KIPS Trans. Softw. and Data Eng.
Vol.12, No.9 pp.387~398
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2023.12.9.387

Predicting the Number of Confirmed COVID-19 Cases Using Deep Learning Models with Search Term Frequency Data

Sungwook Jung†

ABSTRACT

The COVID-19 outbreak has significantly impacted human lifestyles and patterns. It was recommended to avoid face-to-face contact and over-crowded indoor places as much as possible as COVID-19 spreads through air, as well as through droplets or aerosols. Therefore, if a person who has contacted a COVID-19 patient or was at the place where the COVID-19 patient occurred is concerned that he/she may have been infected with COVID-19, it can be fully expected that he/she will search for COVID-19 symptoms on Google. In this study, an exploratory data analysis using deep learning models(DNN & LSTM) was conducted to see if we could predict the number of confirmed COVID-19 cases by summoning Google Trends, which played a major role in surveillance and management of influenza, again and combining it with data on the number of confirmed COVID-19 cases. In particular, search term frequency data used in this study are available publicly and do not invade privacy. When the deep neural network model was applied, Seoul (9.6 million) with the largest population in South Korea and Busan (3.4 million) with the second largest population recorded lower error rates when forecasting including search term frequency data. These analysis results demonstrate that search term frequency data plays an important role in cities with a population above a certain size. We also hope that these predictions can be used as evidentiary materials to decide policies, such as the deregulation or implementation of stronger preventive measures.

Keywords : Deep Learning Algorithms, Number of Confirmed COVID-19 Cases, Search Term Frequency Data, Time-Series Data

검색어 빈도 데이터를 반영한 코로나 19 확진자수 예측 딥러닝 모델

정 성 욱†

요 약

코로나 19 유행은 인류 생활 방식과 패턴에 큰 영향을 주었다. 코로나 19는 침 방울(비말)은 물론 공기를 통해서도 감염되기 때문에 가능한 대면 접촉을 피하고 많은 사람이 가까이 모이는 장소는 피할 것을 권고하고 있다. 코로나 19 환자와 접촉했거나 코로나 19 환자가 발생한 장소에 있었던 사람이 코로나 19에 감염되었을 것을 염려한다면 구글에서 코로나 19 증상을 찾아볼 것이라고 충분히 예상해 볼 수 있다. 본 연구에서는 과거 독감 감시와 관리에 중요 역할을 했었던 구글 트렌드(Google Trends)를 다시 소환하고 코로나 19 확진자수 데이터와 결합하여 미래의 코로나 19 확진자수를 예측할 수 있을지 딥러닝 모델(DNN & LSTM)을 사용한 탐색적 데이터 분석을 실시하였다. 특히 이 연구에 사용된 검색어 빈도 데이터는 공개적으로 사용할 수 있으며 사생활 침해의 우려도 없다. 심층 신경망 모델(DNN model)이 적용되었을 때 한국에서 가장 많은 인구가 사는 서울(960만 명)과 두 번째로 인구가 많은 부산(340만 명)에서는 검색어 빈도 데이터를 포함하여 예측했을 때 더 낮은 오류율을 기록했다. 이와 같은 분석 결과는 검색어 빈도 데이터가 일정 규모 이상의 인구수를 가진 도시에서 중요한 역할을 할 수 있다는 것을 보여주는 것이다. 우리는 이와 같은 예측이 더 강력한 예방 조치의 실행이나 해제 같은 정책을 결정하는데 근거 자료로 충분히 사용될 수 있을 것으로 믿는다.

키워드 : 딥러닝 알고리즘, 코로나 19 확진자수, 검색어 빈도 데이터, 시계열 데이터

1. 서 론

2020년 초 시작된 코로나 19 팬데믹으로 사람들의 일상 생활이 크게 변하면서 코로나 이전의 세상과 코로나 이후의 세상으로 나누어야 한다는 말이 회자 됐었다[1]. 이후 3년 동안은 대면 접촉을 줄이는 대신 온라인 소통을 늘리는 방식으로 인간 관계가 이루어졌다. 이처럼 대면 접촉을 줄이는 이유는 코로나 바이러스가 접촉을 통해 호흡기로 들어가 감염되기 때문이다[2]. 따라서 한국 정부는 5인 이상 모임을 금지하고 음식점에서의 식사 시간을 저녁 9시로 제한했으며, 시민들도 대중 교통 이용을 자제했다. 이처럼 시민들의 협조로 코로나 19

※ 이 논문은 2018년 대한민국 교육부와 한국연구재단 그리고 서울대학교 언론정보연구소의 지원을 받아 수행된 연구임(NRF-2018S1A5B8070398).
† 정 회 원 : 서울대학교 언론정보연구소 선임연구원, 성균관대학교 강사
Manuscript Received : May 9, 2023
First Revision : July 24, 2023
Accepted : August 4, 2023
* Corresponding Author : Sungwook Jung(jj4863@naver.com)

확진자가 감소하자 정부는 2021년 11월 사회적 거리두기를 완화했다. 하지만 코로나 확진자수가 한 달여 만에 다시 크게 증가하자 이전의 강력한 사회적 거리 두기 정책으로 회귀했고, 그제야 의료 체계의 범위 안에서 다시 안정되게 관리되었다. 대면 접촉이 많으냐 적으냐가 확진자수 증감에 영향을 미친다는 확실한 실제 사례였다. 인공 지능을 사용하여 코로나 19 확진자수를 예측하려는 연구는 팬데믹으로 선포된 이후 지금까지 지속되고 있다. 전통적 시계열 분석과 마찬가지로 인공 지능을 사용한 대부분의 연구도 한 가지 변수(feature), 즉 과거의 확진자수를 바탕으로 미래의 확진자수를 예측했다[3-9]. 일부 논문만이 대면 접촉의 이동 방향을 알려주는 열차나 비행기에 의한 인구 이동 데이터를 사용하거나[10], 온도나 습도 데이터[11], 또는 소셜 미디어에서의 검색 지수[12]를 사용하여 확진자수 예측을 시도했다.

소셜 미디어에서의 검색 지수처럼 구글에서의 키워드 검색 활동이 전염병 발병을 관리하고 예측하는데 사용될 수 있음을 보여주는 연구는 쉽게 찾아볼 수 있다[13,14]. 특히 구글 플루 트렌드(GFT)는 한때 독감과 같은 병의 활동을 감지해 낼 수 있는 시스템으로 많은 관심을 받았었다. GFT에서 생성되는 데이터는 인간 행동의 측면을 반영하기 때문이다. 예를 들어 독감에 걸렸을 것으로 예상되거나 독감을 염려하는 사람들이 독감의 증상, 독감 예방 방법, 독감의 위험성 등을 알아보고자 한다면 구글 검색을 통해 정보를 얻을 것이라고 예상해 볼 수 있다. 따라서, 이와 같은 검색어가 급증하는 지역에서는 독감이 시작된 것으로 판단할 수 있고, 이를 보여주는 실제 사례도 있었다[15]. 하지만 GFT에 의한 독감 예측과 CDC에 의해 측정된 실제 환자 수 사이에는 커다란 차이가 있는 것으로 나타나면서 언론에서 처음으로 GFT 시스템에 문제점을 지적하기 시작했다[16,17]. 초기 GFT는 크게 몇 가지 방법론적인 문제를 드러냈는데, 특히 검색어 선택과 제외 같은 가장 중요한 측면에 있어서도 경험적 증거를 바탕으로 하지 않았다[18]. 즉, 독감과 가장 관련이 있는 45개 검색어의 평균적인 추세를 하나의 변수로 사용해 예측했지만 45개의 검색어에는 단순히 전염병에 대한 일반적인 관심을 나타내는 단어(어떤 단어인지는 알려지지 않았다)까지 포함되면서 예측 정확도를 크게 떨어뜨린 원인으로 지목되었고[18], 이후 GFT 사이트는 폐쇄되었다.

그렇다면 검색어를 통한 예측은 전혀 불필요한 것인가? 그렇지 않다. 검색어를 통한 예측은 초기 정보로서 여전히 필요한 것으로 여겨지고 있다. CDC 인플루엔자 감시 및 발병 대응 팀장은 어떤 데이터도 데이터가 전혀 없는 것보다는 도움이 된다는 생각에 “all the time” 구글 플루 트렌드를 모니터하고 있다고 말한다[17].

따라서 본 연구는 과거 인플루엔자 감시 및 관리에 큰 역할을 했던 구글 검색어 빈도 데이터를 다시 소환하고 과거 코로나 19 확진자수 데이터와 결합하여 미래의 코로나 19 확진자수를 예측할 수 있을지 분석하는 것이다. 특히 구글 트렌드를 이용한 검색어 빈도 데이터는 인구 이동 데이터나 임상 데이터와는 달리, 사생활 침해의 우려가 전혀 없으며 공개된 사이트에서 누구나 이용할 수 있다. 코로나 환자와 접촉했거나 코로나 환자가 발생한 장소에 있었던 사람이 코로나에 전염됐을 것을 걱정한다면 반드시 검색해야 할 단어인 ‘코로나 증상’, 이 하나의 검색어만을 사용하여 매일매일의 검색 빈도수를 구글 트렌드를 통해 확보했다.

코로나 19 잠복기는 5일에서 6일 정도로 알려져 있다[19]. 만약 오늘 코로나 바이러스에 감염된다면 평균적으로 5일에서 6일 이후에는 코로나 19 증세가 발현하여 코로나 19 환자가 되는 것이다. 이는 바꿔 말하면 코로나 바이러스에 감염돼도 최소 5일 동안은 증상이 없기 때문에 감염된 사실을 모른 채 직장 출근 등 사회 생활이 가능하다는 것을 의미한다. 하지만 코로나 바이러스는 잠복기에도 전염이 일어난다[20]. 만약 예측 모델이 개발돼 5일 후 코로나 19 확진자 수가 급격하게 증가할 것으로 예상된다면, 이는 앞으로 5일 동안은 코로나 바이러스에 감염돼 잠복기에 해당하는 환자들이 많다는 것을 의미한다. 따라서 국민들에게 주의를 당부하거나 좀 더 강한 접촉 금지 조치를 선제적으로 시행할 수 있는 근거가 될 수 있다.

이 연구의 목적은 과거 확진자수와 검색어 빈도 데이터를 결합하여 5일 후의 코로나 19 확진자 수를 선제적으로 예측하는 모델을 개발하는 데 있다. 이를 통해 향후 또 다른 팬데믹이 발생할 경우 해당 팬데믹의 과거 확진자수와 검색어 빈도 데이터를 본 연구에서 개발한 예측 모델에 적용하여 팬데믹 초기 확진자수 예측 정보를 정책 담당자에게 제공하기 위함이다.

이 모델을 통해 팬데믹 초기 확진자 수의 증가 폭과 기간 등을 현저하게 경감시킬 수 있을 것이다[21].

코로나 바이러스는 앞에서도 언급했지만 인간 생활 모든 방식과 패턴에 큰 영향을 주었기 때문에 의학 관련 데이터는 물론 구글에서의 키워드 검색 활동에도 영향을 주었을 것으로 충분히 예상 가능하다. 그리고 이 같은 데이터와 최근 급속도로 발전하고 있는 AI가 코로나 19 확산은 물론 향후 또 다른 전염병을 관리하고 대처하는데 잠재력이 있을 것으로 예상되고 있다[22-23]. 따라서 본 연구는 의학적 방법이나 임상적 데이터가 아닌 구글 트렌드를 바탕으로 생성된 검색어 빈도 데이터를 가지고 딥 러닝(DNN & LSTM)을 이용해 코로나 19 확진자수를 예측할 수 있는지 탐색적 데이터 분석을 시도해 보았다.

과거 메르스[24], 독감[25-26] 그리고 뎅기열[27] 등의 경우에도 발병이나 확산되는 상황을 감시하기 위해 검색어 빈도 데이터를 사용한 연구는 많았지만 본 연구처럼 딥러닝을 가지고 분석하지는 못했다. 딥러닝이 다시 주목받게 된 계기는 컴퓨팅 자원이 발전하고 학습을 위한 데이터가 폭발적으로 증가

하는 2010년대 중반부터였으며, 그중에서도 본 연구처럼 시계열 분석에 사용될 수 있는 LSTM 모델이 미래 예측에 본격 사용되기 시작한 것은 2018년 이후이기 때문이다. 또한, 코로나 19와 관련해서도 앞서 언급한 [12]와 [28] 그리고 [29] 등의 연구 모두 미래 확진자수를 예측하기 위해 검색어 빈도 데이터를 사용했지만 분석을 위해서는 회귀 분석 또는 상관 관계 분석 같은 전통적 통계 기법을 사용했다.

Prasanth et al.[30]의 구글 트렌드를 사용한 코로나 19 확산 예측 연구와 Pan et al.[31]의 코로나 19 팬데믹에 대한 구글 트렌드 분석 연구는 모두 본 논문처럼 검색어 빈도 데이터를 사용하고 딥러닝을 적용해서 미래의 확진자수를 예측했다. 하지만 Prasanth et al.[30]의 논문은 LSTM 모델을 통해 분석했으며, Pan et al.[31]의 논문은 DNN 모델을 통해 분석한 것으로 LSTM 모델과 DNN 모델을 모두 적용해 예측 능력을 비교하고 두 모델의 장단점을 살펴본 본 논문과는 역시 차이점이 있었다.

2. 확진자수 예측을 위한 입력 데이터

좀 더 긴 기간의 시계열 데이터를 사용할수록 어떤 트렌드나 계절에 따른 변동성 등을 파악할 수 있기 때문에 좀 더 정확한 예측이 가능하다[3, 9]. 하지만 우리는 전염병이 발생하고 1년 정도 후부터는 정책 결정권자에게 전염병에 대처하고 관리할 수 있도록 확진자수 예측 정보를 제공하는 것을 목표로 했다. 따라서 약 1년 정도의 시계열 데이터만을 사용하고도 정확하게 예측할 수 있는 예측 모델을 만들고 싶었다. 서울을 포함한 6개 도시에서 2020년 1월 30일부터 2021년 2월 28일까지의 데이터를 입력 데이터로 사용했으며, 2021년 2월 한 달간 확진자수 예측을 목표로 했다. 입력 데이터는 확진자수와 검색어 빈도수 등 두 개의 변수(feature)로 구성됐다. 이 기간 동안의 확진자수는 중앙방역대책본부가 발표하는 일일 확진자수를 사용했으며, 검색어 빈도수는 구글 트렌드(Google trends)를 사용했다. 구글 트렌드는 사용자가 지정된 기간(2020년 1월 30일부터 2021년 2월 28일까지)과 검색어('코로나 증상')를 입력하면 지역별로 그 기간 동안 해당 검색어가 구글 플랫폼에 입력된 빈도수(해당 용어의 절대 검색량)를 기반으로 점수(scores)를 산정한다. 점수는 지수(indexes)로 정량화되며 '100'은 해당 기간 동안 '코로나 증상'이란 검색어가 최다 빈도로 사용됐음을 의미한다[28]. 반대로 '0'은 '코로나 증상'이 검색어로 한 번도 사용되지 않았음을 보여준다. Table 1에 해당 기간 동안 입력 데이터의 표본을 제시했다.

이와 같은 데이터를 확보한 후 특정일을 기준으로 특정일 포함 과거 7일(time step)간의 데이터를 사용해 5일 후(future)의 확진자 수를 예측하였다. 코로나 19는 비말에 의한 전염병이고 따라서 최근의 감염 추세가 앞으로의 확진자수 증감에 영향을 줄 가능성이 크기 때문에 과거 7일간의 데이터를 사용한 것이다. 또한, 이론에서 논의한 것처럼 잠복기를 고려해 5일 후의 확진자 수를 예측하고자 한다. 예를 들어 2월 1일 확진자 수를 예측하기 위해서는 1월 27일을 기준으로 과거 7일간의 데이터 즉, 1월 21일부터 27일까지의 데이터를 사용해 5일 후인 2월 1일 확진자 수를 예측해야 된다(Fig. 1 참조). 따라서 본 연구에서는 2020년 1월 30일부터 2021년 1월 20일까지의 데이터를 훈련 데이터(training data)로 사용하고, 2021년 1월 21일부터 2월 28일까지 39일간의 데이터를 테스트 데이터(test data)로 사용했다.

훈련데이터 = 2020.5.1 ~ 2020.10.31
테스트 데이터 = 2021.1.21~2021.2.28
예측 목표 = 2021.2.1 ~ 2021.2.28

배치 사이즈(batch size)는 데이터를 몇 개씩 끊어 학습할 것인지를 결정해주는 하이퍼파라미터다[32]. 본 연구에서 독립 변수(feature)는 8개(더미 변수[1] 포함)이고 타임스텝(time step)은 7일 이기 때문에 배치 사이즈를 10으로 주었다면, 8X7=56개의 데이터를 하나의 묶음으로 10개씩 묶은 데이터가 하나의 미니 배치(mini-batch)가 되고 이렇게 묶은 데이터

Table 1. Input Data Sample

Date	Number of confirmed cases	search term frequency
20200130	1	42.57
20200131	3	40.71
20200201	0	38.86
20200202	0	37.00
20200203	0	34.43
…	…	…
20210224	138	10.86
20210225	114	10.14
20210226	129	9.43
20210227	130	8.71
20210228	117	8.00

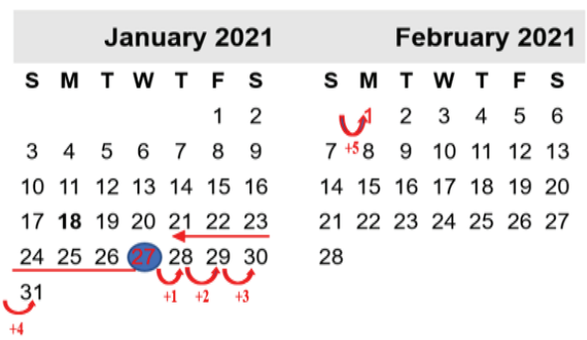

Fig. 1. Prediction of the Number of Confirmed Cases 5 Days After January 27 and Time Step

1) 요일 변수를 더미 변수(6개)로 변환하여 입력 데이터에 포함시켰지만 이후 논문에서 변수 개수에는 포함시키지 않았다.

만큼 끊어서 학습을 하게 된다[33]. 배치 사이즈가 클 경우 학습 속도는 빨라지지만 정확도가 떨어질 수 있고, 배치 사이즈가 작을 경우 속도는 떨어지지만 정확도는 향상될 수 있다. 하지만 너무 많은 데이터에 배치 사이즈를 작게 줄 경우 과적합(overfitting)이 일어나 오히려 정확도가 떨어질 수도 있다. 따라서 본 연구에서는 정확한 예측값을 구하기 위하여 배치 사이즈를 2에서부터 2씩 증가해 18까지 튜닝해 가며 학습시켜 보았다.

3. 확진자수 예측 모델 구성

시계열 데이터를 분석해 미래 예측을 하는데도 딥러닝 기법을 사용할 수 있다. 입력층(input layer)과 출력층 사이에 많은 은닉층을 포함하여 다양한 비선형적 관계를 학습할 수 있는 심층신경망(Deep Neural Network: DNN)과 심층신경망의 한 종류로서 시계열 데이터 처리에 적합한 순환 신경망(Recurrent Neural Network: RNN) 등 두 가지 알고리즘을 사용해 2021년 2월 한 달간 확진자 수를 예측해 보았다.

3.1 심층신경망(Deep Neural Network) 모델

딥러닝 프레임워크인 케라스(keras)를 기반으로 심층신경망 모델을 구성했다. 본 연구에서는 층(layer)을 차례대로 쌓는 순차적 방법(sequential method)을 사용해 1개의 예측값을 발생시키는 출력층과 5개의 은닉층으로 이루어져 총 6개의 Dense 층을 가지는 심층신경망 모델을 구성했다(Fig. 2 참조).

은닉층의 활성화 함수로는 딥러닝에서 가장 인기 있는 활성화 함수이며 은닉층을 깊게 만드는 장점을 살릴 수 있는 ReLU 함수를 사용했다[34]. 활성화 함수는 딥러닝의 예측력에 매우 큰 영향을 미칠 수 있다[35]. 최종 출력값(확진자수 예측)은 선형 회귀 문제로도 볼 수 있기 때문에, 출력층에는 선형 함수(Linear)를 사용했다.

Fig. 2에서와 같이 5개의 은닉층과 각각의 은닉층에 30개가 넘는 유닛이 있기 때문에 각 입력값에 가중치를 곱하고 바이어스를 더해 다음 은닉층으로 보내기 위해서는 상당한 연산 비용이 발생한다. 따라서 좀 더 단순하고 효율적으로 계산하기 위해 Equation (1)과 (2)처럼 벡터화를 이용, 내적을 통해 연산을 수행하도록 했다.

본 연구에서 훈련 데이터의 개수는 346개이고 데이터 하나당 첫 번째 은닉층에 들어가는 입력값은 모두 14(feature=2 x time step=7)개 이다. 따라서 Equation (1)처럼 14개의 입력값(x)에 각각의 가중치(w)를 곱하고 바이어스를 더한 가중합($Z^{[1]}$)을 생성한 후 Equation (2)처럼 활성화 함수(Linear)를 거치면 총 64개(unit)의 결과값($A^{[1]}$)이 산출된다. 이 결과값이 다음 은닉층의 입력값으로 사용된다.

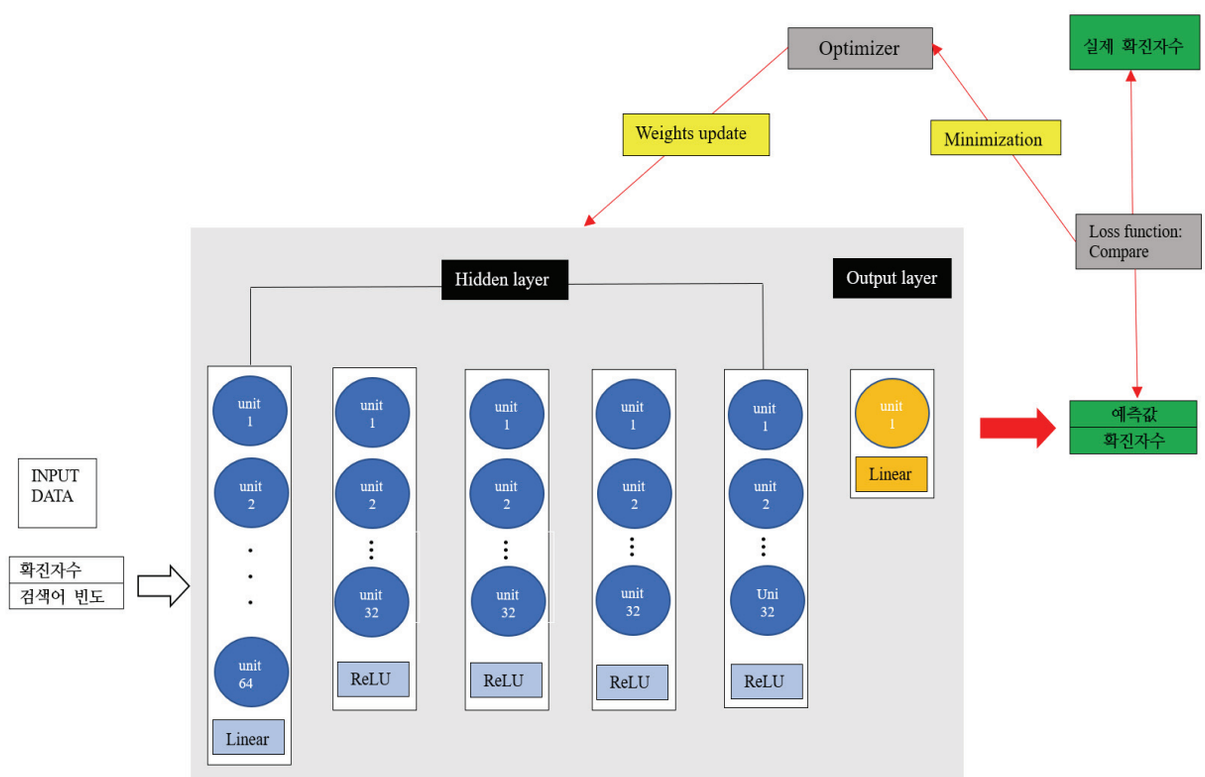

Fig. 2. The Learning Process of the DNN Model Used in This Study

$$Z^{[1]} = X^{T}W + b \quad (1)$$

$$A^{[1]} = \text{Linear}(Z^{[1]}) \quad (2)$$

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}, \quad W=\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1k} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ w_{n1} & w_{n2} & \cdots & w_{nk} \end{pmatrix}$$

· X= 입력값 행렬
· W= 가중치 행렬
· b= 바이어스 벡터
· $Z^{[1]}$=첫번째 은닉층의 가중합
· $A^{[1]}$=첫번째 은닉층의 결과값
· m: 데이터의 개수(346)
· n: 입력값의 개수(14)
· k: 첫번째 은닉층의 유닛 수(64)

이처럼 모델을 구성했으면 다음으로 컴파일(compile) 함수를 호출하여 어떻게 학습시킬지 구체적 과정을 설정해 주어야 한다. 컴파일 함수에서는 크게 3가지, 즉 손실 함수(loss function), 최적화 알고리즘(optimizer), 그리고 평가 지표(metrics)를 지정해주어야 한다[34]. 본 연구에서는 평균제곱오차(MSE)를 적용해 손실 함수를 결정했고, 최적화 알고리즘으로는 아담(Adam)이 사용됐다. 평가 지표(metrics)란 학습 과정이 잘 진행되고 있는지 모니터링 하기 위해 설정하는 지표이다. 회귀 문제에는 평균제곱오차(MSE)와 평균절대오차(MAE)가 사용되는데 본 연구에서는 MAE를 사용하여 손실함수를 모니터링 하였다[34]. Fig. 2에서와 같이 학습 과정을 통해 딥러닝 모델은 예측값과 실제값의 차이를 최소화하는 가중치와 편향의 조합을 찾기 위하여 끊임없이 두 값의 차이를 비교한다[36]. 따라서 손실 함수가 작으면 작을수록 그 모델의 수행 능력은 좋아지게 된다. 또한, 학습 과정에서 일어날 수 있는 과적합을 막기 위해 본 연구에서는 Early Stopping기법을 도입했다. Early Stopping은 과적합이 발생하기 직전에 학습을 중단시키는 기법이다. 본 연구에서는 학습을 총 400번(epochs) 반복하도록 설정했다. 하지만, 모델 성능이 추가 40번 동안 반복하여 학습해도 개선되지 않으면, 400번 이전에도 학습을 중단하도록 설정해 놓았다.

마지막으로 이와 같은 예측의 정확성을 평가하기 위해 본 연구에서는 MAPE(Mean Absolute Percent Error)를 사용했다. MAPE는 예측 오차(이하 오류율)를 측정할 때 쓰이는 가장 일반적인 방법으로 오차의 정도를 백분율 값으로 나타내기 때문에 모델의 성능을 직관적으로 이해하기 쉬운 장점이 있다[37]. Equation (3)에 정의된 것과 같이, 계산 방식은 실제값에서 예측값을 빼준 후, 이를 실제값으로 나눠 준 값들의 평균을 구한다. 그리고 백분율로 변환하기 위해서 100을 곱해준다. 이처럼 실제값으로 나눠줘야 하기 때문에 실제값에 0이 존재한다면 MAPE는 정의되지 않는 단점 또한 가지고 있다[38].

$$MAPE = \frac{\sum \left| \frac{y-\hat{y}}{y} \right|}{n} \times 100\% \quad (3)$$

3.2 순환신경망(Recurrent Neural Network) 모델

순서가 있는 데이터(Sequential data) 학습에 사용하면 좋은 결과를 만들어 낼 수 있는 순환신경망은 심층신경망의 한 종류로서[39], 내부에 순환 구조가 들어 있는 특징이 있다[34]. 심층신경망에서는 각 층의 결과값이 출력층으로 전달(feed-forward neural networks) 되지만, 순환신경망에서는 결과값이 출력층으로 전달되는 동시에 순환 구조를 통해 현재 층의 다음 계산을 위해 사용된다. 이것은 순환신경망이 과거의 모든 정보를 기억할 수 있는 은닉 상태, hs(실수로 이루어진 벡터)를 가지고 있기 때문이다[40]. 순환신경망 모델은 직전의 은닉 상태와 현재의 입력 데이터를 고려하여 출력값을 발생시키며, 은닉 상태는 새로운 입력 데이터가 이와 같은 방식으로 처리될 때마다 업데이트 된다.

본 연구에 사용된 RNN 모델에서는 입력 데이터를 지정해주는 입력층을 따로 만들어 입력층과 출력층 사이에 5개의 은닉층을 갖도록 구성했다. 첫 번째 은닉층에서는 순환 구조를 가지고 있는 신경망을 내부에 배치했고, 나머지 은닉층과 출력층은 DNN과 동일하게 Dense 층으로 구성했다. 따라서, RNN모델은 한 개의 순환층과 5개의 피드-포워드(feed-forward) 층으로 형성됐다. Fig. 3에서와 같이, 입력 순서상의 각 타임 스텝(time step)에서 순환층은 해당 층의 은닉 상태를 직전의 은닉 상태 hs0와 현재의 입력 데이터[2] $[x_a, x_b]_1$를 고려하여 hs1으로 업데이트 시킨다. x_a는 확진자수, x_b는 검색어 빈도이다. 훈련 데이터는 모두 346개 이므로 n은 346이다. 그런 다음 hs1은 해당 타임 스텝의 확진자수를 예측하기 위해 다음층으로 보내진다. 다음 순서의 입력 데이터, $[x_a, x_b]_2$가 순환층으로 들어오면, 그 업데이트된 hs1이 $[x_a, x_b]_2$와 결합하여 다음 은닉 상태 hs2를 발생시킨다.

RNN은 이론적으로 순서가 있는 데이터를 처리하기에 강력한 모델이지만 기간이 긴 시계열 데이터를 학습할 때는 기울기 소실 문제(vanishing gradient problem)가 종종 발생한다

2) 입력 데이터는 모두 2개의 변수(feature:확진자수 + 검색어 빈도) x 7(time step)로 14개이지만, 그림의 단순화를 위해 변수만을 표시했다.

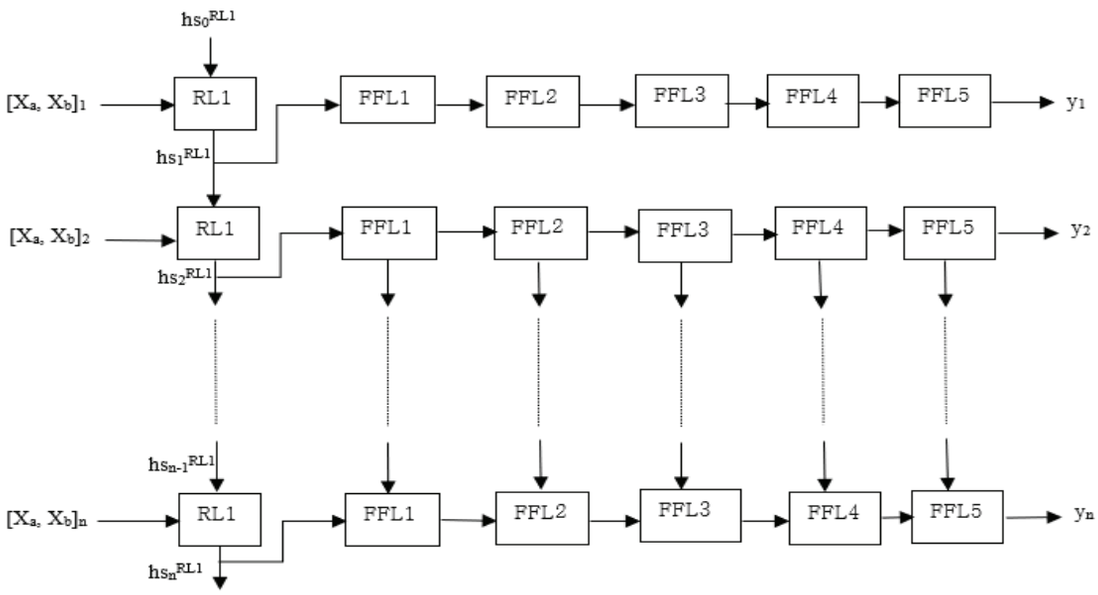

Fig. 3. RNN Model Architecture Used in This Study. Two Features ($[x_a, x_b]_n$) Come in. Then, They Pass Through One Recurrent Layer and Five Feed Forward Layers. Subsequently, the Number of Confirmed Cases (y_n) is Generated as an Output Value

[34, 41]. 따라서 이 문제를 해결하기 위해, RNN에 약간의 변형이 더해진 LSTM 또는 GRU 네트워크가 사용된다[42]. 기존 연구 결과를 살펴보면 LSTM과 GRU의 성능에는 큰 차이가 없기 때문에[42, 43], 본 연구에서는 LSTM을 순환층에 사용하였다. 다음으로 컴파일 함수를 호출하기 위하여 손실함수는 평균제곱오차(MSE), 최적화 알고리즘은 Adam, 평가 지표(metrics)는 평균절대오차(MAE)를 지정했다. 에폭(epochs)과 배치 사이즈, early stopping 같은 하이퍼 파라미터는 DNN 모델과 동일하게 설정했다. 확진자수 예측 성능을 평가하기 위한 오류율 역시 DNN 모델과 동일하게 MAPE를 통해 계산했다.

4. 딥러닝을 통한 분석 결과

4.1 DNN 모델 적용

대부분의 기존 연구처럼 과거 확진자수 만을 가지고 확진자수를 예측하는 것과 과거 확진자수와 검색어 빈도 데이터를 함께 고려하여 확진자수를 예측하는 것 중에 어떤 방식이 오류율을 낮추는지 또한 이 연구의 중요 관심사이다. 따라서 입력 데이터 세트를 2가지 형태(확진자수 vs 확진자수 + 검색어 빈도)로 구분했고, 또한 모든 데이터가 시계열 데이터이기 때문에 정확성을 높이기 위해 요일 변수를 더미 변수로 전환하여 2가지 데이터 세트에 각각 포함시켰다.

대한민국 수도 및 광역시 가운데 가장 많은 인구수를 가지고 있는 서울과 두 번째로 인구수가 많은 부산은 검색어 빈도를 포함하는 데이터 세트(확진자수 + 검색어 빈도)에서 오류율이 더욱 낮아졌다(Seoul: 21.1 → 20.7; Busan: 37.9 → 33.9). 반면, 인구수가 250만 명 이하인 도시(대전, 대구, 광주) 중에서는 대구에서만 검색어 빈도를 포함하는 데이터 세트에서 오류율이 감소했다(Table 2 참조). 또한, 인구수가 250만 명이 넘는 도시에서는(Table 3 참조) 오류율이 35% 미만이었지만, 인구수가 250만명 이하인 도시에서는 오류율이 35%를 넘어섰다.

Table 2. Error Rate When Predicting the Number of Confirmed Cases by Applying the DNN Model

Names of input data set	Combinations	SEL[3)]	BUS	INC	DAE	DAJ	GWJ
Coronics alone	coronics	21.1	37.9	27.2	37.8	120.6	51.3
Searches addition	coronics+searches	20.7	33.9	28.0	36.0	129.0	56.3

Note. Coronics is a neologism that refers to a person, who has contracted the coronavirus (https:// www. urbandictionary.com/define.php?term=Coronic). Searches mean search term frequency data

3) 6개 도시 영어 약어:서울(SEL), 부산(BUS), 인천(INC), 대구(DAE), 대전(DAJ), 광주(GWJ)

Table 3. Population by Metropolitan City in S. Korea (Unit: ten thousand)

CITY	Population
Seoul	960
Busan	340
Incheon	290
Daegu	240
Daejeon	150
Gwangju	140

이처럼 확진자수를 100% 정확하게 예측할 수는 없었지만, 인구수 250만 명 이상의 도시에서는 확진자수 예측 그래프(이하 예측 그래프: Fig. 4의 주황색과 회색 실선)와 실제 확진자수 그래프(이하 실제 그래프: Fig. 4의 파란색 실선) 사이에 변화 추세가 유사함을 확인할 수 있다. Fig. 4는 6개 도시에서 각각의 데이터 세트를 사용해 예측된 확진자수를 보여주는 그래프인데, 서울에서 주황색 실선을 보면, 예측 그래프의 변화 추세와 실제 그래프의 변화 추세가 거의 일치했다. 여기서 변화 추세가 유사하다는 말은 실제 확진자수가 감소(증가) 추세에서 증가(감소) 추세로 변화할 때, 예측된 확진자수도 감소(증가) 추세에서 증가(감소) 추세로 변화하는 경향을 의미한다. 실제로 2021년 2월 한 달간 서울에서 확진자수는 3일, 6일, 11일, 17일 그리고 24일 등 모두 5번의 고점을 찍는데, 예측 그래프의 고점일과 실제 그래프의 고점일은 정확히 일치하거나 하루 정도의 오차만이 발생했다. 부산에서는 2021년 2월 한 달 동안 5일, 11일, 13일, 18일 그리고 25일 등 총 5번의 고점을 찍는데, Fig. 4의 부산에서 주황색 실선을 보면 예측 그래프의 변화 추세와 실제 그래프의 변화 추세가 거의 일치함을 입증해주고 있다. 특히 예측 그래프와 실제 그래프 사이에 고점일 뿐만 아니라 저점일도 거의 유사함을 확인할 수 있다. 이처럼 예측 그래프 가운데 검색어 빈도를 포함하는 그래프와 실제 그래프 사이에 변화 추세의 일치 정도가 더욱 뚜렷했다.

4.2 LSTM 모델을 적용한 예측

LSTM 모델에서도 확진자수를 예측하기 위해 입력되는 데이터 세트는 DNN 모델에서와 같이 2가지 형태(확진자수 vs 확진자수 + 검색어 빈도)이다.

Table 4에서와 같이 6개 도시 중 서울과 인천, 대구에서는 확진자수만을 사용한 데이터 세트에서 오류율이 더 낮았다.

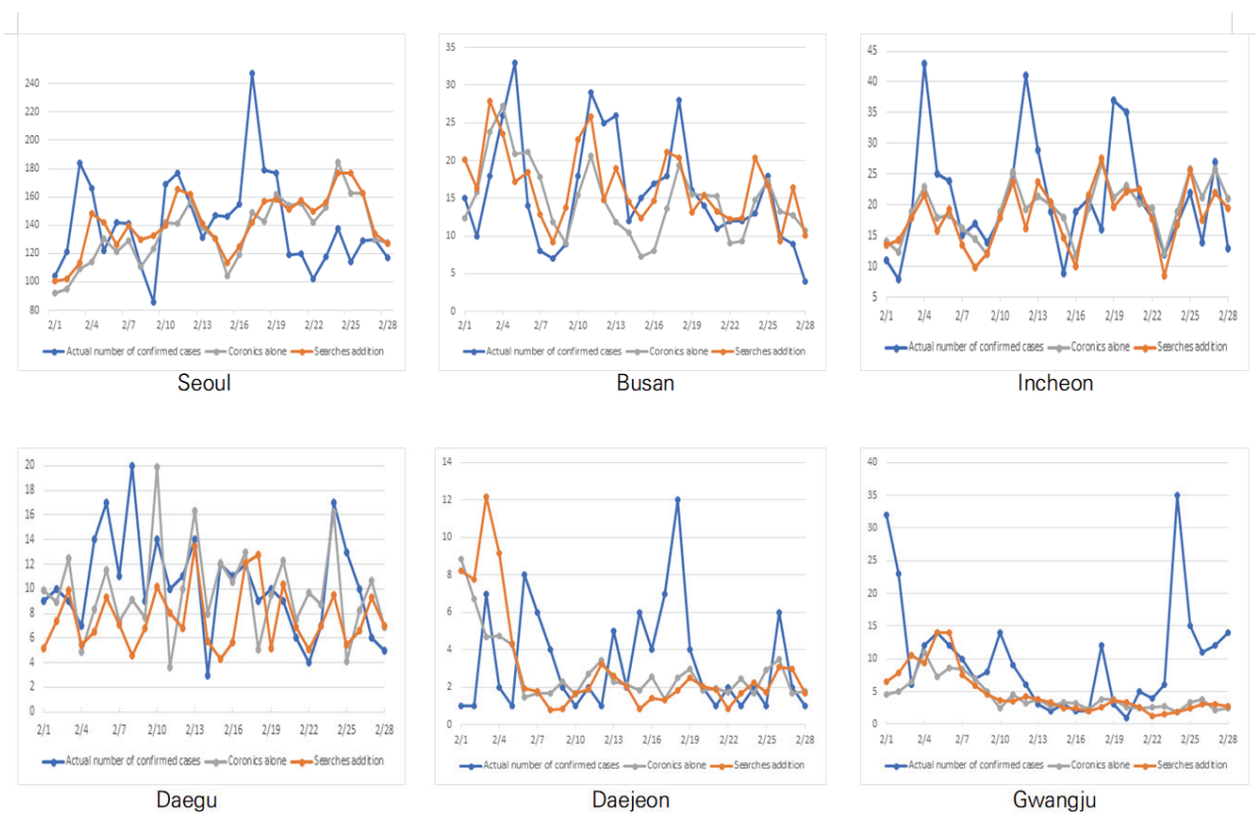

Fig. 4. A Graph of the Actual Number of Confirmed Cases(Blue Line), a Graph of the Number of Confirmed Cases Predicted Using the 'Coronics Alone' Set(Gray Line), and a Graph of the Number of Confirmed Cases Predicted Using the 'Searches Addition' Set(Orange Line) in Six Cities in February 2021(with DNN Model).

Table 4. Error Rate When Predicting the Number of Confirmed Cases by Applying the LSTM Model

Names of input data set	Combinations	SEL	BUS	INC	DAE	DAJ	GWJ
Coronics alone	coronics	20.0	35.1	31.7	36.3	96.0	94.0
Searches addition	coronics+searches	21.2	32.2	32.3	39.0	90.0	84.7

반면 부산과 대전, 광주에서는 검색어 빈도 데이터를 추가한 데이터 세트에서 오류율이 더 낮았다. 하지만 서울과 인천의 경우 오류율의 차이가 1% 포인트 안팎인 점을 고려하면 6개 도시 가운데 대구를 제외한 5개 도시에서는 검색어 빈도를 포함하는 데이터 세트에서 오류율이 더 낮아지거나 확진자수 데이터 만을 가지고 예측한 경우와 큰 차이가 없는 것으로 분석됐다. 또한 DNN 모델에서와 유사하게 인구수 250만 명 이상의 도시에서는 오류율이 35%를 넘지 않았지만, 250만 명 이하의 도시에서는 35%를 초과하는 것으로 나타났다.

또한, 250만 명 이상의 도시에서 검색어 빈도를 포함하는 데이터 세트를 가지고 확진자수를 예측했을 때도 확진자수 예측 그래프와 실제 확진자수 그래프 사이에 변화 추세가 거의 유사했다(Fig. 5 참조).

Fig. 5의 서울에서 주황색 실선을 보면 실제그래프와 예측 그래프 사이에 고점일 뿐만 아니라 저점일도 거의 일치했다. 또한 Fig. 5의 인천에서 주황색 실선을 보면 5번째 고점일(2월 25일)의 확진자수가 거의 정확하게 예측됐다.

5. 논의 및 결론

본 연구는 과거 확진자수와 검색어 빈도 데이터를 사용하여 5일 후의 코로나 확진자 수를 미리 예측할 수 있을지 탐색적 데이터 분석을 실시했다. 딥러닝을 통한 확진자수 예측 결과 서울은 오류율이 20% 내외, 부산과 인천에서는 30% 내외를 기록했다.

정부와 학계도 지난해 12월 1일 2주 후(2022년 12월 14일)의 확진자수를 예측한 '수리모델링으로 분석한 코로나 19 유행 예측' 보고서를 발표했다[44]. 이들 연구팀은 KT 이동통신 데이터로부터 집계된 지역 간 이동량과 과거 확진자수를 고려해 2주 후의 확진자수를 예측했다. 국가수리과학 연구소(NIMS)는 11

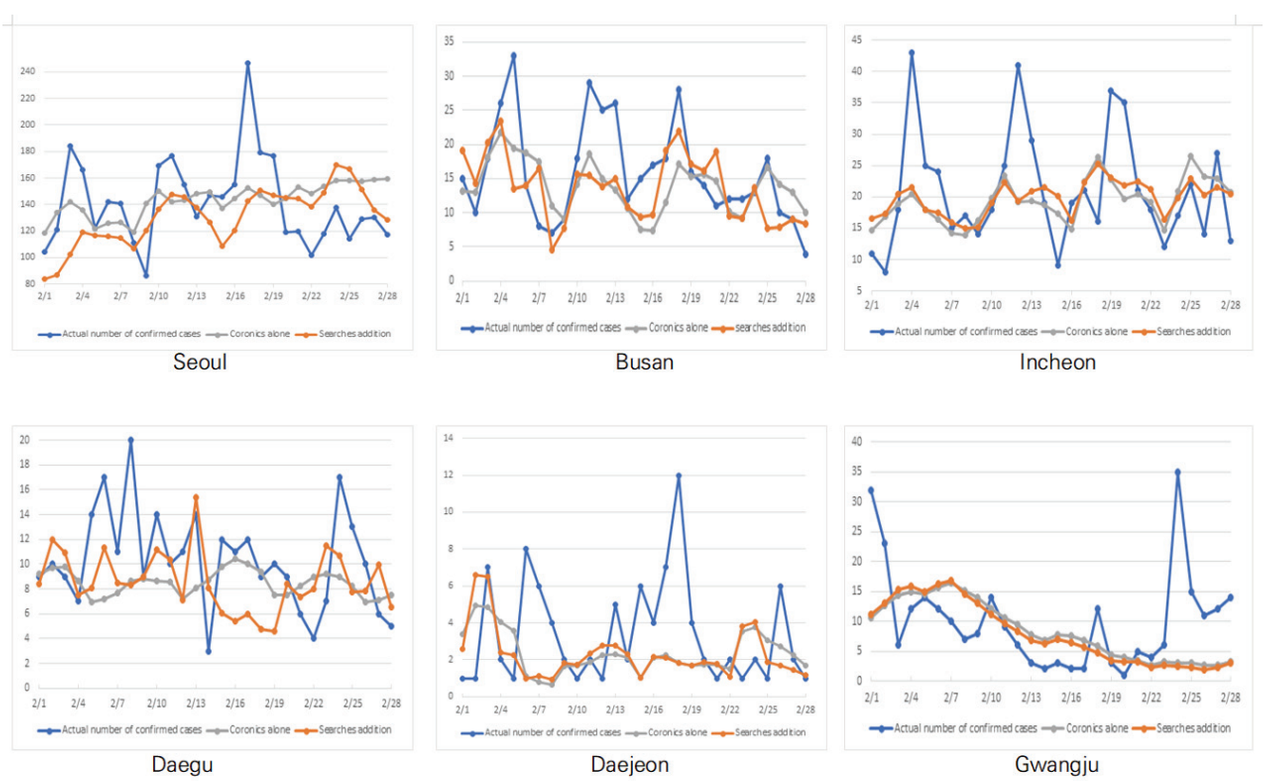

Fig 5. A Graph of the Actual Number of Confirmed Cases(Blue Line), a Graph of the Number of Confirmed Cases Predicted Using the 'Coronics Alone' Set(Gray Line), and a Graph of the Number of Confirmed Cases Predicted Using the 'Searches Addition' Set(Orange Line) in Six Cities in February 2021(with DNN Model)

Table 5. Error Rate When Predicting the Number of Confirmed Cases by Applying Mathematical Modeling

	NIMS	UNIST	Soongsil
MAPE(%)	70.4	42.8	22.3

월 30일 6만 7415명이던 신규 확진자수가 2주 후 2만 5000명 수준, 울산과학기술원(UNIST)은 4만 8401명, 숭실(Soongsil) 대학교 연구팀은 6만 5666명으로 각각 줄어들 것으로 전망했다. 이와 같은 예측치는 본 연구에서 사용한 오류율(MAPE)로 계산할 경우 22.3%에서 70.4%에 해당한다(Table 5 참조).

물론 정부와 학계의 예측 성능과 본 연구의 예측 성능을 결과만을 가지고 비교하기는 힘들다. 데이터의 특성, 입력 데이터의 기간, 예측 시기 그리고 예측에 사용된 방법론이 모두 다르기 때문이다. 하지만, 1년 정도의 데이터만을 가지고 특히 서울과 부산, 인천 등에서 기록한 20%에서 30% 내외의 오류율 수준은 정책 결정자들이 충분히 참고할 만한 수준의 예측력이라고 생각한다.

이 연구를 통해 우리는 미래의 전염병 확진자수 연구에 기여할 수 있는 3가지 주목할 만한 결과를 확인할 수 있었다. 첫째, 검색어 빈도 데이터는 전염병 초기의 감염자수 예측 연구에 여전히 영향력이 있음을 확인할 수 있었다.

DNN모델을 적용한 경우, 250만 명 이상의 3개 도시 중 2개 도시(서울과 부산)에서 검색어 빈도를 추가한 데이터 세트를 통해 확진자수를 예측했을 때 오류율이 낮아졌다. 나머지 한 개 도시(인천)에서는 오류율이 낮아지지는 않았지만, 확진자수 만을 가지고 예측했을 때와 큰 차이가 없었다(1% 포인트 미만). LSTM 모델을 적용한 경우에는, 대구를 제외한 5개 도시에서 검색어 빈도를 추가한 데이터 세트를 가지고 예측했을 때 오류율이 낮아지거나 확진자수 만을 가지고 예측했을 때와 큰 차이가 없었다. 이와 같은 분석 결과는 검색어 빈도 데이터가 확진자수 예측에서 중요한 역할을 하고 있음을 보여주는 것이다.

그렇다면 왜 DNN 모델에서는 검색어 데이터가 특히 250만 명 이상의 도시에서 오류율을 낮추는데 더 효과적이었을까? 서론에서 언급했듯이 '코로나 증상' 이란 검색어를 사용하는 빈도수가 증가한다면 물론 단순히 지적 욕구 때문에 검색하는 경우도 있을 수 있다. 하지만 빈번한 대면 접촉을 한 사람들, 대중 교통을 이용한 사람들, 또는 번잡한 식당을 방문한 사람들이 코로나에 감염되었을 가능성을 우려한 나머지 감염되면 어떠한 증상이 나타나는지 알기 위해 검색했을 가능성이 더욱 높아지는 것이다. 그리고 대면 접촉에 자주 노출된 경우는 코로나 19에 감염됐을 가능성을 증가시키는 조건과 밀접히 관련돼 있다. 따라서 검색어 빈도 데이터는 이와 같은 상황을 간접적으로 알려주는 것이다. 인플루엔자와 관련된 검색어의 영향력은 주(state) 단위 수준의 인구수와 매우 높은 상관관계가 있음이 확인되었다[45]. 따라서, 인터넷 사용자가 많으면 많을수록, 정보를 검색하는 활동이 의미 있는 시그널(signal)을 발생시킬 가능성을 더욱 높여주는 것이고[46, 47], 이러한 이유로 본 연구에서도 250만 명 이상의 도시에서 검색어 데이터가 더 큰 영향력을 발휘하였다. 즉 검색어 빈도 데이터만으로는 확진자수 예측에 큰 영향력을 미칠 수 없었으며 어느 정도의 인구수가 담보되었을 때 데이터 속에 잠재된 시그널이 활동하며 영향력을 미치게 되는 것이라고 생각해 볼 수 있겠다.

반면 LSTM 모델을 적용한 경우, 본 연구와 같은 시계열 데이터 분석에 특화된 모델이기 때문에 검색어 빈도를 추가한 데이터 세트를 가지고 확진자수를 예측했을 때 250만 명 이하의 도시에서도 오류율이 더 낮아진 것이라는 예측이 가능하다.

둘째, 누구나 사용할 수 있는 공개적 데이터를 가지고 이와 같은 예측력을 달성할 수 있었다는 것이다. 코로나 19는 물론 전염병의 확진자수를 예측하는 많은 연구가 임상 데이터를 사용하고 있다[23, 48]. 하지만, 이와 같은 임상 데이터는 사생활 침해의 가능성 때문에 많은 논란이 되고 있다. 따라서 이 연구에서는 개인 정보 침해의 우려가 전혀 없고 누구나 사용할 수 있는 구글 트렌드를 사용하여 오류율이 낮아질 수 있는지 확인하고 싶었고 실제로 낮아질 수 있음을 확인하였다.

셋째, 새로운 기술이 적용되지 않아도 시계열 데이터에 특화된 예측 모델을 통해 강력한 예측 성능을 달성할 수 있었다. 서울과 부산 등에서의 낮은 오류율은 새로운 알고리즘 개발 같은 기술적 발전 때문에 달성된 것은 아니었다. 3장 모델 구성에서 언급한 것처럼 본 연구에 사용된 예측 모델들은 온라인에서 누구나 사용할 수 있는 오픈 소스를 사용해 개발됐다. 특히 은닉층을 만들고 몇몇 파라미터들을 튜닝하는데도 딥러닝의 가장 기본적인 방법들이 사용되었다.

따라서 후속 연구에서는 최근에 개발된 알고리즘을 사용하거나 더 좋은 성능을 만들어낼 수 있는 파라미터를 적용해 예측 성능을 비교해 볼 것을 제안하는 바이다. 예측 성능이 추가로 향상될 여지는 충분하다고 생각된다. 또한, 중소 도시에서의 오류율도 서울 같은 대도시 수준으로 떨어뜨리기 위해서는 작은 인구 이동에도 정확하게 반응할 수 있는 새로운 데이터를 발굴하여 예측 능력을 향상시키는 것이 필요하겠다.

코로나 바이러스는 아직 종식된 것이 아니다. 코로나 바이러스에 대처하기 위한 백신과 치료제가 나오고 있지만 동시에 많은 변종 바이러스가 등장해 일상으로 돌아가려는 우리의 노력을 지속적으로 방해할 것이다. 본 연구에 사용된 확진자수 예측 모델을 이러한 후속 연구 등을 통해 계속해서 발전시킨다면 추후 또 다른 팬데믹이 발생할 때 정책 결정권자들에게 시기적절하고 더욱 신뢰할 수 있는 확진자수 추정치를 제공할 수 있을 것이다.

References

[1] M. Vallee, "Doing nothing does something: Embodiment and data in the COVID-19 pandemic," *Big Data & Society*, Vol.7, No.1, pp.1-12, 2020.

[2] Centers for Disease Control and Prevention. How COVID-19 Spreads, 2021. Available at www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html.

[3] F. Petropoulos and Makridakis, S. "Forecasting the novel coronavirus COVID-19," *PloS one*, Vol.15, No.3, pp. e0231236, 2020.

[4] J. Stubinger and L. Schneider, "Epidemiology of coronavirus covid-19: Forecasting the future incidence in diferent countries," *Healthcare*, Vol.8, No.2, pp.99, 2020.

[5] A. Tobías, "Evaluation of the lockdowns for the SARS-CoV2 epidemic in Italy and Spain after one month follow up," *Science of the Total Environment*, Vol.725, pp.138539, 2020.

[6] Y. Li, M. Liang, X. Yin, X. Liu, M. Hao, Z. Hu, Y. Wang, and L. Jin, "COVID-19 epidemic outside China: 34 founders and exponential growth," *Journal of Investigative Medicine*, Vol.69, No.1, pp.52-55, 2021.

[7] T. Chakraborty and I. Ghosh "Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis," *Chaos, Solitons & Fractals*, Vol.135, pp.109850, 2020.

[8] M. Perc, N. Gorišek Miksić, M. Slavinec, and A. Stožer, "Forecasting covid-19," *Frontiers in Physics*, Vol.8, pp.127, 2020.

[9] S. J. Fong, G. Li, N. Dey, R. G. Crespo, and E. Herrera Viedma, "Finding an accurate early forecasting model from small dataset: A case of 2019-nCoV novel coronavirus outbreak," *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol.6, No.1, pp.132-140, 2020.

[10] Z. Yang et al., "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," *Journal of Thoracic Disease*, Vol.12, No.3, pp.165, 2020

[11] B. Pirouz, S. Shaffiee Haghshenas, S. Shaffiee Haghshenas, and P. Piro, "Investigating a serious challenge in the sustainable development process: Analysis of confirmed cases of COVID-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis," *Sustainability*, Vol.12, No.6, pp.2427, 2020.

[12] L. Qin, Q. Sun, Y. Wang, K. F. Wu, M. Chen, B. C. Shia, and S. Y. Wu, "Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index," *International Journal of Environmental Research and Public Health*, Vol.17, No.7, pp.2365, 2020.

[13] M. Santillana, E. O. Nsoesie, S. R. Mekaru, D. Scales, and J. S. Brownstein, "Using clinicians' search query data to monitor influenza epidemics," *Clinical Infectious Diseases*, Vol.59, No.10, pp.1446-1450, 2014a.

[14] S. Yang, M. Santillana, and S. C. Kou, "Accurate estimation of influenza epidemics using Google search data via ARGO," *Proceedings of the National Academy of Sciences*, Vol.112, No.47, pp.14473-14478, 2015.

[15] M. Helft, "Google uses searches to track flu's spread," The New York Times, 11 November, 2008. Available at https://www.nytimes.com/2008/11/12/technology/internet/12flu.html?_r=0#.

[16] S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi, "Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic." *PloS one*, Vol.6, No.8, pp.e23610, 2011.

[17] D. Butler, "When Google got flu wrong," *Nature News*, Vol.494, pp.155-156, 2013.

[18] M. Santillana, D. W. Zhang, B.M. Althouse, and J. W. Ayers, "What can digital disease detection learn from(an external revision to) Google Flu Trends?" *American Journal of Preventive Medicine*, Vol.47, No.3, pp.341-347, 2014b.

[19] World Health Organization, "Transmission of SARS-CoV-2: implications for infection prevention precautions," 2020. Available at https://www. who. int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions.

[20] World Health Organization, "Coronavirus disease (COVID19): How is it transmitted?," 2021. Available at https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted.

[21] A. de Fátima Cobre et al., "Diagnosis and prediction of COVID-19 severity: Can biochemical tests and machine learning be used as prognostic indicators?," *Computers in Biology and Medicine*, Vol.134, pp.104531, 2021.

[22] A. Daoud, R, Kim, and S. V. Subramanian, "Predicting women's height from their socioeconomic status: A machine learning approach," *Social Science & Medicine*, Vol.238, pp.112486, 2019.

[23] N. L. Bragazzi, H. Dai, G. Damiani, M. Behzadifar, M. Martini, and J. Wu, "How big data and artificial intelligence can help better manage the COVID-19 pandemic," International Journal of Environmental Research and Public Health, Vol.17, No.9, pp.3176, 2020.

[24] D. W. Seo and S. Y. Shin, "Methods using social media and search queries to predict infectious disease outbreaks," *Healthcare Informatics Research*, Vol.23, No.4, pp.343-348, 2017.

[25] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data." *Nature*, Vol.457, No.7232, pp.1012-1014, 2009.

[26] D. J. McIver and J. S. Brownstein, "Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time," *PLoS Computational Biology*, Vol.10, No.4, pp.e1003581, 2014.

[27] E. H. Chan, V, Sahai, C, Conrad, and J. S. Brownstein, "Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance," *PLoS Neglected Tropical Diseases*, Vol.5, No.5, pp.e1206, 2011.

[28] S. Yousefinaghani, R. Dara, S. Mubareka, and S. Sharif, "Prediction of COVID-19 waves using social media and Google search: A case study of the US and Canada," *Frontiers in Public Health*, Vol.9, pp.656635, 2021.

[29] S. Ben et al., "Global internet search trends related to gastrointestinal symptoms predict regional COVID-19 outbreaks," *Journal of Infection*, Vol.84, No.1, pp.56-63, 2022.

[30] S. Prasanth, U. Singh, A. Kumar, V. A. Tikkiwal, and P. H. Chong, "Forecasting spread of COVID-19 using google trends: A hybrid GWO-deep learning approach," *Chaos, Solitons & Fractals*, Vol.142, pp.110336, 2021.

[31] Z. Pan, H. L. Nguyen, H. Abu-Gellban, and Y. Zhang, "Google trends analysis of covid-19 pandemic," In *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, pp.3438-3446, 2020.

[32] J. Brownlee, "How to control the stability of training neural networks with the batch size," In: Machine Learning Mastery, 2020. Available at: https://machinelearningmastery.com/how-to-control-the-speed-and-stability-of-training-neural-networks-with-gradient-descent-batch-size/

[33] Tensorflow: Recurrent Neural Networks (RNN) with Keras, 2021. Available at: https://www.tensorflow.org/guide/keras/rnn.

[34] F. Chollet, "Deep learning with Python," Shelter Island: Manning Publications Co., 2017.

[35] N. A. Zambri, A. Mohamed, and M. Z. C. Wanik, "Performance comparison of neural networks for intelligent management of distributed generators in a distribution system," *International Journal of Electrical Power & Energy Systems*, Vol.67, pp.179-190, 2015.

[36] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records." *Scientific Reports*. Vol.6, No.26094, pp.1-10, 2016.

[37] D. Hudgeon and R. Nichol, "Machine learning for business: Using Amazon SageMaker and Jupyter," Shelter Island:Manning Publications Co., 2019.

[38] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *International Journal of Forecasting*, Vol.32, No.3, pp.669-679, 2016.

[39] J. Lago, F. De Ridder, and B. De Schutter, "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms," *Applied Energy*, Vol.221, pp.386-405, 2018.

[40] J. L. Leevy, T. M. Khoshgoftaar, and F. Villanustre, "Survey on rnn and crf models for de-identification of medical free text," *Journal of Big Data*, Vol.7, No.73, pp.1-22, 2020.

[41] A. Géron, "Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems," Sebastopol: O'Reilly Media, 2019.

[42] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv: 1412.3555*, 2014.

[43] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," *International Conference on Machine Learning*, pp. 2342-2350, 2015.

[44] R. Han, "COVID-19 confirmed after 2 weeks↓" Mathematicians predict…Quarantine authorities cautious. (2022. 12.01). Retrieved 12/28/2022 from https://news.jtbc.co.kr/article/article.aspx?news_id=NB12105367

[45] F. S. Lu, M. W. Hattab, C. L. Clemente, M. Biggerstaff, and M. Santillana, "Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches," *Nature Communications*, Vol.10, No.1, pp.1-10, 2019.

[46] L. Poole, “Seasonal influences on the spread of SARSCoV-2 (COVID19), causality, and forecastabililty,” 2020. Available at http://dx.doi.org/10.2139/ssrn.3554746.

[47] P. Pequeno et al., “Air transportation, population density and temperature predict the spread of COVID-19 in Brazil,” *PeerJ*, Vol.8, pp.e9322, 2020.

[48] C. Poirier et al., “Influenza forecasting for the French regions by using EHR, web and climatic data sources with an ensemble approach ARGONet,” medRxiv: 19009795, 2019.

정 성 욱

https://orcid.org/0000-0003-4284-4052
e-mail : jj4863@naver.com
2009년 Columbia Univ.(석사)
2020년 서울대학교 언론정보학과(박사)
2021년 ~ 현 재 서울대학교 언론정보연구소 선임연구원
2022년 ~ 현 재 성균관대학교 강사
관심분야 : Data Journalism & Media Big Data Analysis