Homo Migrans Vol.24(May 2021): 135-157

■ 논문 ■

텍스트마이닝을 통한 코로나 발생 전후 시기의 다문화에 대한 인식변화*

최 강 화

Ⅰ. 머리말

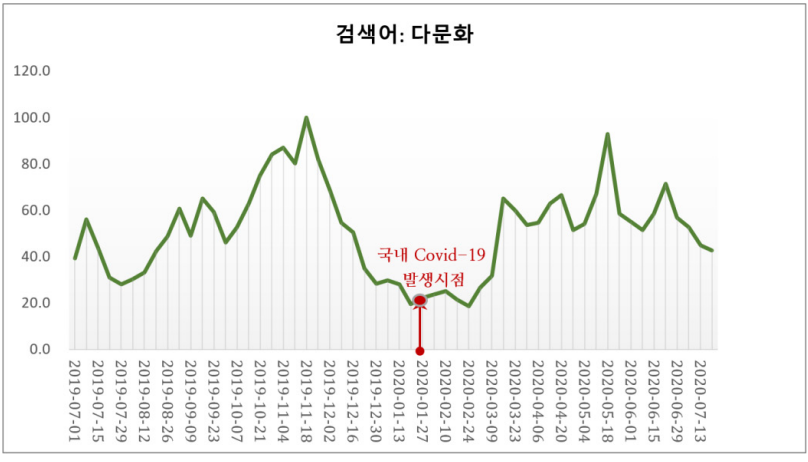
현 시대의 우리는 지금까지 한 번도 경험하지 못한 신종 코로나 바이러스 감염증(이하 코로나 19)으로 인해 전 세계가 혼란과 혼돈 속에 빠져들고 있으며, 현재도 진행 중인 코로나 팬데믹(COVID pandemic) 현상으로 우리 일상생활에 상당히 많은 부분이 변화하고 있다. 특히, 전세계적으로 감염병 공포와 더불어 동아시아계 외국인에 대한 혐오 수준은 극에 달해 있고 아시아인을 대상으로 한 혐오 범죄가 계속해서 증가하고 있는 상황이다. 미국 뉴욕의 경우에는 2019년에 아시아계를 대상으로 한 혐오 범죄가 3건이었던 것이 2020년에는 무려 28건으로 증가하며 특정 인종을 대상으로 한 혐오범죄가 급증하고 있다.[1] 국내에서도 특정 국가의 외국인의 입국을 금지하는 청원에 서명하는 사람들의 숫자가 늘어나는 등의 외국인 혐오 및 차별에 대한 대규모의 움직임이 포착되고 있는 상황이다.

이와 같은 상황이 지속되면서 다문화 가정은 우리 사회로부터 소외되고 외면받기 시작하였고, 코로나 19로 인하여 가장 큰 피해를 보는 영역 중의

* 이 논문은 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2019082514).

1) 본 내용은 2021년 3월 13일자 연합뉴스 기사를 토대로 작성함.

하나가 되었다. 특히, 경제가 어려워지면서 이주 노동자들은 직장 내 해고 1순위가 되기도 하고, 다문화 가정은 사회 공동체에서 소외되기도 하였다. 특히, 국내 거주하는 외국인에 대한 부정적 인식이 확산되면서 다문화 가족들의 삶은 더욱 더 위축되고 어려워지고 있는 상황이다. 네이버의 데이터 랩(datalab.naver.com)에서 '다문화'라는 키워드를 입력하여 검색된 검색량을 살펴보면, 다음의 [그림 1]과 같이, 국내에서 코로나 19가 처음 발생한 2020년 1월 21일을 기점으로 '다문화'라는 키워드 검색의 코로나 발생 이전 및 이후의 차이가 명확한 차이를 보여주고 있다.



[그림 1] 국내 코로나-19 발생 전후의 다문화 검색량 변화[2)]

즉, 코로나 발생 이전에는 다문화 관련 검색어의 빈도가 많은 반면, 코로나 발생 시점 전후 약 3개월간에는 다문화 관련 검색어의 비중이 급격히 감소하며, 우리의 관심에서 멀어지고 있는 양상을 보이고 있었다. 이후 2020년 3월 이후에는 예전 수준으로 증가하고 있는 양상을 보이고 있기는 하지만, 다문화에 대한 인식이 질적으로 어떻게 변화하였는지 또는 다문화 관련 키워드들은 코로나 전에 비해 어떤 변화를 가져오는 지에 대한

2) 네이버 데이터랩 검색어 트렌드 분석 [(1) 주제어: 다문화, (2) 검색기간: 2019. 07. 21~2020. 07. 20, (3) 검색범위: PC, 모바일 전체, 성별 전체, 연령 전체]

추가적인 분석이 필요한 상황이다.

따라서 본 연구에서는 코로나 19와 같은 범세계적 팬데믹 위기 상황하에서 국내외 포털 뉴스들이 바라 본 다문화와 관련한 인식들이 질적으로 어떻게 변화하고 있는가를 측정하고자 한다. 이를 통해 이러한 국내외 환경변화에 대응하여 다문화 정책의 일관성을 유지하고 보다 효율적인 다문화 정책관리를 위한 정책적 대안을 제시하고자 한다. 본 연구에서는 우선 텍스트마이닝(Text mining)을 통해 국내 코로나 발생 시점을 전후해서 가장 많이 논의되고 있는 키워드를 도출하고, 이러한 키워드들 간의 CONCOR 분석을 수행한다. 또한 추가적으로 이러한 키워드들의 언어 감성분석을 통해 긍정적 또는 부정적 감성들의 변동을 살펴봄으로써 이러한 국내외 위기상황에 따른 감성단어들의 변화를 추적해 보고자 한다.

본 연구에서는 다음과 같은 연구 주제를 수행하고자 한다.

·연구주제 1: 코로나 팬데믹 전후 시점에 다문화라는 키워드 분석을 토대로 CONCOR 분석에서 동일한 관계패턴을 가진 언어 클러스터(cluster)들 간에 어떤 변화가 있는가?

·연구주제 2: 코로나 팬데믹 전후 시점에 다문화라는 키워드 분석을 토대로 감성언어 분서 결과에서 긍정적 또는 부정적 감성 단어들 간에 어떤 변화가 있는가?

II. 코로나 상황 하에서의 다문화 관련 기존 연구

최근에 다문화를 주제로 텍스트마이닝 기법을 활용한 다양한 연구들이 활발히 진행되고 있다. 특히, 안명숙(2018)과 윤희진(2020)의 연구는 다문화에 대한 연구 가운데에서 대중들이 가지고 있는 다문화에 대한 인식을 분석하였고, 이수정과 최두영(2020)의 연구는 한국 사회에서 이주민 또는 이민자에 대해 가지는 대준의 인식을 연구하였다. 또한, 강진구와 이기성(2018)의 연구는 이주 난민에 대한 대중들의 인식을 텍스트마이닝으로

측정하였으며, 김수정, 마경희, 윤성은(2020)의 연구와 김태종(2020)의 연구는 2019년 말부터 시작된 코로나 19 사태가 다문화 가족이나 다문화 사회에 미치는 영향을 상세히 분석하였다.

우선, 본 연구의 분석 방법과 유사한 텍스트마이닝 기법을 활용하여 다문화를 분석한 기존의 연구들을 살펴보면, 김세현(2018)은 다문화와 관련한 기존의 연구들의 동향을 분석하기 위해 약 17년 동안 국내의 다문화 아카이브(CSMR Archive)에 수록된 다문화 논문 초록을 텍스트마이닝 기법을 활용하여 분석하였으며, 김세현의 연구(2018)에서는 다문화 연구 토픽을 크게 교육, 이주 그리고 정책과 같이 세 가지로 분류하고 잠재 디리클레 할당(LDA: latent dirichlet allocation) 방법으로 개별 토픽의 비정형적 자료를 분석하였다. 안명숙(2018)의 연구는 다문화 정착기에 접어든 한국사회에서 다문화와 관련한 인식을 살펴보기 위해 '결혼이주여성'이라는 키워드 분석을 통해 연관된 핵심단어들의 네트워크 분석을 수행하였다. 특히, 이 연구에서는 한국사회에서 결혼이주 여성들에게 처해 있는 차별적 요인들에 대해 조명하며, 결혼이주여성들이 한국에서 보다 안정적으로 정주하기 위한 다양한 요인들을 제시하였다. 또한 윤희진(2020)의 연구에서는 국내 다문화 관련 멘토링을 텍스트마이닝으로 분석하였는데, 이 연구는 국내에서 발간되었던 학술지 및 학위논문의 초록과 인용 정보를 활용하여 토픽 모델링을 분석하였고, 다문화 멘토링을 다문화 지원 프로그램, 다문화 사회의 이슈, 멘토링 경험 그리고 한국 문화의 이해와 같은 네 가지 범주로 분석을 시도하였다. 이수정과 최두영(2020)의 연구는 '이주'와 '이민'이라는 키워드를 활용하여 논문 및 언론 기사에 나타난 다문화 관련 이슈들을 분석하였다. 이 연구에서 이주 및 이민 관련 논문들이 다문화 사회에 어떤 영향을 미치고 있는가를 해석하였다. 강진구와 이기성(2019)은 제주도에 입국한 예멘 난민들을 대상으로 한국의 네이버 댓글은 난민에 대해 어떠한 인식을 가지고 있는가를 텍스트마이닝 기법을 이용하여 분석하였다. 이 연구에서는 빈도수 분석과 토픽 분석 그리고 감정분석을 수행하였으며, 상당수의 네이버 댓글이 예멘 난민에 대한 부정적 인식을 보여주고 있음을 제시하고 있다. 이 연구는 난민에 대한 대중들의 인식을 텍스트마이닝

기법을 통해 분석하였다는 점에서 연구의 의의가 있다.

연구자	주제(키워드)	분석대상	텍스트마이닝 기법
김세현(2018)	다문화 관련 연구의 동향	다문화 아카이브(CSMR Archive)에 수록된 다문화 논문정보	LDA 분석, 텍스트 네트워크 분석
안명숙(2018)	결혼이주여성	네이버, 다음, 페이스북, 트위터, 유튜브 등 9,023개 단어	키워드 분석, 텍스트 네트워크 분석
강진구와 이기성(2019)	예멘 난민	네이버 뉴스에 달린 댓글 분석(제주도 난민, 예멘 난민, 정우성 난민, 난민 가짜뉴스)	빈도수 분석, 토픽 분석, 감정분석
윤희진(2020)	다문화 멘토링	학술지논문 104편, 학위논문 102편	LDA 분석
이수정과 최두영(2020)	이주, 이민	KCI 논문 46 편과 조선/중앙/동아일보 /한겨레신문 기사 16편	키워드 분석, 공기어 네트워크 분석

[표 1] 텍스트마이닝 기법을 활용하여 다문화를 분석한 기존의 연구들

한편, 코로나 19가 확산되고 있는 현 시점에서 다문화와 관련한 기존의 다양한 연구들이 진행되어 왔는데, 김태종(2020)의 연구에서는 코로나 19와 관련된 뉴스 빅데이터를 이용하여 언론에서 언급되는 코로나 관련 의제들이 무엇인가를 토픽 모델링하였다. 이 연구에서는 47,816 건의 뉴스 데이터를 감염병 위기 경보에 따라 4단계로 구분하여 약 20개의 핵심 토픽을 찾고, 언론 보도의 방향성을 제시하고 있다. 또한, 김수정, 마경희, 윤성은(2020)의 연구에서는 다문화 가족지원 센터의 코로나 19 대응 및 정책적 과제를 탐색하기 위해 다문화 가족지원 센터의 센터장을 대상으로 포커스 그룹 인터뷰를 진행하였다. 이 연구에서는 사회적 재난 환경 하에서 다문화 가족지원 센터가 제공하는 다양한 지원 프로그램에 대한 정보를 제공하고 포스트 코로나 시대에 필요한 정책적 지원의 새로운 방안을 질적 연구를 통해 탐색하였다. 특히, 다문화 가족지원 센터의 비대면 온라인 서비스의 확대와 지원 대상자 맞춤형 혼합 서비스의 개발 그리고 포용적 정책 개발의 필요성을 제시하고 있다.

III. 텍스트마이닝(Text mining) 기법

텍스트마이닝은 텍스트 형태로 이루어진 정형 또는 비정형의 데이터들을 자연어 처리방식과 문서처리 방법을 적용하여 유용한 정보를 추출하고 가공하는 빅데이터를 처리하는 기법이다. 여기서 텍스트는 일반 문서나 도서뿐만 아니라 웹페이지, 블로그, 전자 저널, 이메일 등 전자문서 등을 포괄하는 자료원천으로 우리 일상에서 가장 흔하게 접할 수 있는 형태의 정보들이다. 텍스트마이닝은 주로 데이터로부터 유용한 의미나 통찰(insight)을 발굴하는 데이터마이닝, 언어를 정보로 변환하기 위한 자연어 처리(natural language process), 정보 검색 등 다양한 전문 영역들이 접목되어 발전한 융합 기술이자 분석 도구이다. 즉, 텍스트마이닝은 비정형 또는 반구조화된 텍스트 집합에서 자연어 처리기술과 기계학습, 인덱싱(indexing), 온톨로지(ontology) 등의 기술을 이용하여 유용한 정보를 획득하고 의미를 정제하고 범주화하는 과정이다.[3)]

이와 같이 텍스트마이닝은 논문, 신문 그리고 보고서 등과 같은 문서의 요약, 자동 범주화와 같은 문서 분류, 유사 단어 또는 유사 문서 간의 군집 분석, 주요 키워드의 추출과 같이 다양한 분야에서 활용될 수 있다. 일반적으로 텍스트마이닝은 텍스트 전처리 단계로 문자 열을 단어, 문장, 단락으로 조각화하는 토큰화(tokenization)와 어절 또는 문장을 최소의미 단위인 형태로소 분절하여 불필요한 단어 또는 문자를 제거하는 정제(cleaning) 및 같은 의미이면서 표현이 다른 단어를 통합하는 정규화(normalization) 과정을 거쳐 문서와 같은 자료원천으로부터 주요 핵심 키워드를 분류하고 특징을 추출한다.[4)] 이를 통해 궁극적으로 핵심단어들 간의 관계구조 분석 및 연관성 분석, 감성단어 분석, 토픽 모델[5)] 등의 결과물을 만들어 낸다.

---

3) 텍스트마이닝에 대한 정의는 곽기영, 『소셜네트워크 분석』 (청람, 2014)과 딘 러셔, 요한 코스키넨, 개리 로빈스, 최수진 역, 『사회 네트워크 통계 모형(EGRM)』 (한울, 2020)의 연구를 일부 인용함.

4) 김용학, 김영진, 『사회 연결망 분석』 (박영사, 2016).

5) 토픽 모델(topic model)은 기계 학습 및 자연언어 처리 분야 중의 하나로, 문

IV. 텍스트마이닝 실증분석 결과

본 연구에서는 '다문화'의 키워드를 활용하여 텍스트마이닝을 수행하였다. 데이터 수집은 웹과 소셜 네트워크 분석(social network analysis) 전문 프로그램인 텍스톰(TEXTOM 5.0)을 이용하였고, 수집 채널은 네이버(Naver), 다음(Daum) 그리고 구글(Google)의 뉴스를 검색하였다. 데이터 수집 기간은 코로나 발생 이전 6개월(2019년 7월 21일 ~ 2020년 1월 20일)과 코로나 발생 이후 6개월(2020년 1월 21일 ~ 2020년 7월 20일)을 지정하여 월 단위로 데이터를 수집하였으며, 코로나 발생 이전에는 총 11,153건(네이버: 5,152건, 다음: 4,372건, 구글: 1,629건)의 다문화 관련 문서를 수집하였고, 코로나 발생 이후에는 총 11,141건(네이버: 5,121건, 다음: 4,440건, 구글: 1,580건)의 문서를 수집하였다.

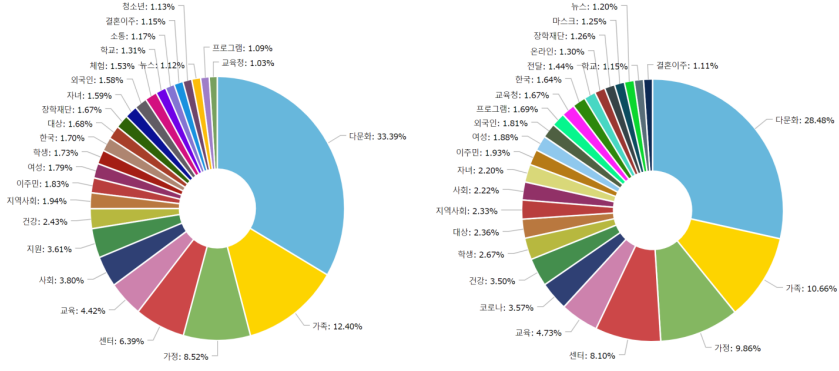
다문화와 관련되어 있는 문서들을 수집 후에 형태소 분석에서는 고유명사, 복합명사를 그대로 결과값에 반영하는 Expresso K를 활용하였으며, 분석 품사는 체언의 일반 명사, 교유 명사, 의존 명사, 단위 명사, 수사, 대명사 등을 위주로 추출하였다. 또한 본 연구와 전혀 상관없는 무의미한 조사들과 단음절의 용어들을 모두 제거한 후에 분석을 시작하였다. 또한 본 연구에서는 데이터가 가지고 있는 정확한 의미를 해석하는 데 방해가 되거나 오류를 발생시킬 수 있는 노이즈 데이터를 제거하는 정제과정과 의미가 같은 단어들을 통합하는 데이터 정규화 과정을 진행하였다. 본 연구에서 정제된 데이터들은 단어들 간의 관계를 분석하기 위해 단어와 단어의 1-mode의 공출현 매트릭스(co-occurrence matrix)로 구성되었으며, 이러한 매트릭스 데이터는 행과 열이 같은 단어들로 구성된 전형적인 매트릭스 형태이다.

1. 키워드 빈도분석 결과

텍스트마이닝에서 키워드 또는 핵심어를 추출하는 방법 중에서 TF(단어

서 집합의 추상적인 주제를 발견하기 위한 통계적 모델임. 특히, 텍스트 본문의 숨겨진 의미 구조를 파악하기 위해 사용되는 텍스트마이닝 기법 중 하나이다.

빈도, Term Frequency)는 문서 내 특정 단어의 빈도를 의미하고, IDF(역문서 빈도, Inverse Document Frequency)는 DF(Document Frequency)의 역수로써, DF는 한 단어가 전체 문서 집합 내에서 얼마나 공통적으로 많이 등장하는지를 나타내는 값을 의미한다. 본 연구에서‘다문화’를 중심으로 한 상위 25개의 단어 빈도(TF) 분석 결과는 다음의 [그림 2]와 같다. 단어 빈도분석 결과를 살펴보면, 코로나 발생과 상관없이 다문화 가족의 삶과 관련한 ‘가족’, ‘가정’, ‘센터’, ‘건강’, ‘사회’, 그리고 ‘이주민’ 등과 같은 단어들의 빈도가 높았고, ‘교육’, ‘학교’, ‘학생’, ‘체험’, ‘프로그램’, ‘자녀’ 등과 같은 다문화 자녀 교육 관련 키워드들도 다수를 구성하고 있는 것으로 분석되었다. 한편, 코로나 발생 이전에는 ‘체험(1.53%)’이나 ‘소통(1.17%)’ 등과 같은 단어들이 나타나고 있는 반면, 코로나 발생 이후에는 ‘건강(3.50%)’, ‘뉴스(1.20%)’이라는 키워드의 빈도가 늘어나고 있다. 또한, 코로나 이전에는 ‘지원’의 빈도가 3.61%로 상당히 높은 빈도를 보였으나, 코로나 이후에는 ‘지원’이라는 키워드의 빈도가 매우 낮아져 기존의 상당수의 다문화 관련 지원 정책들이 부재함을 확인할 수 있었다. 또한 코로나 발생 이후에는 대부분의 초중고 학교들이 온라인 개학을 시작함에 따라 ‘온라인(1.30%)’과 ‘마스크(1.25%)’라는 코로나 팬데믹 현상과 관련한 키워드가 추가적으로 증가하고 있다.



(A) 코로나 발생 이전 (B) 코로나 발생 이후

[그림 2] 코로나 발생 전후의 키워드 단어빈도 파이차트

또한 TF-IDF는 주로 문서의 유사도를 구하는 작업이나 검색 시스템에서 검색 결과의 중요도를 정하는 작업, 그리고 문서 내에서 특정 단어의 중요도를 구하는 작업 등에 주로 활용되는 데, 본 연구에서 키워드들 간의 TF-IDF와 연결 중심성을 측정한 결과는 다음의 [표 2]와 같다. TF-IDF와 연결 중심성의 분석 결과를 살펴보면, 코로나 발생 이전이나 이후에 '다문화,' '가족', '가정', '센터' 그리고 '교육'에 대한 연결 중심성은 모두 높게 나타나고 있다. 한편, 코로나 발생 이전에는 '지원'이라는 단어의 빈도와 TF-IDF는 높은 반면 연결 중심성은 낮게 나타났으며, 코로나 발생 이후에는 '코로나', '온라인', 그리고 '마스크' 등과 같은 국내 코로나 19 팬데믹 상황의 특수성을 반영한 단어들의 키워드 빈도는 상대적으로 높게 나타난 반면, TF-IDF와 연결 중심성이 낮게 나오고 있다. 일반적으로 TF-IDF는 텍스트마이닝에서 활용하는 가중치 알고리즘(algorithm)으로, 여러 문서로 구성된 문서 군에서 어떤 특정 단어가 각각의 특정 문서 내에서 어느 정도 중요한 것인지를 측정하는 통계적 분석 수치로써, '코로나', '온라인', 그리고 '마스크' 등과 같은 코로나 관련 단어들은 특정 문서 내에서 단어 빈도가 낮고, 전체 문서상으로는 이러한 단어를 포함한 문서가 많아서 TD-IDF 값은 작게 나온 것으로 해석할 수 있다.

또한 연결정도 중심성(degree centrality)은 네트워크 구조에서 한 노드에 직접적으로 연결되어 있는 노드의 개수를 의미하는 것으로, 노드의 수가 많을수록 연결 중심성이 높다. 본 연구에서 특정 단어와 직접적으로 연결된 기타 단어의 수를 기준으로 하면, '코로나', '온라인', 그리고 '마스크' 등과 같은 단어들이 특정 단어와 직접적으로 연결된 관계가 적고 서로 독립적으로 사용되어 연결 중심성이 낮게 나온 것으로 분석할 수 있다.

코로나 발생 이전 (2019년 7월 21일 ~ 2020년 1월 20일)			코로나 발생 이후 (2020년 1월 21일 ~ 2020년 7월 20일)		
키워드	TF-IDF	연결중심성	키워드	TF-IDF	연결중심성
가족	9,309.2	0.0874	**가족**	9,309.2	0.0874
가정	7,364.3	0.0713	**가정**	7,364.3	0.0713
센터	7,586.4	0.0950	**센터**	7,586.4	0.0950

교육	7,219.6	0.0815	교육	7,219.6	0.0815
사회	5,599.2	0.0642	코로나	65.1	0.0007
지원	5,453.6	0.0030	건강	4,583.1	0.0329
건강	4,583.1	0.0329	학생	4,376.7	0.0389
지역사회	3,796.3	0.0432	대상	3,489.1	0.0450
이주민	3,882.7	0.0315	지역사회	3,796.3	0.0432
여성	3,946.2	0.0467	사회	5,599.2	0.0642
학생	4,376.7	0.0389	자녀	3,793.2	0.0301
한국	3,750.3	0.0639	이주민	3,882.7	0.0315
대상	3,489.1	0.0450	여성	3,946.2	0.0467
장학재단	4,275.8	0.0193	외국인	3,925.2	0.0361
자녀	3,793.2	0.0301	프로그램	2,890.6	0.0405
외국인	3,925.2	0.0361	교육청	3,090.7	0.0187
체험	3,812.1	0.0459	한국	3,750.3	0.0639
학교	3,528.7	0.0316	전달	1,885.0	0.0253
소통	3,135.9	0.0169	온라인	121.5	0.0017
결혼이주	2,882.0	0.0359	장학재단	4,275.8	0.0193
청소년	3,262.3	0.0283	마스크	455.9	0.0049
뉴스	2,791.9	0.0624	뉴스	2,791.9	0.0624
프로그램	2,890.6	0.0405	학교	3,528.7	0.0316
교육청	3,090.7	0.0187	결혼이주	2,882.0	0.0359
축제	3,054.5	0.0295	서비스	2,405.9	0.0152
실시	2,555.4	0.0348	청소년	3,262.3	0.0283
이중언어	2,859.0	0.0245	학부모	2,348.7	0.0209
복지	2,637.8	0.0199	한국어	1,770.5	0.0173
서울	2,438.7	0.0397	복지	2,637.8	0.0199

[표 2] 코로나 발생 전후 TF 키워드의 TF-IDF와 연결 중심성

2. CONCOR 분석 결과

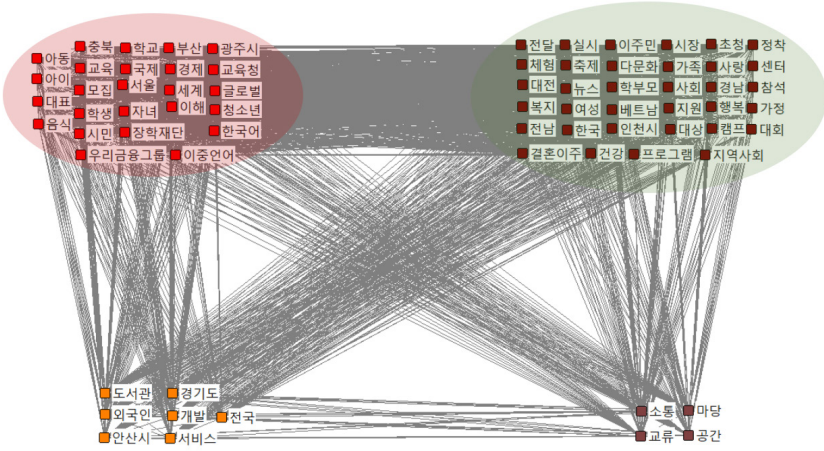
본 연구에서는 도출된 주요 키워드들 사이에 상관계수를 기반으로 구조적 등위성을 확인하기 위해 CONCOR 분석을 실시하였다. CONCOR 분석은 키워드들 간의 상관관계를 토대로 구성요소들 간의 관계를 파악하고 적정한 수준의 유사성을 찾아내어 구조적 등위성을 측정하는 방법으로, 행위자간 관계 패턴의 도출을 위해 행위자들 간 상관관계(correlation)를 사용한다. 여기서 구조적 등위성은 한 네트워크에서 다른 행위자들과 직접적인 관계는

없지만 동일한 관계패턴을 가지는 경우로, 구조적 등위성 분석은 유사한 지위를 가진 행위자들을 블록(block)화하고, 그러한 군집들 간의 묘사하는 방식이다. CONCOR 분석은 단어들 간의 유사성과 연관성을 기준으로 관계가 높은 단어들을 하나의 그룹으로 블록화하고, 중심성 지수 분석을 통하여 각각의 네트워크 구조에서 개별 단어들이 가지고 있는 영향력이나 중요성을 파악하게 된다.

본 연구에서는 '다문화'라는 키워드를 중심으로 UCINET의 넷드로우(Netdraw)를 활용하여 CONCOR 분석을 수행하였다. 우선, 본 연구에서 CONCOR 분석을 수행하기 전에 덴드로그램(dendrogram)을 통해 단계별 군집 수를 확인하였고, 덴드로그램 결과를 기반으로 2개의 속성을 선택하여 군집화하였다. 또한 클러스터 구성 노드들 간의 분산배치 정도를 설정하는 스크런치(Scrunch) 요인을 '8'로 지정하여 노드 간의 간격을 조정하여 CONCOR 분석을 시행하였고, CONCOR 분석 결과를 기반으로 중심 클러스트(cluster, 군집)과 주변 클러스트로 구분하였다.

우선 코로나 발생 이전 시점의 CONCOR 분석 결과는 다음의 [그림 3]과 같이 크게 2개의 중심 군집과 2개의 주변 군집으로 분류할 수 있다. 다문화 삶과 관련된 중심 군집은 총 34개의 키워드로 구성되어 있고, 다문화 자녀들의 교육과 관련된 키워드는 총 25개로 클러스트화되었다. 다문화 삶과 관련된 주요 키워드로는 결혼이주, 지역사회, 여성, 사회정착, 이주민, 지원 프로그램과 같은 키워드들이 있고, 다문화 자녀의 교육과 관련된 키워드로는 청소년, 아동, 교육, 학생, 학교, 이중 언어 등이 대표적인 키워드들이다. 한편, 주변 클러스트에는 소통, 교류, 공간, 마당이라는 그룹과 경기도, 서비스, 도서관, 외국인, 안산시, 개발이라는 주변 키워드가 군집화되었다.

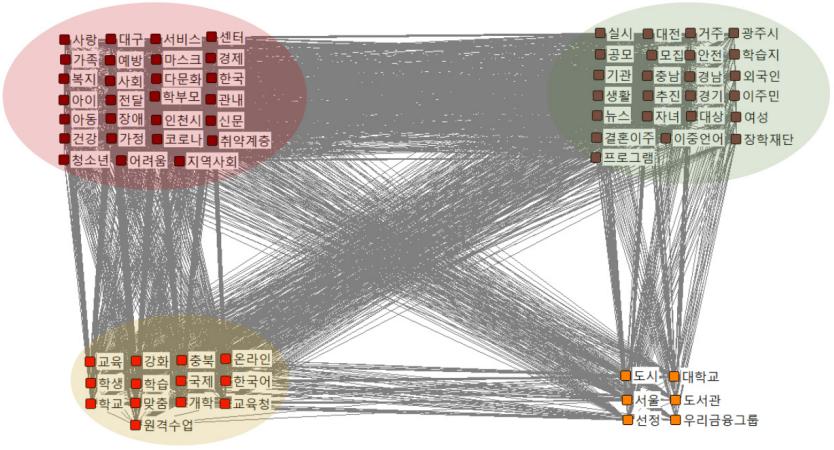
한편, 다음의 [그림 4]와 같은 코로나 발생 이후의 CONCOR 분석 결과를 살펴보면, 코로나 발생 이후에는 크게 세 개의 중심 군집과 한 개의 주변 군집으로 분류되었다. 코로나로 인한 어려움이나 취약성을 보여주는 다문화



[그림 3] 코로나 발생 전의 다문화 인식에 대한 CONCOR 문석 시각화 결과

가정의 삶 관련 중심 클러스트는 총 27개의 키워드가 군집화되었고, 일반적인 다문화 사회 관련 키워드는 총 24개가 군집화되었다. 또한 코로나 발생 이후의 다문화 가족의 자녀 교육과 관련한 키워드는 13개가 추출되었다. 주변 그룹으로는 우리금융그룹, 도서관, 대학교, 도시, 서울, 선정 등과 같은 단어들이 주변 군집을 형성하고 있다. 코로나 발생으로 인하여 다문화 가족이 경험하게 되는 취약한 삶 관련 키워드로는 경제, 취약계층, 어려움, 예방, 마스크 등의 단어가 추가되었고, 코로나로 인하여 다문화 자녀들이 겪는 원격수업, 온라인, 개학, 맞춤 등과 같은 키워드들이 하나의 군집으로 그룹화되었다.

하나의 군집으로 분류되었는데, 이러한 결과는 코로나 발생에 따른 온라인 개학과 원격 수업의 증가와 이로 인한 교육의 공백 문제가 코로나 상황 하에서의 다문화 가정에서도 중요한 문제로 부각되고 있음을 제시하고 있다. 또한 다문화 학생의 수준별 맞춤 학습이나 교육의 필요성, 그리고 교육청의 역할 강화 등이 요구되고 있는 상황이다. 한국여성정책연구원(2020)은 코로나 발생 이후에는 특히, 다문화 가정의 학생들에 대한 교육은 또 다른



[그림 4] 코로나 발생 후의 다문화 인식에 대한 CONCOR 분석 시각화 결과

「코로나 19로 인한 가족의 변화와 정책과제」의 토론회[6]에서 코로나 19의 확산이 영향을 미치는 가족 생활에서 가장 심각한 부분 중의 하나로 초중고 학교 및 각종 보육 시설의 휴원에 따른 자녀들의 돌봄 문제라고 지적하고 있다. 또한 코로나 시대에 가장 우선적으로 해결해야 할 정책 과제로 돌봄 공백에 대한 적절한 대응이 필요함을 제시하고 있다. 이와 같은 자녀들의 돌봄 공백의 문제는 국내의 일반 가정뿐만 아니라 다문화 가정에서도 중요한 이슈로 부각되고 있으며, 본 연구에서의 CONCOR 결과처럼 코로나 발생 이후에 다문화 가정의 자녀 교육과 관련한 이슈들이 가장 중요하고 시급한 대책이 요구되는 다문화 정책의 운영 방안임을 제시하고 있다.

3. 코로나 발생 전후의 언어 감성단어 분석

감성 분석(sentiment analysis)은 텍스트 안에 나타난 부정과 긍정을

6) 한국여성정책연구원은 '코로나19로 인한 가족의 변화와 정책과제'라는 주제로 코로나 관련 여성·가족 분야별 릴레이 토론회를 진행하였고, 이 토론회에서는 코로나19로 인한 돌봄 공백 문제, 경제적 상황 변화, 가족관계 스트레스 등을 파악하여 이러한 문제를 해결하기 위한 관련 대응 정책 수립을 위한 토론회임.

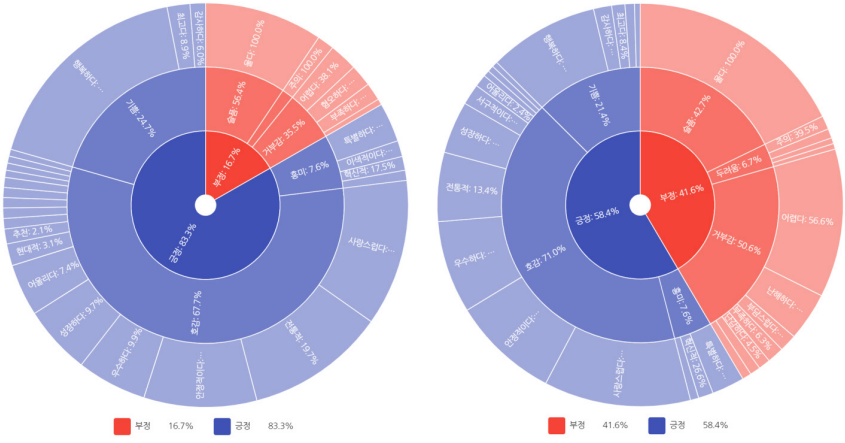
구분하는 극성 분석이고, 사람들의 슬픔이나 기쁜, 분노와 같은 다양한 감성과 태도나 의견, 성향 등과 같은 사람들의 감성에 대한 주관적인 자료를 분류하는 자연어 처리 방식으로 대표적인 오피니언 마이닝(opinion mining) 기법 중의 하나이다. 감성 분석은 우선 감성어 사전을 구축하고 규칙을 개발하는 방법으로, 사람들의 감성을 표현하는 단어를 찾아 분류하고 감성어 사전을 구축한 후에 개별 비정형 텍스트 데이터를 분해하여, 이렇게 분해된 단어를 사전에 미리 준비한 감성어 사전과 매칭(matching)하여 각각의 개별적 단어들의 감성을 판정하는 방법이다.

본 연구에서의 감성단어 빈도분석은 원문 데이터 안에 내제된 감성과 관련된 키워드를 중심으로 이러한 감성단어들의 빈도를 객관적이고 체계적으로 수치화하는 것으로 본 연구에서는 텍스톰(Textom)에서 자체 제작된 감성어 어휘사전을 기반(lexicon-based approach)으로 감성언어 분석을 수행하였다. 텍스톰의 감성어휘 사전은 긍정 또는 부정이라는 단어의 카테고리 안에 긍정의 키워드는 흥미, 호감, 그리고 기쁨의 단어가 있고, 부정의 키워드에는 통증, 슬픔, 분노, 두려움, 놀람, 그리고 거부감이라는 6개의 단어가 포함되어 있다.[7] 이러한 감성단어 빈도는 시각화 결과를 통해 키워드 빈도뿐만 아니라 감성단어의 강도와 세부 감성 분석 시각화를 동시에 제공하기 때문에 감성언어 해석의 다양한 인사이트(insight)를 도출할 수 있다. 즉, 본 연구에서는 감성어 어휘사전 기반의 접근 방법은 텍스톰(Textom) 5.0 시스템 내에 구축되어 있는 사전에 정의된 감성사전과 분석 대상 텍스트의 감성어와의 출현 빈도 매칭을 통해 긍정과 부정 등의 감성을 분류하는 접근 방법이다.

다음의 [그림 5]는 코로나 발생 이전과 이후 시점에서 다문화와 관련한 단어들의 감성단어 분석의 결과이다. 코로나 발생 이전에는 긍정적 단어의 비중이 부정적 단어의 약 5배에 달할 만큼 긍정적 단어들이 많았으나, 코로나 발생 이후에는 긍정적 단어가 약 58.4%이고 부정적 단어가 41.6%로 긍정적

7) 본 내용은 텍스톰(www.textom.co.kr)의 감성분석 100% 활용법을 기반으로 작성한 내용임.

단어가 부정적 단어의 약 1.4배 비율로 부정적 단어가 급격히 증가하였음을 확인할 수 있다.



(A) 코로나 발생 이전 (B) 코로나 발생 이후

[그림 5] 다문화 중심의 단어 감성분석

[그림 5]의 감성단어의 빈도를 보다 자세히 살펴보면, 코로나 발생 전에는 중분류에서 기쁨 관련 단어가 20.57%였으나, 코로나 발생 이후에는 12.50%로 대폭 감소하였음을 알 수 있고, 호감 관련 단어도 56.39%에서 41.46%로 대폭적으로 감소하였음을 확인할 수 있다. 반면에, 슬픔으로 중분류된 부정적 감성어는 코로나 발생 이전에는 9.42%이었으나, 코로나 발생 이후에는 거의 두 배 가까이 증가한 17.76%를 보이고 있다. 또한 거부감 관련 단어는 코로나 발생 이전에는 5.93%였으나, 코로나 발생 이후에는 21.05%로 폭발적으로 증가하여 약 4배에 가까운 증가 폭을 보이고 있다. 이와 같이 다문화에 대한 부정적 단어의 급증은 다양한 언론매체들이 신종 코로나 바이러스의 발생이 중국을 비롯한 아시아 지역에서 발생하고 확산되었다는 부정적 보도가 많아짐에 따라서, 다문화 이주민/이주자에 대한 거부감이 급격히 증가한 것으로 추정할 수 있다. 또한 현실적으로 다문화 가족을 포함한 대다수의 국민들이 코로나 발생으로 인한 사회적 거리두기의 강화로 인하여 전반적인

우울과 불안, 그리고 스트레스의 과중으로 인하여 슬픔이나 두려움, 공포와 같은 부정적 단어들이 많이 도출된 것으로 판단할 수 있다.

또한, 추가적으로 코로나 발생 전후시기에 다문화를 중심으로 어떠한 감성단어들이 많이 도출되어 있는가를 시각화한 단어 워드 클라우드(word cloud) 결과를 살펴보면, 다음의 [그림 6]과 같다. 워드 클라우드의 결과를 살펴보면, 코로나 발생 이전에는 '사랑스럽다', '행복하다', '안정적이다' 등의 긍정적 단어의 빈도가 높은 반면, 코로나 발생 이후에는 '울다', '어렵다' 등의 부정적 단어의 빈도가 높게 나타나고 있다. 또한 코로나 발생 이후에는 기존의 '행복하다'와 '사랑스럽다'와 같은 긍정 감성어들의 빈도가 다소 축소되어 있음을 가시적으로 확인할 수 있다. 한편, 코로나 발생 이후에는 '난해하다', '걱정하다', '위축되다', '부담스럽다' 그리고 '난감하다'라는 부정적 단어들이 추가적으로 도출되었다. 이상과 같은 결과는 코로나 발생 이전에는 긍정적 단어가 많고, 코로나 발생 이후에는 부정적 단어가 많다는 동국대 이주다문화통합연구소의 「코로나19 전후 다문화 관련 언론 보도 양상」 보고서[8]의 결과와 매우 유사한 결과를 보이고 있다. 이와 같은 결과는 다문화 가정의 이주민들이나 다문화 가족 구성원들이 코로나 발생 이전에 비해 코로나 발생 이후에 한국에 거주하고 있는 환경이 현실적으로 매우 열악한 상황이며, 현실적으로 한국에서 다문화 이주민을 대한 시선들이 다소 부담스럽고 위축되어 있음을 제시하고 있다.

8) 동국대 이주다문화통합연구소에서는 2019~2020년 국내 주요 언론사 54곳에서 보도한 다문화 기사 4천여 개를 바탕으로 코로나 19를 기점으로 다문화를 중심으로 한 감성단어를 분석하였음. 분석 결과에 따르면 다문화 이슈와 관련한 '호감'은 2019년 51.5%에서 2020년 38.2%로 13.3%포인트, '기쁨'도 같은 기간 13.7%에서 10.3%로 3.4%포인트 각각 줄었다. 반면 같은 기간 거부감(9.7%→19.3%)과 두려움(1.1%→2.1%)은 각각 두 배 정도 증가했다고 보도함. 연구를 진행한 김동진 동국대 이주다문화통합연구소 초빙교수는 "최대한 정제된 표현을 쓰는 언론 보도에서조차 코로나19를 기점으로 다문화와 관련해 거부감과 두려움 등의 표현이 증가한 것을 볼 때 대중 인식은 더 나빠졌을 가능성이 있다"며 "차별과 편견을 줄일 방안을 마련하고 다문화 이해를 돕는 인식 개선 프로그램을 개발하는 것이 필요하다"고 제안했음. (연합뉴스, 2021년 3월 9일자).

(A) 코로나 발생 이전 (B) 코로나 발생 이후

[그림 6] 다문화 중심의 감성 단어 워드 클라우드

V. 맺음말

2019년 말부터 시작되었던 신종 코로나 바이러스 감염증은 지난 해 우리 사회의 많은 것들을 변화시켰으며, 현재도 진행 중인 전 지구적 사회 재난이다. 이러한 사회적 재난 상황 하에서 사회적 약자 계층 중의 하나인 다문화 가정들은 사회적 편견뿐만 아니라 경제적인 어려움과 고통에 직면해 있다. 특히, 코로나 19 시기에 대부분의 초중고 학교들이 온라인 개학으로 전환됨에 따라 다문화 가정에서의 돌봄 공백이 새로운 문제로 대두되고 있고, 재택근무가 확산됨에 따라 해고나 무급 휴직 등과 같은 일자리 부족 현상이 심화되고 있다. 또한 코로나 19로 인하여 다문화 가족에 대한 사회적 편견이나 배제 그리고 차별이 가중되고 있으며, 일부 다문화 가정 중에 의료 사각지대에 놓인 외국인 이주민들은 의료 공백이라는 심각한 문제에 직면하면서 사회적 소외계층으로 전락하는 등의 다문화 가정을 둘러 싼 내·외부적 환경에 많은 변화나 나타나고 있다. 이와 같이 다문화를 둘러 싼 외부환경 변화 속에서도 다문화 사회의 기본적인 공동체 가치와 시민적, 합리적 규범 그리고 평등 이념은 다문화주의의 가치를 실현하기 위해 매우 중요하다. 즉, 다문화 사회 속에서 다양성의 가치와 더불어 문화 간 이해, 그리고 관점과 전개 방식의 공유는 다문화 사회가 지향해야 하는 매우 중요한 가치이다. 따라서 포스트 코로나 팬데믹(post Corona-pandemic)

시대를 대비하여 한국에서 다문화 이주민의 삶의 질을 높이고 다문화 운영 정책의 바람직한 방향성 정립을 위해 탄력적이고 다각적인 다문화 정책 운영 방안이 필요하다.

따라서 본 연구는 코로나-19라는 전세계적 팬데믹 현상에 의해 다문화의 인식이 시점별로 어떻게 변화되고 있는가를 비교 추적한 연구로서 다문화 사회에서 외부적 환경 요인의 변화에 따라 다문화가 지향하고 있는 다양한 가치를 측정했다는 측면에서 연구의 의의가 있다. 즉, 본 연구에서는 코로나 발생 전후 시점에 다문화에 대한 인식 변화를 살펴보기 위해 '다문화' 관련 텍스트 자료를 수집하였고, 수집된 출현빈도 텍스트 자료를 활용하여 각 추출 단어들 간의 중심성 분석과 감성 언어분석을 수행하였으며, 추가적으로 CONCOR 분석을 통해 각 단어들의 군집분석을 수행하였다.

본 연구의 연구 결과로부터 도출된 시사점은 다음과 같다. 첫째, 언어 감성분석의 결과를 살펴보면, 코로나 발생 전에는 긍정적 의미의 단어들이 많이 도출되었던 반면에 코로나 발생 이후에는 부정적 의미의 단어들이 많아지고 있음을 확인할 수 있었다. 국내에 코로나라는 사회 전반의 큰 변화에 따라 다문화 사회에서 이주민 또는 이민자들에 대한 감성언어들이 크게 변동될 수 있음을 제시하며, 향후 다문화 사회 속에서 이들을 보호하고 함께 살아나가기 위한 전략적 접근이 필요함을 제시하고 있다. 둘째, CONCOR 분석결과를 살펴보면, 코로나 발생으로 인하여 다문화 관련 핵심어들 간의 클러스트에 변화가 발생하였다. 즉, 코로나 19라는 외부환경 변화가 기존의 우리 사회의 자녀들에 대한 교육 영역뿐만 아니라 다문화 가정에서의 교육 영역에서도 새로운 변화가 발생하였고, 이러한 새로운 영역에 대한 보다 세부적인 이해와 추가적인 지원이 필요함을 제시하고 있다.

본 연구는 이와 같은 연구 성과에도 불구하고, 다음과 같은 연구의 한계도 내제하고 있다. 첫째, 본 연구에서는 다문화라는 키워드 텍스트마이닝을 통해 다양한 분석을 진행해 왔지만, 다문화라는 대표적인 키워드 이외에 다문화 가족이나 다문화 사회 등을 대변할 수 있는 다양한 키워드들을 반영하지

못한 한계가 있다. 따라서 추후 연구에서는 다문화를 대표할 수 있는 보다 다양한 키워드를 개발하여 보다 세밀하고 정교한 분석이 필요하다. 또한, 본 연구에서는 키워드 네트워크의 거시적인 구조 효과는 살펴볼 수 있었지만 거시적 구조 내의 개별 행위자들의 효과를 통계적으로 검증하지 못한 한계가 있다. 따라서 향후 추가적인 연구에서는 네트워크 구조 내의 다양한 행위자들의 미시적인 관계성 구조분석을 통해 보다 상세한 네트워크 분석을 수행하고자 한다.

한성대학교 미래융합사회과학대 경영학부 정교수, khchoi@hansung.ac.kr

주제어(Key words):
코로나-19(COVID-19), 다문화(multicuture), 빅데이터(big data), 감성분석(Sentiment Analysis), 구조적 등위성 분석(Structural Equivalence Analysis)

투고일: 2021.04.11, 심사일: 2021.04.25, 게재확정일: 2021.05.03.

<국문 초록>

텍스트마이닝을 통한 코로나 발생 전후 시기의 다문화에 대한 인식변화

최 강 화

전 세계를 공포에 떨게 한 코로나 19 펜데믹은 지난해에 이어 올해까지 한국 사회의 여러 분야에 많은 부정적 영향을 미치고 있다. 특히, 이러한 외부 환경변화는 국내 다문화 관련 정책에도 많은 변화가 필요함을 제기하고 있다.

본 연구에서는 코로나라는 외부적 환경 변화에 따라 다문화와 관련한 핵심 키워드들에 어떠한 질적 변화가 있었는가를 추적했다. 즉, 코로나 발생 전후 시점에 다문화와 관련된 키워드들의 변동을 측정하고, 감성단어들의 변동을 측정하였다. 또한 본 연구에서는 CONCOR 분석을 통해 코로나 전후 시점의 다문화 관련 키워드들의 군집 변동을 측정하였다. 이와 같은 분석결과를 토대로 본 연구에서는 코로나 19와 같은 외부적 환경 변동에 따라 다문화 관련 정책들을 보다 효율적으로 관리할 수 있는 정책적 방향을 제시한다. 또한 포스트 코로나 팬데믹 시대를 대비하여 한국에서 다문화 이주민의 삶의 질을 높이고 다문화 운영 정책의 바람직한 방향성 정립을 위해 탄력적이고 다각적인 다문화 정책 운영 방안을 제시한다.

<Abstract>

A Study on the Change in Perception of Multiculturalism Before and After Corona Outbreak Using the Text Mining

Choi, Kanghwa

Corona-19 Pandemic, which has terrorized the world, has had a negative impact on various fields of Korean society until now. This study was conducted to investigate the effects of external environmental changes such as Corona-19 on the qualitative changes of critical keywords focused on multiculturalism. That is, this study measured the changes in keywords related to multiculturalism before and after Corona-19 occurrence and investigated the changes in sentimental words. In addition, this study traced the variation of clusters on multi-cultural keywords before and after Corona-19 occurrence using the CONCOR analysis. Based on the results of this analysis, this study suggests the strategic initiatives of multicultural policy that can more effectively cope with external environmental changes such as Corona-19. In preparation for the post-corona-19 pandemic era, this study suggests flexible and multi-faceted multicultural policy plans to improve the quality of life for multicultural immigrants in Korea and establishes desirable direction for multicultural policy initiatives.

참 고 문 헌

1. 단행본

곽기영, 『소셜네트워크 분석』 (서울: 청람, 2014).
김용학, 김영진, 『사회 연결망 분석』 (서울: 박영사, 2016).
데렉 한센, 벤 슈나이더마, 마크 스미스, 권상희 역, 『노드엑셀을 이용한 소셜 미디어 네트워크 분석』 (서울: 컴원미디어, 2019).
딘 러셔, 요한 코스키넨, 개리 로빈스, 최수진 역, 『사회 네트워크 통계 모형(ERGM)』 (파주: 한울, 2020).
Cherven, K., *Mastering Gephi Network Visualization: Produce advanced network graphs in Gephi and gain valuable insights into your network datasets* (BIRMINGHAM - MUMBAI: PACKT Publishing, 2015).
Wasserman, S., and Faust, K., *Social Network Analysis: Methods and Applications (Vol. 8)* (Cambridge: Cambridge University Press, 1994).

2. 논문

강진구, 「텍스트마이닝 기법을 통해 본 〈다문화콘텐츠연구〉의 연구 경향 분석」, 『다문화컨텐츠연구』, 제32권(2019).
강진구, 이기성, 「텍스트마이닝(Text Minning)을 통해 본 제주 예멘 난민: 네이버 뉴스 댓글을 중심으로」, 『다문화컨텐츠연구』, 제30권(2019).
김세현, 「비정형자료분석을 통해 살펴본 한국의 다문화 연구」, 『한국인구학』, 제41권 1호(2018).
김수정, 마경희, 윤성은, 「다문화가족지원센터의 코j로나19 대응 및 과제 탐색-센터장 대상 포커스 고룹 인터뷰를 중심으로」, 『생명연구』, 제58집(2020).

김용희, 「소셜네트워크분석(Social Network Analysis)기법의 이해와 적용: 네트워크 구조와 클러스터링 그리고 QAP」, 『Korea Institute of Public Administration』, 제34집(2020).
김태종, 「뉴스 빅데이터를 활용한 코로나19 언론보도 분석: 토픽모델링 분석을 중심으로」, 『한국콘텐츠학회논문지』, 제20권 5호(2020).
안명숙, 「빅데이터를 활용한 다문화 핵심단어 및 네트워크 분석」, 『융복합지식학회논문지』, 제6권 2호(2018).
윤희진, 「텍스트마이닝을 활용한 다문화 멘토링 관련 연구 동향 분석」, 『문화교류와 다문화교육』, 제9권 1호(2020).
이수상, 「언어 네트워크 분석 방법을 활용한 학술논문의 내용분석」, 『정보관리학회지』, 제31권 4호(2014).
이수정, 최두영, 「사회과학을 위한 양적 텍스트마이닝: 이주, 이민 키워드 논문 및 언론기사 분석」, 『한국콘텐츠학회논문지』, 제20권 5호(2020).
이주호, 「다문화주의 관점에서 COVID-19 재난지원 동향에 관한 탐색적 고찰」, 『한국민간경비학회보』, 제19권(2020).

3. 연구/정책 보고서

한국여성정책연구원, 코로나19로 인한 가족의 변화와 정책과제, 제3차 코로나 관련 여성·가족 분야별 릴레이토론회(2020).
우춘희, 코로나19로 인한 국경 이동의 제한이 이주민의 삶에 미치는 영향: 캄보디아 이주농업노동자 사례를 중심으로, 서울특별시 청년허브(2020).