,256건이 어느 주제에도 포함시키기에 적합하지 않은 것(None)으로 나타났고, 보육, 육아를 포함한 교육 카테고리(education)에 104,314건의 문서가 포함, 가장 큰 비중을 차지했음.
  - 머신러닝에 의거한 주제 카테고리 예측을 위해 227,100개의 문서 중 부적합으로 나타난 112,256건을 제외한 114,686건을 최종적으로 활용하였음.

〈표 1〉 머신러닝 예측 결과

class_ prediction/ year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	SUM
None	8928	8257	10118	11586	12502	12837	12432	11282	12338	11976	112256
Education	12777	13314	11549	10180	9551	8534	9406	10624	9223	9156	104314
Education X Housing	228	206	177	122	156	158	96	157	171	406	1877
Education X Labor	3	4	20	3	7	7	2	4	10	10	70
Gender	41	16	11	30	23	21	21	14	39	41	257
Housing	60	53	83	49	68	38	27	143	290	510	1321
Job	568	645	659	657	410	706	590	357	554	697	5843
Labor	46	47	85	61	72	145	111	126	173	138	1004

◆ 랜덤 포레스트(Random Forest) 모형[1]에서 교육환경, 젠더, 주거, 노동, 일자리의 5개 요인을 트레이닝(training) 세트의 독립 변수로, (저)출산 관련 표현을 종속 변수(end feature)로 설정한 경우의 정확도를 계측한 결과임.

- 랜덤 포레스트 모형을 적용한 결과, 5개 변인 중 '젠더', '주거', '노동'은 저출산 현상과 유의미한 관련을 보였음.
- 기계학습 데이터(training)를 모수의 데이터(inference)에 적용했을 때의 종합적인 정확도(F1-score)[2]에서 교육 부문이 매우 저조(6.7%)하게 나와, 교육 변인은 종속변수를 단독으로 설명하는 독립변수로 기능하지 못하였음.
- '일자리(job)' 역시 F1-score 값이 28.5%로 매우 낮아, 저출산을 설명하는 직접적인 변수로 간주하기에는 무리가 따랐음.

1) 랜덤 포레스트 모형은 변수 간의 직접적인 상관관계뿐만 아니라 변수들을 합친 메타 변수와 종속 변수 간의 상관관계도 계산할 수 있는 장점이 있어, 연구자의 선행적 가설을 벗어나는 상황을 발견하기 위하여 주로 사용된다. 해당 머신러닝 모형은 저출산과 관련된 변인들이 복잡하게 상호작용하는 사회 환경을 반영하기에 적합한 것으로 판단되어 본 연구에서의 분석 도구로 채택하였다.

2) F1-score는 70% 이상을 유의미한 것으로 인정 https://en.wikipedia.org/wiki/Precision_and_recall

〈표 2〉 랜덤 포레스트 정확도 테스트 결과

	Class	Accuracy(%)	Recall(%)	F1-score(%)
Training	Education	100.0	100.0	100.0
	Gender	100.0	100.0	100.0
	Housing	100.0	100.0	100.0
	Labor	67.8	100.0	0.0
	Job	67.8	100.0	0.0
Inference	Education	54.7	96.7	6.7
	Gender	99.8	39.8	49.7
	Housing	99.8	74.0	52.2
	Labor	98.8	66.4	65.4
	Job	96.9	68.3	28.5

◆ 그러나 교육 변인은 다른 변인과 상호작용할 때 저출산 현상에 대하여 유의미한 관련을 맺고 있는 것으로 나타남.

- 서브 카테고리 분석 결과, '교육×주거'와 '교육×노동'이 유의미한 복합 변수로 인정됨. 즉, 교육문제만으로 출산환경을 설명할 수는 없지만, 교육과 연계된 주거문제 혹은 교육과 노동 환경이 복합적으로 야기하는 삶의 질은 출산에 유의미한 영향을 미친다고 해석할 수 있음.
- 정리하자면 '젠더', '주거', '노동'의 세 가지 단일 변인과 '교육×주거', '교육×노동'의 두 가지 복합 변인이 출산환경을 설명하는 독립변수가 될 수 있을 것으로 판단되었음.

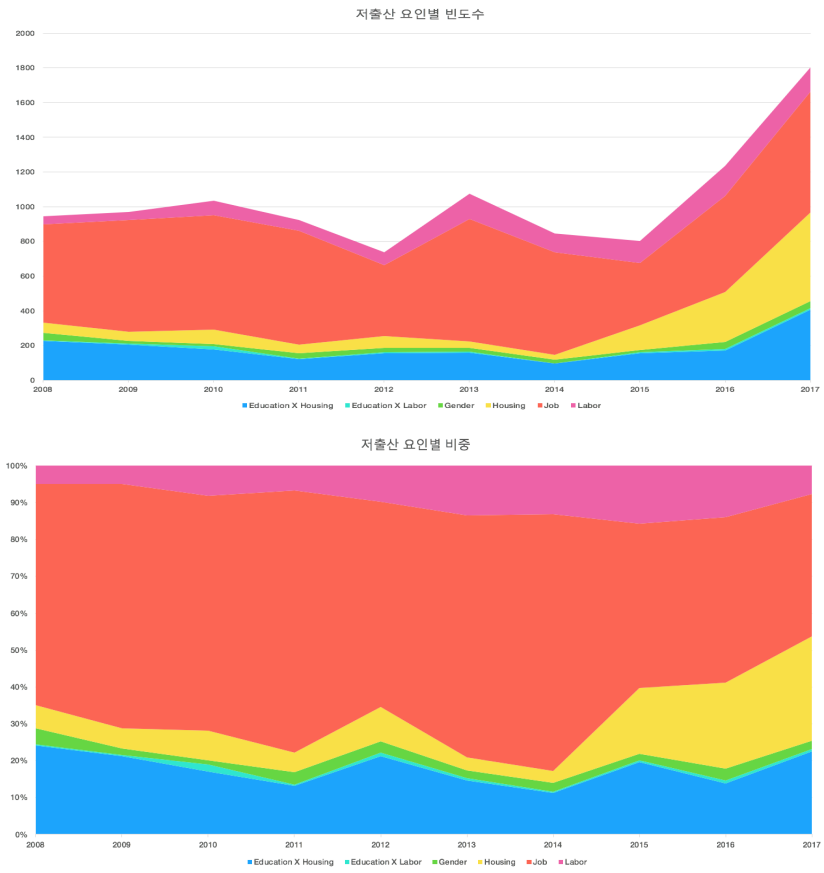
〈표 3〉 서브 카테고리 분석 결과

label	content	education	gender	job	housing	labor
	219244	0	0	0	0	0
Education	3087	3087	0	0	0	0
EducationGender	56	56	56	0	0	0
EducationGenderHousing	1	1	1	0	1	0
EducationGenderLabor	5	5	5	0	0	5
EducationHousing	**468**	**468**	**0**	**0**	**468**	**0**
EducationHousingJob	70	70	0	70	70	0
EducationHousingLabor	6	6	0	0	6	6
EducationJob	42	42	0	42	0	0
EducationLabor	**104**	**104**	**0**	**0**	**0**	**104**
Gender	448	0	448	0	0	0

◆ 아래의 그림은 5개 변인을 대변하는 소셜 미디어 포스팅의 연도별 개수 및 비율을 그래프로 표현한 것으로, 전반적으로 2016년부터 저출산을 설명하는 요인들에 대한 언급(빈도수)이 크게 높아지고 있음.

- 요인별 비중으로 보았을 때, 일자리 부족 등 '일자리'에 대한 언급은 2014년 이후로 눈에 띄게 줄어들고 있는 반면, '주거'에 대한 언급 비중이 2016년 이후로 크게 늘었고, '주거×교육'을 연계한 표현도 2016년 이후 증가하고 있음.
- '노동'과 '젠더'에 대한 문제의식은 별다른 변동 없이 적은 비중으로 나타나고 있고, '교육×노동'과 관련된 언급에 대한 비중은 매우 미미한 편임.

[그림 14] 저출산 연간 요인별 언급 추이



- 따라서 소셜 미디어에서는 앞서 머신러닝을 통해 추출된 5개 독립변수 중 '주거' 및 '주거×교육'의 두 가지 변인이 출산 환경을 직접적으로 설명하는 데 있어 가장 유의미한 변수로 검증되었음.
  - '일자리'는 머신러닝 정확도(F1-score)가 낮고 소셜 미디어에서의 언급이 감소세로 나타남. '노동', '젠더', '교육×노동'과 관련한 언급은 미미하거나 눈에 띄는 변동 없이 일정한 비중으로 나타났음.
  - 이러한 양상을 고려해볼 때, 주거난 혹은 교육과 연계된 주거 마련의 어려움이 가중될 때 저출산에 직접적으로 영향을 미치는 미시적인 요인들이 가장 뚜렷한 변동세를 보인다고 예측할 수 있음.
  - 랜덤 포레스트 모형 적용 결과, 5개 주제 키워드 중 '젠더문화', '주거문제,' '노동환경'의 세 가지 키워드 및 '교육환경×주거문제', '교육환경×노동환경'의 복합 변인이 출산 환경을 설명하는 독립변수로 나타남.

■ 예측 통계 모형 구축 및 시뮬레이션을 통한 검증

- '주거', '주거×교육', '고용', '노동환경'의 네 가지 주제 카테고리를 예측 모형에 적용하여 향후 출산율 변동을 예측하고 그 의미를 해석함.
- 예측 모형에 조출산율(fertility) 및 합계출산율(fertility rate)을 반응 변수로 설정하고 연간 고용률(employment), 가구당 소득(income)을 설명변수에 추가하는 경우, '주거×교육'과 연간 고용률만이 모형상의 유의성을 보임.
- '주거×교육'과 연간 고용률을 기반으로 포아송 자동 회귀모형(Poisson Autoregression model)을 활용하여 예측모형을 구축한 결과, 조출산율은 다소 반등한 후 2019년~2020년에 낮은 수준에서 정체될 것으로 예측됨. 이는 최근 고용률이 소폭 개선되는 추세가 반영되었기 때문인 것으로 보임.
- 빅데이터 예측 결과를 통하여, 단순한 소득 증대 혹은 고용률 신장보다는 주거문제와 교육 문제의 상호작용으로 나타나는 다양한 문제들에 대한 대책에 우선순위를 두어야 함을 알 수 있음.
- 소셜데이터에 대한 머신러닝 및 키워드 분석 결과, '교육', '일자리', '젠더'가 검증 과정에서 각각 단독적인 독립변수가 아니거나, 머신러닝 적합도가 낮게 나오거나, 혹

은 영향력이 미미한 것으로 도출되었음.

◆ 특히 주거 문제는 유의미한 설명력을 갖는 변수로 검출되었으며, 주거 중에서도 '주거' 변수 이외에 '주거와 교육' 문제가 결부된 복합변수 역시 중요한 독립변수로 도출되었음.

〈표 4〉 출산 환경 예측을 위한 주요 변수

주요 문제	검증 결과	도출 변수
주거 문제	가장 중요한 변수	주거, 주거×교육
교육 환경	단독 독립변수 아님	없음
일자리 문제	머신러닝 부적합	없음
노동 환경	작은 영향	노동 환경
젠더 문화	미미함	없음

◆ 데이터 처리 및 분석 결과를 예측 모형과 연계하여 적용함.

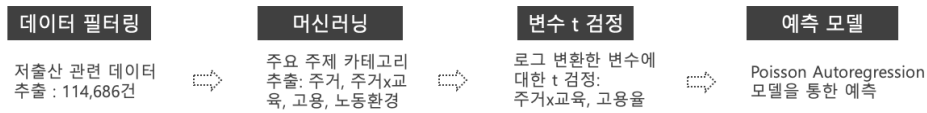
◆ '주거', '주거×교육', '고용', '노동환경'의 네 가지 주제 카테고리를 예측 모형에 적용하여 향후 출산율 변동을 예측하고 그 의미를 해석함.

- 저출산 관련 사회 환경 키워드를 포함하는 114,686건의 데이터에 머신러닝 모델을 적용하여 유의미한 관계를 갖는 주제 카테고리(주거, 주거×교육, 일자리, 노동)를 도출하였음.
- 주거 등 절대적으로 높은 버즈량을 갖는 변수와 노동환경 등 작은 버즈량을 갖는 변수간 절대값의 차이를 완화하기 위해 로그 변환을 수행한 후, t검정(t-test)을 통해 유의한 변수만을 추출하고, 해당 변수들을 예측 모형에 적용하여 향후의 출산율 변동을 예측하고자 하였음.
- 연도별 출산율 변동을 예측해야 하기에 일반화 선형 모델(generalized linear model, GLM)[3]을 이용하였음. 본 연구의 빅데이터 저출산 예측모형에서는 연도별 출생아 수를 종속변수로 고려하기에 GLM 중 대표적 시계열 모형인 포아송 자동 회귀모형(Poisson Autoregression model) (Fokianos, Rahbek & Tjøstheim, 2009)

3) GLM은 종속변수가 정규분포를 따르지 않을 경우 주로 사용되는 모형으로, 일반적으로 종속변수가 이항변수 (합격/불합격, 신용/불량 등)로 표현되는 경우, 종속변수가 사건의 수(월별 교통사고 발생 건수, 연간 출생아 수 등)로 표현되는 경우GML을 주로 사용함.

을 활용하였음[4].

[그림 15] 분석 및 예측 모형의 연계 과정



- 조출산율및 합계 출산율(fertility rate)을 반응변수로 설정한 후, 소셜 미디어 데이터를 통해 수집된 '주거', '주거×교육', '교육', '근로환경(labor)', '일자리(job)' 관련 포스팅 추세를 수집하였음. 또한 통계청의 연간 고용률(employment)[5], 가구당 소득(income)[6]을 설명변수에 추가하였음[7].
- 이후, 변수간 편차를 최소화하기 위해 로그 변환(log transformation)을 하여 t검정을 수행하였음. 반응변수와 설명변수간의 1년간 시차를 둔 모형(반응변수의 결과가 종속변수에 1년 후 영향을 미치는 것으로 상정)의 검정 결과, 1) '주거×교육' 변수와 통계청 자료인 2) 연간 고용률만이 모형상의 유의성을 가짐. 가구당 소득이 변수로 들어가는 경우 전체모형은 유의하지 않았음.

4) 포아송 자동회귀모형을 사용한 이유는 ① 종속변수인 조 출산율은 "천 명당 출생아 수"로서 사건의 수에 대한 회귀분석이 필요하며 ② 연도별 시계열 자료의 특성을 반영하여야 하기에 auto-regressive model이 필요하기 때문임. 상기 예측모형을 통해 향후 3년간의 천 명당 출생아 수를 시간에 대한 상관관계를 고려하여 추정할 수 있음.

5) 연간 고용률은 15세 이상 생산가능 인구(군인, 재소자 제외)에서 취업자가 차지하는 비율로, 본연구에서는 경제활동인구 총괄조사를 통해 수집된 자료(시계열 보정값)를 활용하였음.

6) 가구당 소득은 한 가구의 근로소득, 사업소득, 재산소득, 경상이전소득, 비경상소득의 합으로 소득세, 건강보험료, 국인연금 기여금 등을 공제하기 전 소득을 의미하며, 본 연구에서는 2017년 가계동향조사를 통해 수집된 자료를 활용하였음.

7) 통계청 데이터는 국가통계포털(http://kosis.kr/index/index.do)을 통해 수집된 자료를 활용하였음. 통계청의 연간 고용률은 소셜미디어의 고용 관련 언급을 보완하는 객관적인 지표로 검증의 가치가 있음. 가구당 소득은 사회 통념상 중요하게 받아들여지는 소득 증대 수준이 예측에서도 출산율과 직접적인 상관관계를 갖는지를 검증하는 의미를 지님. 본 연구에서는 이러한 통계청 데이터를 가미함으로써 주관적인 기술의 한계를 갖는 소셜 미디어 버즈 데이터를 보완하였음.

〈표 5〉 예측모형을 위한 설명변수 및 반응변수

year	fertility	fertility_rate	housing	edu_housing	education	labor	job	employment	income
2008	94	1.192	60	228	12777	46	568	64	4728
2009	90	1.149	53	206	13314	47	645	63	4656
2010	94	1.226	83	177	11549	85	659	63.4	4788
2011	94	1.244	49	122	10180	61	657	63.9	4872
2012	96	1.297	68	156	9551	72	410	64.3	5052
2013	86	1.187	38	158	8534	145	706	64.6	5088
2014	86	1.205	27	96	9406	111	590	65.6	5196
2015	86	1.239	143	157	10624	126	357	65.9	5244
2016	79	1.172	290	171	9223	173	554	66.1	5232
2017	70	1.052	510	406	9156	138	697	66.6	5385

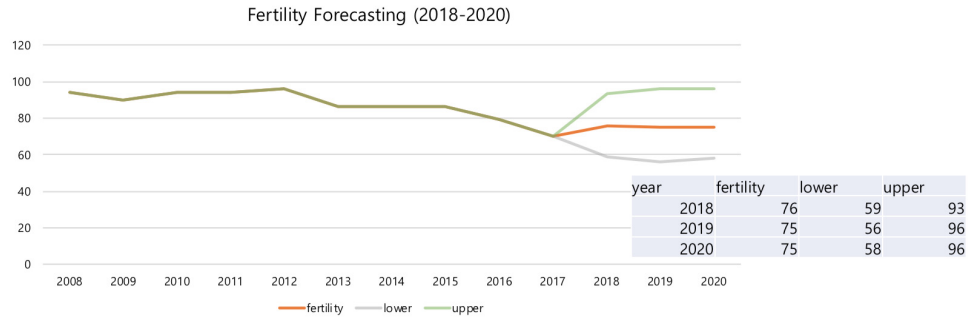
〈표 6〉 변수들의 로그 변환값

year	fertility	fertility_rate	housing	edu_housing	education	labor	job	employment	income
2008	94	1.192	4.094	5.429	9.455	3.829	6.342	4.159	8.461
2009	90	1.149	3.970	5.328	9.497	3.850	6.469	4.143	8.446
2010	94	1.226	4.419	5.176	9.354	4.443	6.491	4.149	8.474
2011	94	1.244	3.892	4.804	9.228	4.111	6.488	4.157	8.491
2012	96	1.297	4.220	5.050	9.164	4.277	6.016	4.164	8.528
2013	86	1.187	3.638	5.063	9.052	4.977	6.560	4.168	8.535
2014	86	1.205	3.296	4.564	9.149	4.710	6.380	4.184	8.556
2015	86	1.239	4.963	5.056	9.271	4.836	5.878	4.188	8.565
2016	79	1.172	5.670	5.142	9.129	5.153	6.317	4.191	8.563
2017	70	1.052	6.234	6.006	9.122	4.927	6.547	4.199	8.591

◆ 예측모형은 유의한 반응변수로 도출된 '주거×교육' 및 고용률을 기반으로 일반화 선형모형(generalized linear model)의 시계열 예측 모형인 포아송 자동 회귀모형(Poisson Autoregression model)(Fokianos et al., 2009)을 활용[8]하였음.

8) 포아송 자동회귀모형을 사용한 이유는 첫 번째로 종속변수인 조출산율이 '백명당 출생아 수'이므로 count에 대한 회귀분석이 필요하였으며, 두 번째로는 연도별 시계열 자료의 특성을 반영하여야 하므로 자기 회귀 모형(auto-regressive model)이 필요했기 때문임. 상기 예측모형을 통해 향후 3년간의 천 명당 출생아 수를 시간에 대한 상관관계를 고려하여 추정할 수 있을 것으로 예상하였음.

[그림 16] 향후 조출산율 예측



year	fertility	lower	upper
2018	76	59	93
2019	75	56	96
2020	75	58	96

◆ 위의 그림은 해당 모형을 통해 2018년 ~ 2020년 기간의 조출산율을 예측하였음. 2017년에 70까지 떨어진 조출산율은 다소 반등한 후 2019년 부터 2020년 까지 낮은 수준에서 정체될 것으로 예측됨.

- 반등이 예측된 이유는 최근 고용률이 미미하게나마 개선되는 추세가 반영되었기 때문인 것으로 보임. 그러나 '주거×교육'과 관련된 부정적인 소셜 미디어 포스팅이 증가하는 추세가 심화되면서, 해당 문제를 방치할 경우 출산율 회복을 위한 삶의 여건이 근본적으로 개선되지 못할 것으로 해석됨.

◆ 즉, 주거×교육 변수가 저출산 추세를 설명하는 데 있어 가장 유력한 변수로 도출된 것은 특정 지역의 학군에 대한 선호 집중 등으로 인하여 주거 비용과 교육 비용의 동조화 추세가 심화되었으며, 문제가 되는 환경이 부모들로 하여금 자녀의 육아와 교육을 위한 미래 투자를 결심하거나 실행하는 데 결정적인 걸림돌로 작용할 수 있음을 시사함.

- 빅데이터 예측 결과가 정책적으로 시사하는 바는, 일자리 등 고용률 신장을 위한 정부의 노력만으로는 출산율로 표현되는 삶의 질 개선이 어렵다는 것을 의미함. 이는 단순한 일자리 늘리기나 소득증대만으로는 부부에게 출산을 결심할 만한 수준의 삶의 질을 보장하지 못할 수도 있음.

◆ 교육을 위하여 특정 지역에 대한 선호가 쏠리고, 그러한 선호가 자동적으로 높은 집값을 견인하는 사회문화 환경의 변화가 선결되어야 보다 안정적인 환경 속에서 미래의 삶에 대한 예측 가능성을 가지고 출산을 고려할 수 있음.

◆ 지금까지의 데이터 분석 결과를 살펴보면 물리적인 주거와 교육 환경이 동일시되고, 따라서 주거×교육 비용 역시 동조화 되어 육아와 교육의 부담을 가중시키는 기존 삶의 문법을 4차 산업혁명의 패러다임 속에서 어떻게 바꾸어 나갈 것인가의 방법론에 대한 화두를 주고 있다고 볼 수 있음.

3 4차 산업혁명 시대 육아정책의 방향과 과제

가. 정책방향

■ 주거, 교육, 고용, 노동, 젠더문화 제고를 통한 삶의 질 향상

◆ 주거의 경우, 일-가정 양립, 신혼부부, 결혼, 공급, 서민, 생활의 키워드와 관련이 있었음. 결혼과 출산 키워드와 밀접한 서민 신혼부부들에 대한 주거 공급의 문제를 고민할 필요가 있음.

◆ 교육의 경우, 교육자체의 의미보다, 사회복지정책 내에서의 교육의 의미를 보다 고려하는 방향으로 정책을 개선할 필요가 있음.

◆ 일자리의 경우, 고용 불안 해소, 고용환경 개선이 중요함.

◆ 노동의 경우, 육아휴직과 양육에 대한 아버지 참여, 일과 삶의 균형에 중점을 둘 필요가 있음.

◆ 젠더의 경우, 양성평등의 문제와 여성의 육아와 가사부담, 노동환경 차별의 문제를 해결하는 방향으로 정책을 개선하여 삶의 질 향상에 중점을 둘 필요가 있음.

■ 주거×교육의 동조화 현상 해소를 위한 노력

◆ '주택문제' 및 '주택문제×교육환경' 카테고리를 대상으로 의미망 분석 결과, 중심부에 '저출산 현상'을 둘러싸고 정부 정책 지원의 필요성을 강조하는 내용이 주를 이룸.

◆ 의미 네트워크 분석 결과에도, 일자리 및 주거공간의 '불안정', '가계부채'가 대표적으로 매개 변수값이 높은 키워드로 도출되며, '불안정' 키워드는 '청년', '신혼부부'와 연결되어 있어 불안정한 환경으로 인하여 결혼을 포기하는 상황이 저출산 및 갈등 환

경의 주요 요인으로 작용할 수 있음을 나타냄.

- ‘(저)소득’은 ‘가계부채’와 직접 연결되어 있는데, 이는 주거와 연계된 가계부채 문제가 소득 문제를 포함하고, 주거 문제가 일자리를 포괄한다고 볼 수 있음.
- 의미화 클러스터링을 구조화해 보면, ‘여성’ 클러스터는 ‘주거’, ‘일자리’ ‘청년’ 등의 이슈가 도출되어 있으며 연관어인 ‘강화’는 ‘민간’, ‘협력’과 직접적으로 연결되어 있음. ‘자녀’ 클러스터의 경우, ‘육아’, ‘부담’, ‘경제’, ‘양육’, ‘사교육비’와 같은 연관어가 포함되었으며 ‘부담’이 ‘주거비’ 및 ‘양육비’와 직접 연결되고 ‘사교육비’와 관련된 언급이 ‘집값’과 직접 연결되는 양상은 머신러닝 결과와도 유사함.
- 종합하면, 주거와 교육 동조화 현상은 교육의 출발선 평등이라는 가치를 무색하게 하고, 교육의 공정성을 무너뜨리는 주요 원인이 되고 있음. 이러한 현상은 결국 여러 간접적인 변인들에 영향을 미치면서 저출생 현상에 영향을 미치는 것으로 나타났음. 주거와 교육 동조화 현상에서 나타나는 갈등을 해결하는 방향으로 정책을 개선할 필요가 있음.

나. 정책과제

- 교육격차를 줄이는 공공임대주택 획기적 확대
  - 주거-교육의 공공성을 확보하여 소득 및 계층 격차를 해소할 필요가 있음. 이를 위해 공공임대주택을 획기적으로 확대하고, 교육과 주거의 동조화 현상으로 계층간 격차를 줄이도록 함. 공공임대주택이 저소득층이 이용하는 곳이라는 기존의 프레임을 깨고, 선진국과 같이 주택 자체가 자가소유나 임대 주택이나 질의 측면에서 차이가 없이 공급 배치하여 수요자가 선택할 수 있는 폭을 늘려주는 정책 기획이 필요함.
  - 특히 이러한 정책적 기획은 특별시나 광역시, 강남, 서초, 송파를 중심으로 시행할 수 있도록 하도록 재구조화할 필요가 있겠음. 서울시가 재건축, 재개발시 전체 공급량의 10%를 공공임대주택을 짓도록 하고 있는데, 이 비율을 대폭적으로 높일 필요가 있음.
- 이음새 없는 통합 유아교육-보육-학교돌봄 시스템 구축을 통한 교육·보육의 공공성 강화
  - 유치원과 어린이집의 공공성 강화를 통한 교육보육의 질 제고, 육아휴직 확대, 국공립

기관 확대, 사교육비 경감, 학령기 아동의 방과후돌봄교실 확대 정책을 보다 강화할 필요가 있음. 즉, 교육과 보육의 공공성을 높여 육아부담을 줄이고, 주거와 교육의 동조화 현상에서 나타나는 사회적, 문화적 자본의 편중 현상을 줄이는 것에 중점을 두는 것임.

- 국공립 유치원과 사립 유치원, 어린이집은 시스템과 질의 편차가 크고, 가까운 곳에 믿고 맡길 수 있는 유치원과 어린이집을 구하기가 어렵다는 점에서 양의 문제 뿐만 아니라, 질의 문제도 존재하고 있음.
  - 이에 대한 해결방안으로는 점차적으로 사립 유치원을 학교 법인화하여 유아학교 시스템을 구축하고 의무교육화하는 것이 필요함. 또한 어린이집도 법인화를 유도하여 통합 유아학교 시스템을 구축하고 유치원과 어린이집에 대한 행정지원 시스템을 일원화할 필요가 있음. 이를 초등학교 시스템과 연계하고 방과후 돌봄 시스템을 연계하여 이음새 없는(seamless) 통합 유아교육, 보육 시스템 구축할 필요가 있음.

■ 스마트 워크/텔레워킹/거주지 인근 공용 오피스 근무, 출산 및 육아 유급 휴가 적극 지원

- 저출산과 육아정책에서 가장 중요하다고 판단한 '일자리' 문제의 경우, 다양한 방식의 근로 형태를 강화하는 방향으로의 정책 제고가 반드시 필요함. 즉, 일과 육아를 병행하고, 재택근무 기회를 늘리며 육아휴직의 기회를 실질적으로 활용하는 것이 중요함.
- 이상의 정책들은 기본적으로 정부가 실행하고 있는 정책들이나 실제로 실효성이 있는지 제고해볼 필요가 있으며, 재택근무와 육아휴직 정책이 실질적으로 자리매김할 수 있도록 보다 적극적이고 공격적인 정책 마련이 필요하다고 볼 수 있겠음.
- 스마트 워크와 재택근무를 높일 수 있도록 근무환경을 조성하고, 자녀를 함께 키우는 육아공동체를 온/오프라인으로 손쉽게 구축할 수 있도록 지역수준, 정부수준의 총체적인 지원이 필요함.
- '일과 육아를 병행할 수 있는 일자리 창출', '출산휴가 및 육아휴직 기간 중 재택근무 기회 확대', '출산휴가 및 육아휴직 재원 확대', '출산휴가 및 육아휴직 사용시 대체인력 지원', '경력단절 여성 및 남성에 대한 정책 마련', '다양한 근무환경과 근무방식

도입(재택근무, 스마트워크 등)', '모성보호관련법, 남녀고용평등법 및 근로기준법을 기업이 준수하고 있는지 대한 감독 강화'가 필요함.

- 성별 임금 격차 해소, 다양한 가족 형태 지원
  - 성평등 직장문화 확충을 위한 직장문화 개선, 성별 임금격차 해소, 혼외출산에 대한 지원 정책을 강화하여 노동과 일자리 문제에서 발생하는 여성의 고충을 덜어줄 필요가 있음.
  - 성별 임금 격차 해소를 하는 기업에 인센티브를 주고, 법 개정을 통해 가족의 형태를 보다 다양하게 규정하고, 혼외출산과 동거, 미혼모/부 가정도 결혼한 가정과 같이 동등한 권리를 누릴 수 있도록 지원할 필요가 있음.

Ⅰ참고문헌Ⅰ

박창현·김나영·이유진(2018). 4차 산업혁명 시대 육아정책의 이슈와 과제. 육아정책연구소.

Fokianos, K., Rahbek, A., & Tjøstheim, D. (2009). Poisson autoregression. Journal of the American Statistical Association, 104(488), 1430-1439.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196).

[참고 웹사이트]

국가통계포털(http://kosis.kr/), (검색일: 2018. 3. 6)