한국빅데이터학회지
제7권 제2호, 2022, pp. 103-111 https://doi.org/10.36498/kbigdt.2022.7.2.103

감염병 위기 대응을 위한 소셜 데이터 수집 및 적재 엔진 기반 신뢰도 분석 시스템 개발

Development of Social Data Collection and Loading Engine-based Reliability analysis System Against Infectious Disease Pandemic

정두영[1] · 이상준[1] · 민경일[1] · 정석송[1] · 한현욱[2,3,4,5†]

미소정보기술[1], 차의과학대학교 의학전문대학원 정보의학교실[2],
차의과학대학교 의학전문대학원 정보의학연구소[3], 차미래의학연구원 데이터사이언스연구센터[4],
분당차병원 의료정보빅데이터센터[5]

요 약

감염병 대응과 관련된 기관, 조직, 사이트 등의 다수 운영되고 있으나 코로나-19와 같은 팬더믹 상황이 수년간 지속됨에 따라 초기양상과 현재 양상의 수많은 변화가 있으며 이에 따른 정책과 대응체계도 진화하고 있다. 이에 따른 지역별 격차가 발생하고 정책에 대한 신뢰와 불신, 이행도에 따른 여러 가지 문제들이 산재해 있다. 따라서 본 연구에서는 정보전염이 포함된 소셜 데이터를 분석하는 과정에서 루머가 포함된 데이터를 수집하는 과정에서 팩트 체크가 되는 언론 매체와 다르게 정확한 출처를 알 수 없는 부정확한 정보들이 포함되는 주요 소셜 미디어 플랫폼 중의 하나인 트위터 데이터를 수집하여 사실과 무관한 내용을 사전 차단하는 시스템을 개발했다. 비정형데이터인 소셜데이터를 기반으로 감염병 위협을 자동 감지할 수 있는 알고리즘을 개발하여 감염병 위기 대응과 관련된 객관적인 근거를 창출함으로써 관련 분야 국제경쟁력을 공고히 하고자 한다.

■ 중심어 : 텍스트 마이닝, 머신 러닝, 소셜 데이터, 크롤링 시스템

Abstract

There are many institutions, organizations, and sites related to responding to infectious diseases, but as the pandemic situation such as COVID-19 continues for years, there are many changes in the initial and current aspects, and accordingly, policies and response systems are evolving. As a result, regional gaps arise, and various problems are scattered due to trust, distrust, and implementation of policies. Therefore, in the process of analyzing social data including information transmission, Twitter data, one of the major social media platforms containing inaccurate information from unknown sources, was developed to prevent facts in advance. Based on social data, which is unstructured data, an algorithm that can automatically detect infectious disease threats is developed to create an objective basis for responding to the infectious disease crisis to solidify international competitiveness in related fields.

■ Keyword : Text mining, Machine learning, Social data, Crawling system

2022년 11월 18일 접수; 2022년 12월 07일 수정본 접수; 2022년 12월 08일 게재 확정.

* This research was supported by a grant of the Information and Communications Promotion Fund through the National IT Industry Promotion Agency (NIPA), funded by the Ministry of Science and ICT (MSIT), Republic of Korea (No. S2002-21-1003).

† 교신저자 (stepano7@gmail.com)

Ⅰ. 서 론

1.1 감염병 국내외 현황

1.1.1 감염병 확산과 트렌드

새롭게 출현하는 코로나-19(Coronavirus)와 같은 팬더믹 상황이 수년간 지속됨에 따라 초기양상과 현재 양상의 수많은 변화가 있었고 이에 따른 정책과 대응체계 또한 진화하고 있다. 이렇게 새롭게 출현하는 감염병은 신종감염병으로 정의되며 발생과 확산에는 인간의 생물학적인 요인뿐만 아니라 기후, 생태계, 보건의료체계와 같은 사회, 경제적 요소들이 복합적으로 작용한다[1,2]. 근래 이슈가 되고 있는 코로나-19 발생에 대해 세계보건기구(WHO, World Health Organization)는 단순 전 세계적 확산사태로 인한 공중보건적 요소뿐만 아니라 경제, 사회적 요소에 영향을 미치는 심대한 위기이며, 전세계적 차원에서의 대응을 촉구한다[3]. 특히 신종감염병의 과반수 이상을 차지하는 인수공통감염증(Zoonoses)는 변이성과 적응력이 커서 출현 시기와 인체에의 영향을 예측하기 어렵고, 평상시에 대비 및 대응 계획을 세우고 필요한 체계를 구축하며 자원을 확보하고 인적 역량을 강화하여 신종감염병 위기 발생시 신속하고 효과적으로 대응함으로써 인명 및 사회 경제적 피해를 최소화해야 한다고 말한다[4,5].

1.1.2 국외 현황

1900년대 초반 미국에서 발생한 스페인 독감으로 인해 전 세계적으로 약 4천만 명의 사상자가 발생했다. 1950년대 아시아 독감은 중국에서 시작되어 전 세계적으로 1천 4백만 명의 사망자를 낳았으며, 이후 유행한 홍콩 독감의 경우 전 세계적으로 140만 명의 사상자가 발생했다. 20세기 이후 위생의 개선을 비롯한 환경적 변화와 항생제 및 백신 개발 등 과학 기술적 발전으로 상당한 수준의 감염병을 극복하였으나, 21세기에 들면서 현대화, 세계화가 가속화되고 교통의 발달로 인해 이동량이 급격히 증가하였으며, 기후 변화와 같은 환경적 변화로 인해 기존에 없던 신종감염병이 등장하기 시작했다.[6] 2000년 이래 대표적으로, 전 세계는 2003년 사스(SARS, Severe acute respiraoty syndrome)를 시작으로 2009년 A형 인플루엔자 바이러스(H1N1, Influenza A virus subtype H1N1), 2014년 에볼라(Ebola), 2015년 메르스(MERS, Middle East Respiratory Syndrome), 2017년 고병원성 조류인플르엔자(H5N1, Influenza A virus subtype H5N1)을 겪었다[7-9]. 특히, 2012년 겨울, 중국에서 시작되어 전세계로 확산되어 2014년 에볼라 발생 이후 세계보건기구는 전염병에 대해 9가지 기준을 수립했다; 인간으로의 전염, 치사율, 전염 가능성, 진화 가능성, 의료 대응 조치, 감시 및 통제의 어려움, 영향을 받는 지역의 공공 의료 시스템, 국제적 확산의 위험, 사회에 미치는 영향[10].

1.1.3 국내 현황

국내 또한 2002년 사스, 2009년 A형 인플루엔자 바이러스, 2015년 메르스 등을 겪었고 최근 코로나-19로 인하여 다시금 경제적, 사회적으로 어려움을 겪고 있다. 메르스는 확진환자 186명, 사망 38명에 그쳤으나 장기간에 걸쳐 경기침체・국민 삶의 질 저하 등 막대한 피해를 주었고, 5-7년 주기로 발생이 계속되고 있다. 신종감염병 등으로 인해 감염병의 발생 총량은 증가하고, 메르스, 에볼라, A형 인플루엔자 바이러스, 인체감염증 등의 인수공통감염병 위험 증가, 해외 감염병의 유입 위험 증가, 항생제 내성 및 의료기관 내 의료 관련 감염 증가, 국제행사 및 다중이용시설 증가와 단체급식 확대로 인한 집단감염 등 감염병 관련 건강 위협요인이 다변화되고 있다[11]. 코로나-19 사태에서도 나타났듯이 신종 감염병과의 싸움에서 가장 중요한 무기는 바로 정보력

으로 질병의 원인이 되는 병원체가 무엇이며, 어떤 식으로 전염되는지, 잠복기는 얼마나 되는지, 이병원체가 가지고 있는 약점은 무엇인지 등에 대해서 얼마나 신속하고 정확하게 파악할 수 있는지가 중요하다. '감염병 예방 및 관리에 관한 법률'이 허용되는 범위 내에서 환자발생 지역에 즉각대응팀 출동 및 현장대응을 실시하고, 국토부, 과기부 합동 '코로나-19 역학조사 지원시스템'을 구축하여 신용카드 사용내역, CCTV, 휴대폰 위치정보 등 IT기술을 적극 활용하여 접촉자들을 빠르게 파악하고 격리하였다. 또한 이를 통해 빅데이터 기반 역학조사 기반을 구축하였다[12]. 해당 시스템은 확진자 면접조사 결과를 보완, 빅데이터의 실시간 분석이 가능해져 확진자 이동 동선과 시간대별 체류지점을 자동으로 파악할 수 있게 되고, 대규모 발병지역을 분석하여 지역 내 감염원 파악 등 다양한 통계분석을 지원하여 기존에는 정보수집, 분석 시 질병관리본부를 지원하는 28개 기관 간 공문 작성 및 유선연락 등의 과정이 대부분 수작업으로 이뤄져 왔으나, 이를 스마트시티 기술 시스템으로 전환함에 따라 정보 취득의 신속성과 정확성을 확보한다는 평가가 존재한다. 그러나 현재 확진자 설문, 위치추적, CCTV 등을 활용한 동선 확인 및 접촉자 관리에 인력이 과다 투입되고 있으며, 제2파에 대응하여 역학조사는 효율적인 인력배치와 빠르고 정확한 정보수집이 이루어지는 방향으로 나아가는 것이 필요하다.

Ⅱ. 감염병 대응 분석 사례

2.1 소셜 미디어 분석 방법론

언론보도를 이용하여 감염병을 분석한 국내 KCI 연구들은 A형 인플루엔자 바이러스, 메르스 등을 대상으로 문맥 분석, 키워드 빈도분석, 그리고 소셜 네트워크 분석 기법인 의미 연결망 분석(Semantic network analysis) 등의 분석 방법론들이 사용되어왔다[13,14].

소셜 미디어 분야는 빅데이터 분야 중의 주요 축으로서 활발하게 연구에 이용되어 왔는데, 최근의 연구들은 게시판, 댓글 내의 감정 표현, 소셜 정보 등과 같은 비정형 데이터들을 이용하여 마케팅에 적용하여 활용하거나 선거 예측을 하는 사례들이 있다[15,16]. 하지만 소셜 데이터 분석에 있어서 난관에 되는 이슈는 비정형 데이터의 전처리 방법이다. 특정한 키워드를 필터링한 후 분석하고 시각화하는 네트워크 통계 분석 및 시각화 프로그램들이 있지만[17,18], 데이터 정제를 위하여 수집된 소셜 미디어 데이터의 거짓 정보를 제거하고 분류 할 수 있는 프레임워크를 개발하기 위한 방법을 제시되어야 한다. 관련 연구에 사용되는 방법은 기계 학습 및 자연어 처리 분야에서 문서 집합의 추상적인 주제를 발견하기 위한 통계적 모델 중 하나로, 토픽 모델링(Topic modeling)이 이용된다. 실제 한 연구에서는 국내 뉴스 빅데이터를 토픽 모델링 분석 방법을 이용한 텍스트마이닝(Text-mining)을 한 사례가 있다[19].

2.2 본 연구의 목적과 방법

정보전염(Infodemic)이 미디어를 통해 확산되며 사회, 정치, 경제 안보에 치명적 위기를 초래하고 있다. 정보전염이 포함된 소셜 데이터를 분석하는 과정에서 루머가 포함된 잘못된 데이터가 거짓 신호를 전달할 위험이 있다. 따라서본 연구에서는 데이터를 수집하는 과정에서 사실 확인이 되는 언론 매체와 다르게 정확한 출처를 알 수 없는 부정확한 정보들이 포함되는 주요 소셜 미디어 플랫폼 중의 하나인 트위터(Twitter) 데이터를 수집하여 사실과 무관한 내용을 사전 차단하고 광고 및 거짓정보에 대한 정제 시간을 단축하고자 한다.

Ⅲ. 연구 방법

3.1 데이터 수집 방법과 키워드 분류

본 논문에서는 실질적인 데이터 분석을 통해 팩트 체크를 통한 의미 유사성을 실시간으로 평가하고자 오픈 API (Open Application Programming Interface)을 통해 수집이 가능한 트위터를 활용했다. 학습 데이터 목적으로 수집하고자 하는 트위터 데이터를 비교하기 위해 검색 키워드 코로나로 지정하여 현재 거짓으로 판별된 글의 문장들이 선별된 구글 팩트 체크(Google ToolBox FactCheck) 35개의 문장들을 크롤링하여 문장 명사 단위 키워드를 유의어 사전<표 1>을 구축한 후 형태소 분석을 통해 빈도 수가 높은 키워드를 선별하여 온톨로지 사전으로 조합했다[20].

〈표 1〉 온톨로지 사전

분류	키워드 종류
약품	이버멕틴, 나트륨, 아스트라제네카, 임상, 클로로퀸, 구충제 etc
증상	코로나, 돌연사, 증후군, 불임, 감기, 무증상, 퇴행성 etc
제2감염병	원숭이, 에이즈, 전파, 약화, 변종, 두창 etc
유아	어린이, 영유아, 모유 etc
국가	일본, 한국, 중국 etc
루머	면봉, 기생충, 혈액 etc
마스크	마스크

수집한 구글 팩트 데이터의 명사 단위 분절하여 중요 키워드를 기준으로 그룹화 하였다. 그룹화 예시는 <표 2> 과 같다. 트위터의 데이터 수집 과정은 2020년 1월부터 2022년 10월 27일까지 코로나와 등의 감염병 관련 트위터 문장 1,387,759 건을 수집했다. 수집된 데이터는 코로나, 증상, 특징, 예방, 변이 키워드 기준으로 전화번호, 주소와 같은 고유 식별데이터를 제외하여 적재하였으며 구글 팩트 데이터에서의 온톨로지 사전을 기준으로 동일하게 그룹화했다. 수집된 트위터 문장들을 정제하기 위해 크롤링 시 중복되는 주소들을 제거했다.

〈표 2〉 키워드 그룹화 방법

텍스트	키워드	분류
미국립보건원이 코로나 19 치료에 이버멕틴 사용을 권고했다	국립보건원, 코로나, 치료, 이버멕틴, 사용, 권고	약품
코로나 19 백신 접종이 돌연사증후군을 일으킨다	코로나, 백신, 접종, 돌연사, 증후군	증상
원숭이두창이 화이자사가 개발한 코로나19백신 접종 국가해서만 발생한다	원숭이, 두창, 화이자, 사가, 개발, 코로나, 백신	제 2 감염병
코로나19 백신 접종이 불임을 일으킨다	코로나, 두창, 아스, 코로나, 백신, 부작용	증상
원숭이두창이 아스트라제네카 코로나19 백신의 부작용이다	원숭이, 두창, 아스, 코로나, 백신, 부작용	제 2 감염병
코로나 19 백신이 면역 체계를 손상시켜 에이즈를 유발한다	코로나, 백신, 면역, 체계, 손상, 에이즈, 유발	제 2 감염병

3.2 학습데이터 구축 방법

일반적으로 수집된 글의 경우 트위터의 특성 구어체의 문장 사용으로 맞춤법의 교정이 필요하여 맞춤법 검사기를 활용한 py-hanspell 라이브러리를 활용하여 맞춤법을 교정하였다[21]. 트위터 문장 내 온톨로지 사전 단어 2개 이상 존재하는 경우 가져오는 것으로 선별했다. 최종적으로 7개의 온톨로지 사전 키워드로 분류된 77,722 건의 트위터 데이터를 모델 학습 목적으로 구축했다. 문장 쌍 구성은 같은 라벨을 공유하는 구글 팩트 데이터, 트위터 데이터를 문장쌍으로 1:N 으로 데이터 형성 했다 <표 3>.

수집, 통합, 분석 영역에서의 다양한 소셜 미디

〈표 3〉 문장 쌍 구성

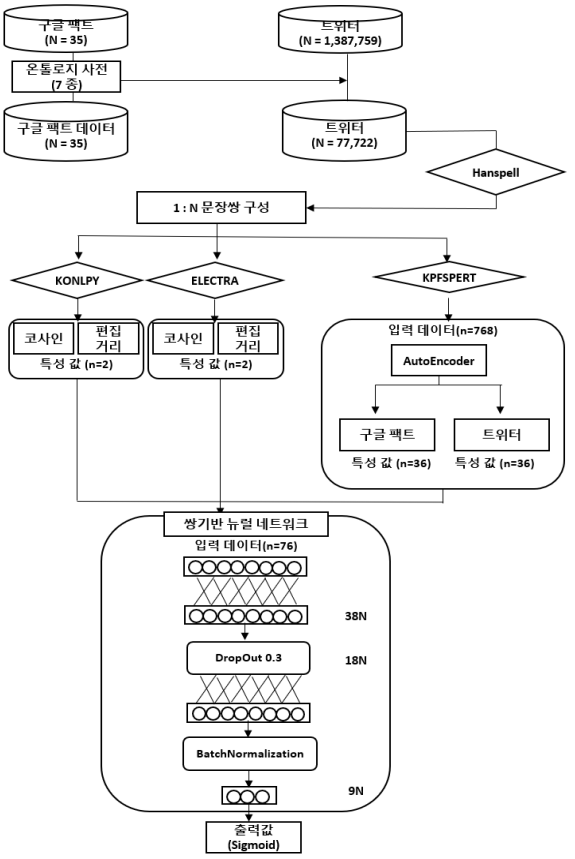
분류	데이터 출처	데이터 내용
증상	구글 팩트	한국 코로나 19 백신 접종이 돌연사 증후군을 일으킨다
	트위터	한국한국에서 코로나 백신 접종 후 돌연사 증후군 일으킨다
		코로나는 위험다하다 어린이에게
		코로나 백신 효과있나 치료
		우한폐렴 충격코로나 검사 백신 면봉에백신 절대로 맞...
		코로나 바이러스 증후군 없습니나 백신 맞아도
		원숭이 두창이 코로나 백신 부작용이다
		코로나 사기진단 백신 인체

어 수집 시 관련 토픽을 수집하는 절차에서의 주요 이슈는 비정형 데이터의 변환이다. 수집 채널의 다양성으로 인한 통합, 추출 대상의 특정 키워드 수집, 지속적 학습을 위한 설계와 같은 기술적 요소를 해결하기 위해 텍스트의 의미적 유사성 기반으로 토픽들을 추출하고 맞춤법 등의 비정형 요소들을 제외 및 분류하는 방법을 이용했다 <그림 1>.

Ⅳ. 모델 개발

4.1 활용 모델

수집된 트위터 소셜 데이터와 같은 비정형 데이터는 마이닝 기법을 사용하여 분석 가능한 형



〈그림 1〉 학습데이터 구축과 모델 개발 방법

태로 형식화하게 된다. 정형 데이터인 35개의 구글 팩트 문장쌍에 대해 7개의 온톨로지 사전 기반으로 분류된 트위터 문장을 코사인 유사도(Cosin Similarity), 편집거리(EditDistance), 자연어 처리 방법을 이용해서 문맥 정확도를 산출했다. BERT(Bidirectional Encoder Representations from Transformers) 모델은 2018년 구글에서 발표하였으며[22], 앞에 나오는 단어로 다음에 올 단어를 예측하는 것이 아니라 문장의 중간 단어를 마스킹한 후 전체 문장에서 해당 단어를 예측하는 MLM(Masked Language Model) 방으로 문장이 이어지는 관계인지 아닌지를 학습하는 NSP (Next Sentence Prediction) 기능을 가진다.

BERT 모델 계열로, 단어 벡터 간의 유사도 비교문장과 문맥을 이해하기 위해 정형 소셜데이터 학습에 사용되는 KPFSPERT(Korean Sentence BERT) 모델과 함께 KonlPy와 ELECTRA (Efficiently Learining an Encoder that Classifies Token Replacements Accurately)를 융합한 쌍 기반뉴럴 네트워크 모델을 인용했다[23-25]. KonlPy 모델은 트위터에서 개발한 Twitter 한국어 처리기에서 파생된 오픈소스 한국어 처리 모델이다. ELECTRA는 MLM 방식 대신 RTD (Replaced Token Detection)을 이용해 문장을 부분적으로 단어가 바뀐 여러 문장으로 생성이 가능하게 하며 '실제'와 '가짜' 입력 데이터를 구별하는 NLP 모델이다.

쌍기반 뉴럴네트워크란 약물-표적 단백질 연관관계 예측을 위한 융합 모델로 알려져 있으며, 쌍 기반 레이어의 설계 방법으로 첫번째 모델로는 단백질의 아미노산 서열에 기반하여 자연어 처리 방법으로 특질을 추출한다. 두 번째 모델로는 화합물 내 원자의 다양한 속성들(원자번호, 연결개수)등의 단일 정수값을 변환한다[26]. 다중 모델을 융합함으로써 특정 특질 값에 편향되지 않도록 조절하는 효과를 가진다. 현재의 소셜데이터 신뢰도 분석 시스템에서는 3개의 모델을 융합하였으며, KPFSBERT는 자연어처리를 통해 문장 속성들을 단일 정수값으로 변환한다. KONLPY, ELECTRA의 코사인 유사도, 편집거리 값을 특질 값으로 변환 및 문장간의 유사도를 평가한 식별자로 할당한다. 편집거리는 두 문자열의 유사도를 판단하는 알고리즘이다. 문자열간의 수정, 삽입, 삭제 기능을 몇 번의 동작이 몇 번 이루어지는가에 따라 문자열의 유사도를 판단하며, 최소 편집거리를 만드는 경로의 값을 제공한다. 코사인 유사도는 두 벡터 간의 코사인 각도를 구하여 벡터 간의 유사도를 판단한다. 문서 단어 행렬화를 하여 코사인 유사도를 구한다<표 4>. 하나의 판별 예시로, 문자 키워드 빈도 수에 따라 문장 1과 문장 2의 유사도의 경우 0.67, 문장 1과 문장 3의 유사도의 경우 0.67, 문장 2와 문장 3의 유사도는 1.00로 산출된다.

〈표 4〉 문장 유사도 판별 방법

	코로나	확진자	구충제	임상
문장1	0	1	1	1
문장2	1	0	1	1
문장3	2	0	2	2

4.2 모델 개발과 성능

제안한 온톨로지 사전 기반 소셜 토픽 기법과 분류 기법을 활용하여 문장 쌍 간 유사도에 대한 정확도 산출을 위해 쌍기반뉴럴 네트워크 모델를 사용했다<그림 1>. KONLPY, ELECTRA를 이용해 기존 문장 쌍을 이용한 코사인, 편집거리 특성값을 추출을 하였다. 코사인 값은 백터화된 행렬로 변환하여 단어 간의 연관관계를 나타내며 편집거리 값은 문장의 단어 개수 및 문자열의 유사도를 판단한다. 자연어 처리를 위해 KPFSBERT에서의 특성값 추출에는 오토인코더(AutoEncoder)를 활용하였다

오토인코더란, 입력값을 압축하여 출력값으로

복사하여 차원을 축소하는 방법으로 압축된 네트워크층 병목(bottleneck)으로 입력데이터의 학습된 압축을 이끌어낸다. KPFSBERT의 임베딩값 768개의 특성값을 구글, 트위터 각각 36개의 특성으로 차원 축소하여 쌍기반 뉴럴네트워크의 특성값으로 활용하였다. 본 연구에서 제시한 쌍 기반 뉴럴 네트워크 모델을 통해 의미 연관성 기반 소셜 문맥에 대한 유사도 평가 성능은 F1 score로 확인하였으며 다음과 같이 산출되었다; 정확도 0.65, 정밀도 1.0, 재현율 0.65, F1 score 0.79.

Ⅴ. 결 론

본 연구에서는 하나의 대표적인 소셜 매체인 트위터 비정형 데이터를 감염병 루머 키워드를 기준으로 수집하여 분석이 가능한 문장쌍 형태로 형식화했다. 이후 KPFSPERT 모델과 함께 자연어 처리 기법에 사용되어지는 KonlPy와 ELECTRA를 이용한 쌍 기반 뉴럴 네트워크를 이용하여 문장쌍 간 유사 정확도를 측정했다. 기존의 KPFSBERT 모델은 언론진흥재단에서 제공하는 정형화된 언론, 뉴스 기사 데이터 혹은 카카오브레인에서 제공하는 문장 쌍 데이터에 특화된 모델이다[23]. 이번 연구로 개발된 딥러닝 모델 결과를 통해 정형데이터가 아닌 비정형데이터 학습을 통한 유사도 분석 정확도가 정형데이터 학습 성능에 대응하는 성능를 확인했으며 소셜 데이터를 이용해 거짓 정보를 차단하고 정보를 수집하는 감염병 예방 및 대응의 가능성을 제시한다.

우리나라 코로나-19 대응은 해외 여러나라와 비교를 하더라도 잘 되어 있다. 그러나 확진자 발생 이후 역학조사를 기반으로 대응을 하다보니 코로나 감염 환자 급증으로 병상이 부족해져 중증 환자가 집에서 입원 대기중에 사망하는 사고가 발생하고 의료 수요 급증 시 의료자원을 효율적으로 관리해야 할 필요성이 대두된다. 또한 코로나-19 같은 호흡기 질환의 경우 날씨, 지역적 특성을 영향을 많이 받는 질환으로 선제적인 대응이 필요하다. 이러한 문제를 해결하기 위해서 전 세계적으로 인터넷이 널리 보급됨에 따라 다양한 정보들이 이전보다 신속하게 퍼지고, SNS와 같은 과거에는 존재하지 않았던 플랫폼을 통해 정보가 공유되기도 하며, 이러한 텍스트 데이터를 분석 하기 위해 자연어 처리기법을 이용하여 실시간으로 발생하는 텍스트 데이터를 분석하고 독감의 발생 예측 등을 시도하는 연구들이 많이 이루어지고 있다.

위기 대응을 위해 '언제 어디로 필요한 의료 자원을 배치해야 하는지'와 같은 질문을 끌어낼 수 있으며 신종 감염병 발병 주기가 짧아지고, 국가간 이동이 많은 현대 사회에서는 발병 후 대응보다 사전발병 추이를 분석하고 선재적 대응하는 것이 중요하다. 의료분야의 경우 고도의전문지식이 필요한 영역으로 역학조사관 인력도 부족한 실정으로 인공지능을 활용하여 선재적으로 신종 감염병 발병 가능성을 예측하여 선제적 대응이 가능하도록 하는 것이 필요하다. 향후연구로, KPFSBERT의 기반인 BERT 모델은 샘플한 문장당 최대 15%의 단어만 마스킹하여 학습하므로 적은 데이터 셋으로는 충분한 학습을 이룰 수 없다. 충분한 데이터 셋을 만들기 위해 GPT(Generative Pre-Trained Transformer)모델을 이용해 유사한 문장을 여러 개 만들어 문장쌍을 구성하는 방법보다 ELECTRA의 전체 단어를 모두 각각 바꾸는 기법을 사용해보고자 한다.

참 고 문 헌

[1] Institute of Medicine. Microbial Threats to Health: Emergence, Detection, and Response. Washington DC: Institute of Medicine, 2003.

[2] Weiss RA, McMichael AJ. Social and environmental risk factors in the emergence of infectious

diseases. Nature medicine. 2004;10(12 Suppl): S70-6

[3] 옥철, “WHO, 코로나19 팬데믹 선언,” 연합뉴스, 2020.3.12.

[4] 천병철. 인수공통감염증의 역학적 특성. 대한의사협회지. 2004;47(11):1019-34.

[5] Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. Nature. 2008;451 (7181):990-3.

[6] Nature Index, By the numbers: counting the costs of infectious illness, 2021.10.27.

[7] World Health Organization. Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003 (Based on data as of the 31 December 2003). Available from: http://www.who.int/csr/sars/country/table 2004_04_21/en/.

[8] World Health Organization. Situation updates - Pandemic (H1N1) 2009. Available from: http://www.who.int/csr/disease/swine-flu/updates/en/.

[9] World Health Organization. Disease outbreak news - Middle East respiratory syndrome coronavirus (MERS-CoV). Available from: http://www.who.int/csr/don/archive/disease/ coronavirus_infections/ en/.

[10] Shin, N.R., Baek, S.J., Yoo, H.S., & Shin, I.S.(2019),Global trends in preparation for future infectiousdiseases. Brief Report. 12(5), 120-126.

[11] 제약바이오협회, 2020년 이후 발생한 신종 감염병 종류 및 특징, http://www.kpbma.or.kr/ attach/KPBMA_Brief_20.pdf

[12] 중앙방역대책본부, COVID-19 대응전략, 200722.

[13] 주영기,유명순, “신문·TV뉴스의 신종 출몰형 질환 및 만성질환 보도 패턴 분석,” 한국언론학보, 제54권, 제 2호, pp.363-381, 2010.

[14] 주영기,유명순, “한국 언론의 신종플루 보도 연구,” 한국언론학보, 제55권, 제5호, pp.30-54, 2011.

[15] Kushin, M.J. and M. Yamamoto, “Did Social Media Really Matter? College Students’ Use of Online Media and Political Decision Making in the 2008 Election”, Mass Communication and Society, Vol.13, No.5, pp.608-630, November, 2010. DOI:10.1080/15205 436.2010.516863

[16] Michaelidou, N., N.T. Siamagka, and G. Christodoulides, “Usage, Barriers and Measurement of Social Media Marketing : An Exploratory Investigation of Small and Medium B2b Brands”, Industrial Marketing Management, Vol.40, No.7, pp.1153-1159. October 2011, DOI: 10.1016/ j.indmarman.2011.09.009

[17] Man-Mo Kang, Sang-Rak Kim, Sang-Moo Park, “Analysis and Utilization of Big Data”, Communications of the Korean Institute of Information Scientists and Engineers, Vol. 30, No. 6, 2012.6, pp. 25-32, June, 2012.

[18] Keun-Tae Kim, “Environment Challenge in Company for Big Data Analysis”, Korea Information Processing Society Review, Vol.19, No.2, March, 2012.

[19] 김태종, 뉴스 빅데이터를 활용한 코로나19 언론보도 분석: 토픽모델링 분석을 중심으로, 한국청소년정책연구원 청소년정책분석평가센터, 2020.

[20] Google ToolBox FactCheck, https://toolbox.google.com/factcheck/explorer/search/%EC%BD%94%EB%A1%9C%EB%82%98;hl=ko

[21] py-hanspell, https://github.com/ssut/py-hanspell

[22] BERT, https://github.com/KPFBERT/kpfbert)

[23] KPFSBERT, https://github.com/KPFBERT/kpfSBERT

[24] KONLPY, https://github.com/open-korean-text/open-korean-text
[25] ELECTRA, https://github.com/google-research/electra
[26] 이문환, 김응희, 김홍기. (2017). 약물-표적 단백질 연관관계 예측모델을 위한 쌍 기반 뉴럴네트워크. 인지과학, 28(4), 299-314.

저 자 소 개

정 두 영(Doo Young Jung)
·2020년 2월 : 순천향대학교 의료 IT 공학과 (학사)
·2021년 5월~현재 : 미소정보기술 헬스케어사업본부
<관심분야>: Digital Healthcare, Artificial Intelligence, Deep learning

이 상 준(Sang-Jun Lee)
·2020년 2월 : 한국외국어대학교 생명공학과 (학사)
·2022년 8월 : 차의과학대학교 정보의학 (석사)
·2022년 9월~현재 : 미소정보기 술 헬스케어사업본부
<관심분야> : Digital Healthcare, Artificial Intelligence, Digital Therapeutics, Metaverse.

민 경 일(MIN KYUNG IL)
·2011년 2월 : 한국외국어대학교 경영대학원 (석사)
·2022년 3월 : 차의과학대학교 의학전문대학원 정보의학 연구교수
·2022년 7월~현재 : 미소정보기술 헬스케어사업본부
<관심분야> : Digital Healthcare Application, Digital Therapeutics, Metaverse.

정 석 송(Seogsong Jeong)
·2022년 2월 : 서울대학교 의과대학 의과학 (박사)
·2022년 3월~현재 : 미소정보기술 연구소
<관심분야> : Data Science, Statistics, Aritificial Intelligence

한 현 욱(HyunWook Han)
·2015년 : 차의과학대학교 의학전문대학원 의료정보학(박사, 의사)
·2018년~현재 : 차의과학대학교 의학전문대학원 정보의학교실(주임교수)
<관심분야>: Healthcare Bigdata, Digital Healthcare, Medical Infomatics, Data Science