J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

다중 웹 데이터와 LSTM을 사용한 전염병 예측☆

Prediction of infectious diseases using multiple web data and LSTM

김 영 하[1] 김 인 환[1] 장 백 철[1*]
Yeongha Kim Inhwan Kim Beakcheol Jang

요 약

전염병은 오래전부터 인류를 괴롭혀 왔으며 이를 예측 하고 예방하는 것은 인류에게 있어 큰 과제였다. 이러한 이유로 지금까지도 전염병을 예측하기 위해 다양한 연구가 진행되고 있다. 초기의 연구 중 대부분은 CDC(Centers for Disease Control and Prevention)의 역학 데이터에 의존한 연구였으며, CDC에서 제공하는 데이터는 일주일에 한 번만 갱신돼 실시간 질병 발생 건수를 예측하기 어렵다는 문제점을 갖고 있었다. 하지만 최근 IT 기술의 발전으로 여러 인터넷 매체들이 등장하면서 웹 데이터를 통해 전염병의 발생을 예측하고자 하는 연구가 진행되었고 이 중 우리가 조사한 연구 중 대부분은 단일 웹 데이터를 사용하여 질병을 예측하는 연구였다. 하지만 단일 웹 데이터를 통한 질병 예측은 "COVID-19" 같이 최근에 등장한 전염병에 대해서는 많은 양의 학습 데이터를 수집하기 어려우며 이러한 모델을 통해 정확한 예측을 하기 어렵다는 단점을 가지고 있다. 이에 우리는 전염병 발생을 LSTM 모델을 통해 예측할 때 여러 개의 웹 데이터를 사용하는 모델이 단일 웹 데이터를 사용하는 모델보다 정확도가 더 높음을 실험을 통해 증명하고 전염병 예측에 적절한 모델을 제안하고자 한다. 본 실험에서는 단일 웹 데이터를 사용하는 모델과 우리가 제안하는 모델을 사용하여 "말라리아"와 "유행성이하선염"의 발생을 예측했다. 우리는 2017년 12월 31 일부터 2019년 12월 28일까지 총 104주 분량의 NEWS, SNS, 검색 쿼리 데이터를 수집했는데, 이 중 75주는 학습 데이터로, 29주는 검증 데이터로 사용됐다. 실험 결과 우리가 제안한 모델의 예측 결과와 단일 웹 데이터를 사용한 모델의 예측 결과를 비교했을 때 검증 데이터에 대해서 피어슨 상관계수가 0.94, 0.86로 가장 높았고 RMSE 또한 0.19, 0.07로 가장 낮은 오차를 보여주었다.

☞ 주제어 : 머신러닝, 전염병 예측, 웹 데이터, LSTM

ABSTRACT

Infectious diseases have long plagued mankind, and predicting and preventing them has been a big challenge for mankind. For this reasen, various studies have been conducted so far to predict infectious diseases. Most of the early studies relied on epidemiological data from the Centers for Disease Control and Prevention (CDC), and the problem was that the data provided by the CDC was updated only once a week, making it difficult to predict the number of real-time disease outbreaks. However, with the emergence of various Internet media due to the recent development of IT technology, studies have been conducted to predict the occurrence of infectious diseases through web data, and most of the studies we have researched have been using single Web data to predict diseases. However, disease forecasting through a single Web data has the disadvantage of having difficulty collecting large amounts of learning data and making accurate predictions through models for recent outbreaks such as "COVID-19". Thus, we would like to demonstrate through experiments that models that use multiple Web data to predict the occurrence of infectious diseases through LSTM models are more accurate than those that use single Web data and suggest models suitable for predicting infectious diseases. In this experiment, we predicted the occurrence of "Malaria" and "Epidemic-parotitis" using a single web data model and the model we propose. A total of 104 weeks of NEWS, SNS, and search query data were collected, of which 75 weeks were used as learning data and 29 weeks were used as verification data. In the experiment we predicted verification data using our proposed model and single web data, Pearson correlation coefficient for the predicted results of our proposed model showed the highest similarity at 0.94, 0.86, and RMSE was also the lowest at 0.19, 0.07.

☞ keyword : Machine Learning, Predict infectious diseases, Web data, LSTM

[1] Department of Computer Science, Sangmyung University, Seoul, 03016, Korea
* Corresponding author (bjang@smu.ac.kr)
[Received 30 June 2020, Reviewed 7 August 2020(R2 7 September 2020), Accepted 24 September 2020]

☆ This work was supported by the National Research Foundation of Korea Grant funded by the Korea Government under Grant NRF-2019R1F1A1058058.

1. 서 론

과거부터 전염병은 인류에게 있어 재앙 그 자체였으며 인류의 역사에 매우 큰 영향을 끼쳐왔다. 그 예시로 인류를 위협한 전염병 중 하나인 흑사병은 1300년대 유럽을 강타하여 4~5년 만에 유럽 인구의 3분의 1 이상의 목숨을 괴롭혀 왔으며 전염병을 예측하고 예방하는 것은 인류에게 있어 큰 과제였다. 이러한 이유로 지금까지도 전염병을 예측하고 예방하기 위해 다양한 연구가 활발히 진행됐다.

초기에는 전염병 예측과 관련한 대부분의 연구가 CDC(Centers for Disease Control and Prevention)로부터 제공되는 질병 발생 데이터에 의존해 왔으나 CDC로부터 제공되는 질병 데이터는 1주일에 1번씩만 업데이트되기 때문에 실시간으로 발생하는 전염병의 예측이 어려웠다. 하지만 IT 기술이 발전하면서 다양한 인터넷 매체들(인터넷 NEWS, SNS, 검색 엔진)이 등장하였으며 이러한 매체들은 전염병이 발생했을 때 대부분 즉각적인 변화를 보이기 때문에 실시간으로 전염병을 예측하기에 용이했다. 따라서 CDC 데이터에만 의존하던 기존의 전염병 연구들은 웹 데이터를 통해 전염병을 예측하는 방향으로 연구가 진행되었다.

그 예시로 [2]는 소셜 미디어 서비스(SNS)와 빅 데이터를 통해 수두, 성홍열, 말라리아 등의 질병을 예측하는 실험을 진행하였으며, [3]은 2019년 우한에서 발생한 COVID-19 질병을 트위터 데이터와 나이브 베이즈, SVM, Decision Tree, LogitBoost, Random Forests 등의 모델을 사용하여 질병을 예측하였다. [4]는 트위터 데이터와 BOW(Bag of Word) 와 마르코프 체인 (Markov Chain State)를 사용하여 트위터에서 키워드를 분류하여 인플루엔자 발병을 예측하는 모델을 제안했다. [5]는 소셜 미디어 서비스(SNS)를 사용하여 지카 바이러스의 발병을 예측하는 실험을 진행했다.

그 이외에도 단일 웹 데이터를 통해 질병을 예측하고자 하는 다양한 연구들이 존재한다 [6][7][8][9].

그리고 2가지 이상의 여러 개의 웹 데이터를 같이 사용하여 질병을 예측하는 연구들도 있었다. [10]은 2개의 신경망 구조를 사용하여 계절 내 관측과 계절 간 관측 데이터를 사용하여 질병을 예측한다. [11]은 실시간으로 인플루엔자의 발병을 예측하는 새로운 모델인 Att-MCLSTM을 제시하였다. 이 모델은 기존의 LSTM 모델을 기반으로 한 새로운 모델로 평균 온도, 최대 온도, 최소 온도, 강우량, 기압, 상대습도 등을 같이 사용하여 질병 발생을 예측한다. 하지만 이러한 연구 중에 단일 웹 데이터를 사용한 연구들은 "COVID-19" 같이 최근에 등장한 질병에 대해서는 데이터의 양이 한정되어 있으므로 많은 양의 데이터를 수집하기 어려우며 예측 정확도가 떨어진다는 단점이 있다.

본 연구에서는 단일 웹 데이터로 질병을 예측하는 것보다는 여러 종류의 웹 데이터를 통해 질병을 예측할 때 더 좋은 결과를 얻을 수 있음을 실험을 통해 증명하고 전염병 예측에 적합한 모델을 제안하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문의 연구와 관련된 연구들을 소개하고, 3장에서는 우리가 실험에서 사용할 두 개의 모델의 구조와 특징, 모델에 사용된 하이퍼 파라미터에 대해 설명한다. 4장에서는 본 연구의 실험 과정에 대해 설명하며 5장은 실험 결과에 대한 분석과 결론, 앞으로의 연구 목표를 제시한다.

2. 관련 연구

2.1 SNS 데이터

[2]에서는 소셜 미디어를 포함한 빅 데이터와 DNN, OLS, LSTM, ARIMA 모델을 사용하여 수두, 성홍열, 말라리아 등의 질병 발생을 예측하는 실험을 진행했다. 실험에서는 DNN과 LSTM 모델을 사용해 1주일 동안 3가지 감염병을 예측하고 ARIMA 모델과 비교했더니 DNN과 LSTM 모델이 ARIMA보다 우수한 성능을 보인 것으로 나타났다. 수두 예측 시 DNN과 LSTM 모델은 ARIMA 모델과 비교했을 때 정확도가 24%, 19% 더 높았으며, LSTM 모델은 전염병이 퍼지고 있을 때 더 정확한 결과를 보여주었다.

[3]은 2019년 우한에서 발생한 COVID-19 를 트위터 데이터와 나이브 베이즈, SVM, Decision Tree, LogitBoost, Random Forests 등의 모델을 사용하여 질병을 예측하였다. 실험에서 트윗의 감정 점수를 반영하여 데이터에 포함시켰으며, LogitBoost 모델이 74%의 가장 높은 정확도를 보여주었다.

[4] 은 트위터 데이터를 가지고서 인플루엔자 질병을 예측하는 방법을 제시한다. 중국의 모든 대도시에서 공유된 트위터 데이터를 기반으로 수집하였으며 트위터 API를 사용하였다. 또한 BOWs(Bag Of Words)를 기반으로 한 MarkovChainState모델을 사용하여 키워드를 3가지로 분류(질병의 시작, 질병의 퍼짐, 질병 차단)하여 질병을 예측하는 모델을 제안하였다.

[5]은 소셜 미디어 서비스(트위터)를 사용하여 지카 바이러스의 발병을 예측하고 예방하는 기술을 제안한다.

CDC에서 질병 발병에 관한 데이터를 수집하고 연관 트윗을 통해 관련 데이터를 수집하였다. 실험 결과 주석이 달린 데이터에 대해 로지스틱 회귀 모델을 훈련했을 때 72%의 정확도를 얻을 수 있었다.

2.2 검색 쿼리 데이터

[6]에서는 홍콩에서 발생한 인플루엔자를 구글 검색 쿼리 데이터와 기상 데이터를 통해 질병 발생을 예측한다. GLM, LASSO, ARIMA, FNN을 이용한 딥러닝(DL) 총 4개의 모델과 각 예측 모델을 합쳐 평균을 내는 BMA가 사용되었으며 인플루엔자 발병을 1주 2주 전에 예측하였다. 1주를 기준으로 예측했을 때 GLM 모델이 65%, BMA는 나머지 4개 모델보다 더 정확한 73%의 정확도를 보여주었다. 하지만 인플루엔자가 유행하는 시즌에 예측한 데이터는 이보다 다소 떨어지는 61%의 정확도를 보였다.

[7]에서는 구글 트렌드 웹사이트에서 이란에서 검색된 코로나 19와 관련한 검색 비율을 수집하고. 선형 회귀 분석 모델과 LSTM 모델을 사용하여 환자 수를 예측했다. 선형 회귀 모델은 RMSE 7.562의 발생률을 예측하였고 LSTM 모델의 RMSE는 27.187이다. 하지만 LSTM 모델에서는 적은 양의훈련 데이터로 인해 과적합 현상이 일어났으나 학습오류가 비교적 적음을 보여주었고 이는 LSTM 모델이 테이터에서 패턴을 추출할 수 있음을 보여 주었다.

[8]은 중국 랴오닝성에서 공식적으로 보고된 2011년 1월부터 2015년 12월까지의 검색 데이터와 기존 인플루엔자 데이터를 병합하여 예측할 때 SVM(Support Vector Machine) regression 모델을 사용하여 예측한다. SVM 모델은 인플루엔자 감시 데이터와 바이두 검색 쿼리 데이터를 통합한 데이터를 기반으로 파라미터($C = 2$, $\gamma = 0.005$, $\varepsilon = 0.0001$)를 사용했을 때 좋은 결과를 얻었다.

[9]에서는 2011년 1 월부터 2017년 6월까지의 월별 시계열 데이터를 기반으로 바이두에서 검색된 에이즈 관련 검색어를 조사하여 MLP 모델을 사용하여 예측하였다. 입력 데이터 중 80%가 훈련 데이터로 사용되었고 나머지 10%씩 검증과 검증 데이터로 사용되었다. 훈련 결과 MAPE와 RMSPE가 모두 0.05보다 작고 IA(Index of Agreement)가 0.68보다 컸으며 이는 매우 좋은 예측 결과를 보여줬다.

2.3 다중 웹 데이터

[10] 은 두 개의 신경망 구조를 구축하여 계절 내 관측 및 계절 간 관측 데이터를 기반으로 단기적이지만 고해상도 예측이 가능하게 하는 합성정보를 이용하여 전염병 질병을 예측한다. 위 논문은 두 개의 LSTM 신경망을 사용하였고 하나는 계절 내 관측치를 인코딩하는 누적 LSTM 계층이고 나머지 하나는 LSTM으로 설계된 단일 계층이며 두 모델을 합쳐서 평균을 내는 MARGE 레이어를 통해 두 개의 신경망을 하나로 합쳐 결과를 산출하는 모델을 사용한다.

[11] 은 중국 광저우에서 발생하는 인플루엔자를 예측하기 위해서, 기존의 LSTM 모델을 기반으로 한 새로운 모델인 Att-MCLSTM을 제시하였다. 위 실험에서는 모델에 들어갈 데이터를 두 가지로 구분한다, 첫 번째는 평균 온도, 최대 온도, 최소 온도, 강우량, 기압, 상대습도를 기후 관련 데이터 범주로 분류하였고 나머지 특징들은 함께 인플루엔자 관련 데이터로 분류된다. 위 실험에선 총 10주간의 기간을 두고 여러 모델을(Att-MCLSTM, MCLSTM, LSTM, RNN) 사용하여 결과를 예측했을 때 Att-MCLSTM이 MAPE 0.0086으로 가장 좋은 값을 보였다.

[12] 은 질병 추정치와 기계학습 방법론을 통해 실시간으로 중국 내 32개의 지방에서 발생하는 COVID-19 활동을 정확하게 예측하는 방법을 제시한다. 학습 데이터로는 중국 CDC에서 제공하는 질병 데이터, 바이두의 COVID-19 관련 인터넷 검색활동, 뉴스 미디어를 수집하였으며 부트스트랩 방법을 사용하여 학습 데이터를 샘플링하고 LASSO 다변량 정규 선형 모델을 사용하여 질병을 예측하였다.

2.4 기타

[13]에서는 미국 지역의 지리적 특징을 포착하고, 인플루엔자의 시간적 역학을 포착한다. [13] 의 실험에서는 기존 모델 CNNRNN-res 과 GCNGRU 의 성능을 비교하는 실험을 진행했다. 각 모델의 최적 파라미터를 찾아서 모델을 훈련시켰더니 87.66%, 80.33%의 정확도를 보였으며, 이는 GCNGRU 모델이 기존의 CNNRNN-res 모델보다 더 정확한 모델임을 보여주었다.

[14]는 소셜 네트워크 서비스(SNS) 데이터와 강수량을 사용하여 말라리아의 발병을 예측한다. SNS 데이터는 트위터 Twitter search Application Programming Interface(API)를 통해 수집하였으며 SVM(Support Vector Machine) 분류 알고리즘을 사용하여 예측한 결과 0.75의 상관관계가 나왔다.

[15]에서는 6개의 서로 다른 모델에서 인플루엔자 예측

성능을 조사했다. 예측 모델로는 ARIMA, SVR, RF, GB, ANN, LSTM 모델이 사용되었다. 총 52주의 기간의 질병을 예측하였고, 모든 모델에서 4계층의 LSTM 모델은 최저 MAPE 5.4%에 도달했으며 정규화를 사용하는 5계층의 LSTM 모델의 예측 결과는 RMSE가 0.00210으로 좋은 결과를 보여주었다.

[16]에서는 뎅기열 발생률이 가장 높은 20개 도시 중 월간 뎅기열 사례와 2005~2018년의 지역 기상 데이터를 바탕으로 실험이 진행되었고 LSTM 모델을 사용하였다. 그 결과 LSTM 모델을 사용했을 때가 SIR 모델 및 ZIGAM 모델을 사용했을 때 보다 RMSE의 예측 측정치가 평균적으로 각각 54.79%, 34.76% 감소함을 보였다.

[17]에서는 Google 독감 트렌드 데이터와 질병 통제 센터 데이터를 기반으로 일련의 모델이 설정되었다. 모델에는 GFT 회귀 모델, 가중 GFT 회귀 모델, GFT + CDC 회귀 모델, CDC 회귀 모델, 가중 CDC 회귀 모델 등 총 5개의 모델이 사용되었고. 미국 10개 지역에서 인플루엔자 발병을 예측하였다. 결과는 CDC 회귀 모델, 가중 CDC 회귀 모델 이 다른 모델보다 우수한 성능을 보여주었다.

[18] 에서는 인도에서 2020년 1월 30일부터 2020년 5월 10일 사이에 발생한 COVID-19 을 분석하고 SEIR 모델과 회귀 모델을 사용하여 전염병 발생을 예측하였다. 그 결과 3주간의 테스트 데이터를 예측했을 때 SEIR모델은 RMSLE 1.52, 회귀 모델은 RMSLE 1.75 으로 SEIR 모델이 회귀 모델보다 더 좋은 성능을 나타냄을 보여주었다.

3. 전염병 예측 모델

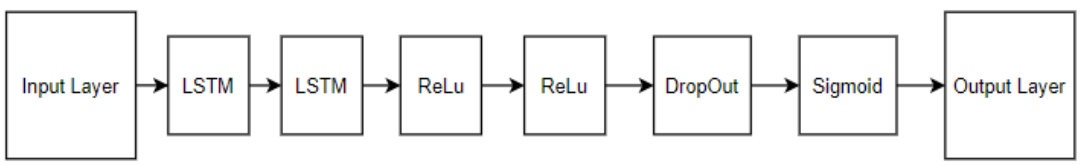
이번 목차에서는 실험에서 사용된 2개의 질병 예측 모델에 대해 설명하며 각 모델에 사용된 하이퍼 파라미터에 대해 설명하고자 한다. 그림 1의 두 개의 모델에는 LSTM(Long Short Term Memory) 레이어가 사용되었다,

여기서 LSTM 이란 RNN(Recurrent Neural Network)의 변형으로, RNN이 학습 데이터에 대하여 관련 정보와 해당 정보를 사용하는 지점의 거리가 멀어질수록 학습 능력이 떨어지는 장기 의존성 문제를 해결하기 위해서 RNN의 Hidden-state에 Cell-state를 추가하여 Input Gate, 와 Forget Gate를 통해 이전의 데이터를 얼마만큼 잊고 새로운 데이터를 얼마만큼 기억하게 할지 결정하게 하여 시계열 데이터가 길어져도 이전의 학습 데이터를 잘 기억할 수 있도록 설계된 신경망이다 [19].

또한 LSTM 은 [2], [7], [15], [16], [19] 등의 연구에 따르면 LSTM 모델과 기존의 모델 (ARIMA, RNN, 선형 회귀 모델, SVR…)등을 비교하여 시계열 데이터를 예측했을 때 LSTM 모델이 나머지 예측 모델보다 정확도가 높다는 점이 입증되었기 때문에, 본 실험에서는 LSTM 모델을 사용하여 실험을 진행하게 되었다.

그림 1의 모델은 2개의 LSTM 레이어와 심층 신경망으로 구성되어 있으며. 입력 레이어에는 NEWS , SNS, 검색 쿼리 시계열 데이터 중에 한가지의 웹 데이터만 들어가며 데이터를 수집한 기간(총 104주) 동안 해당 질병에 대해 언급한 횟수, 검색 쿼리에 대한 데이터가 들어가게 된다.

그림 2는 LSTM 레이어 2개로 이루어진 신경망 3개가 하나로 병합된 구조로 각 신경망의 입력 레이어에는 NEWS, SNS, 검색 쿼리 시계열 데이터가 들어가게 되고 총 104주 동안 해당 질병에 대해 NEWS, SNS에서 언급한 횟수, 네이버에서 해당 질병에 대한 검색 비율(검색 쿼리 데이터)이 데이터가 들어가게 된다. 표 1은 그림 1과 그림 2의 모델에 사용된 하이퍼 파라미터이다. 하이퍼 파라미터는 훈련 횟수, 노드의 개수, 손실함수 등이 들어가며 모델이 학습을 어떻게 할 것인지 결정하는 매개변수이다. 각 모델의 훈련 횟수는 700번으로 하였는데 모델이 한번 학습할 때마다 가중치를 저장하게 하고, 저장된 가중치 중에서 검증 데이터에 대한 손실이 가장 적은 가중치를 선택하기 위해서 적절히 큰 값을 선택하였다. 또한 노드 개수와 학습률은 각각의 값들을 여러 번 바꾸면서 학습을 시도하였고, 그중에서 가장 학습 결과가 좋았던 파라미터 값(노드 개수:16개, 학습률: 0.0005)을 사용하였다. 손실함



(그림 1) LSTM 모델
(Figure 1) LSTM model

(표 1) 모델에 사용된 하이퍼 파라미터
(Table 1) Hyper Parameters Used in Model

모델	LSTM 모델	제안된 모델
훈련 횟수	700번	
노드 개수	16 개	
손실함수	Mean Sequence Error	
옵티마이저	Adam	
드롭아웃	0.0005	
학습률	0.5	
타임스텝	30 Week	

수는 주어진 질병 예측 문제가 회귀 문제이므로 회귀 문제에 보편적으로 사용되는 손실함수인MSE(Mean Sequence Error)를 사용하였으며, [20] 의 실험에 따르면 Adam 이 다른 옵티마이저 기법과 비교했을 때 좋은 성과를 보였으며 비교적 안정적으로 loss-value가 감소함을 보여서, Adam 옵티마이저를 사용하였다. 드롭아웃은 모델의 과적합을 방지하기 위해 신경망의 일부를 무작위로 누락시키는 것으로 자주 사용되는 비율인 0.5로 설정하였다. 타임 스텝은 모델이 과거의 데이터 몇 주를 보고 예측하는지 결정하는 요소로 학습 데이터의 크기를 고려하여 30주로 설정하였다.

4. 실험 방법

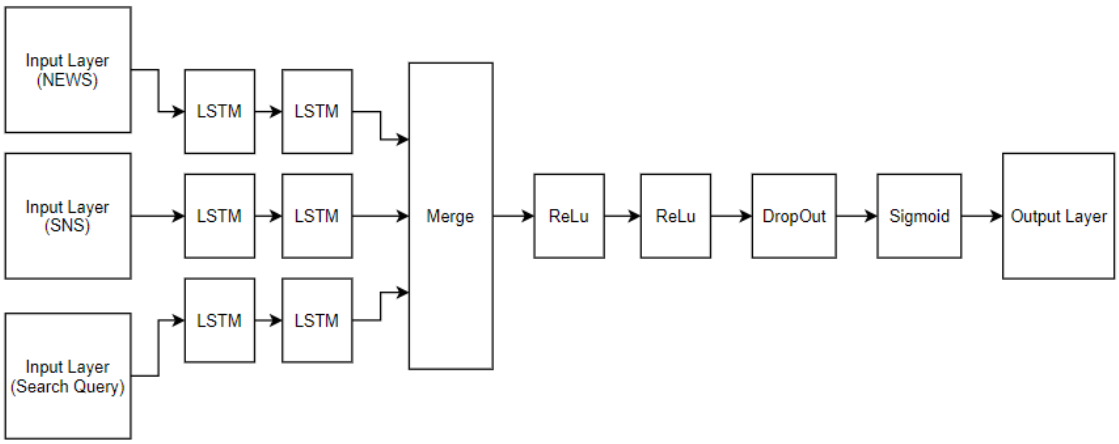
4.1 데이터 수집

우리는 “유행성 이하선염” , “말라리아” 총 2개의 전염병을 대상으로 2018년 12월 31일부터 2019년 12월 28일까지 총 104주간의 전염병의 발생 수, 전염병에 관해 언급된 NEWS, SNS의 수, 검색 비율을 수집하고, 단일 웹 데이터를 사용하는 모델과 우리가 제안하고자 하는 모델을 통해 각 질병을 예측하여 두 모델의 성능을 비교하고자 한다. 실험에 사용된 전염병 발생 데이터는 질병관리본부(KCDC)에서 제공하는 전염병 포털에서 질병 발생 데이터를 주별로 수집하였다.

NEWS 데이터는 네이버 뉴스 API를 통해 선정한 질병 2개의 키워드와 관련된 뉴스 15,064건을 수집하였고, 주별로 질병에 대해 언급된 뉴스의 개수를 세서 CSV 파일로 저장하였다.

(표 2) 수집한 데이터의 양
(Table 2) Amount of data collected

질병이름	발생 횟수	NEWS 개수	SNS 개수	검색 쿼리 평균
유행성 이하선염	35217	5295	155	13.40
말라리아	1136	9769	4637	0.98



(그림 2) 제안된 모델
(Figure 2) Proposed model

SNS 데이터는 트위터 API를 통해 선정한 질병 2개와 관련한 트윗 4,792건을 수집하여 주별로 질병에 대해 언급된 트윗의 개수를 세서 CSV 파일로 저장하였다.

검색 쿼리 데이터는 네이버 데이터 랩에서 제공하는 검색어 트랜드 API를 사용하여 해당 기간 동안 질병과 관련한 검색어 비율을 일별로 수집하고, 주 단위로 나누고 각각의 평균을 구하여 CSV 파일로 저장하였다. (표 2)는 앞에서 언급한 2개의 질병 질병에 대하여 총 104주의 기간 동안 발생한 질병의 발생 횟수, 해당 기간 동안 각 질병이 NEWS, SNS에서 언급된 횟수, 네이버에서 해당 질병이 검색된 검색 비율을 의미한다.

4.2 모델 학습

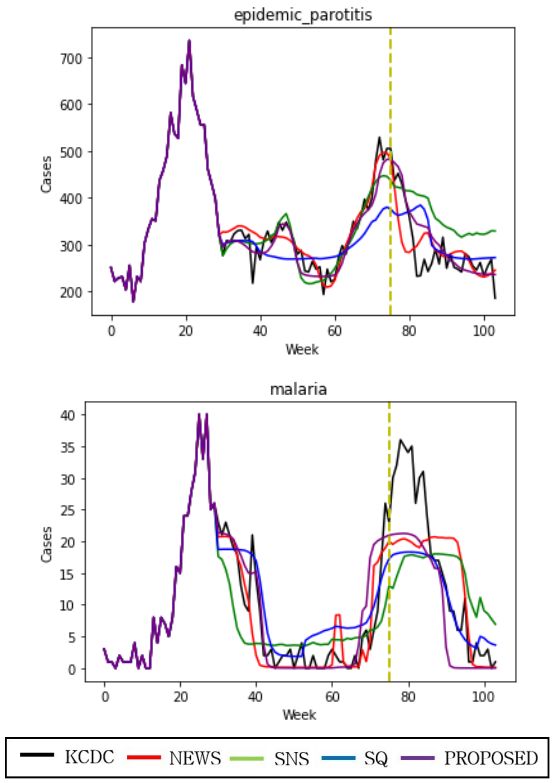
우리는 수집한 데이터 104주 중 75주를 훈련 데이터로 하고 나머지 29주는 모델의 학습 결과를 평가하기 위한 검증 데이터로 나누었다.

학습 데이터를 모델에 훈련시킬 때 표 1에서 설정한 하이퍼 파라미터의 타입스탭 만큼 학습 데이터를 나누고 각 데이터를 0~1 사이의 값으로 정규화(Normalization)시킨 후 모델에 학습 데이터를 넣어 총 700번 훈련시켰다. 그리고 모델이 1번 학습할 때마다 그때의 가중치를 저장하고 다음 학습 때 검증 데이터에 대한 손실을 서로 비교하여 더 좋은 가중치를 선택하도록 하였다. 즉 모델을 700번 학습시키는 동안 가장 검증 데이터에 대한 손실이 적은 가중치를 선택한 것이다. 모델을 학습시키고 난 뒤 우리는 단일 웹 데이터를 사용하는 모델과 우리가 제안하고자 하는 모델을 사용하여 훈련 데이터와 검증 데이터를 예측하였다.

5. 실험 결과 및 분석

그림 3은 앞에서 설명한 2개의 모델을 통해 각각 NEWS, SNS, 검색 쿼리 의 웹 데이터를 사용하여 질병을 발생을 예측한 그래프이다. 그래프에서 황색 점선을 기준으로 왼쪽 영역은 학습 데이터 영역이며 오른쪽 영역은 검증 데이터 영역이다. 학습 데이터 영역은 그림 1, 그림 2의 두 모델이 학습 데이터에 대해서 예측한 결과를 나타내고 검증 데이터 영역은 검증 데이터에 대해서 예측한 결과를 나타낸다.

그림 3에서 검은 선은 질병 발생 데이터를 의미하고 빨간색, 파란색, 초록색 곡선은 각각 그림 1의 모델을 사용하여 NEWS, SNS, 검색 쿼리 웹 데이터를 사용하여 질병을 예측한 결과이다. 보라색 곡선은 우리가 제안하고자



(그림 3) 예측 결과
(Figure 3) Predict Results

하는 모델이 NEWS, SNS, 검색 쿼리 웹 데이터를 모두 사용하여 질병발생을 예측한 결과이다. 그림 3에서 유행성 이하선염에 대한 예측 결과를 보면 학습 데이터 영역에서 파란색 선(SNS)을 제외한 모든 선이 검은색 선(질병 발생 데이터)을 대부분 잘 따라가는 것을 볼 수 있다. 이는 SNS 데이터를 제외하고는 학습 데이터에 대해 훈련이 잘 이루어졌다고 할 수 있다. 검증 데이터 영역에서는 빨간색 선(NEWS)과 보라색 선(Proposed)이 검은색 선(질병 발생 데이터)을 비슷하게 따라가는 모습이 보인다. 이는 모델이 학습을 통해 미래의 데이터를 잘 예측한다고 할 수 있다.

그림 3에서 말라리아에 대한 예측 결과를 보면 학습 데이터 영역에서 모든 선이 검은색 선(질병 발생 데이터)을 비슷하게 따라가는 형태가 보이지만 완벽히 따라가지 않는 모습이 보인다. 이는 학습 데이터에 대해 훈련이 완벽히 이루어지지는 않았지만, 어느 정도의 학습이 이루어졌음을 알 수 있다. 검증 데이터 영역에서는 모든 선이 검

(표 3) 예측 결과 유사도
(Table 3) Predicted Results Similarity

질병명	예측한 데이터	NEWS		SNS		Search Query		Proposed	
		PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE
유행성이하선염	훈련 데이터	0.93	0.04	0.91	0.05	0.82	0.09	0.94	0.04
	검증 데이터	0.72	0.09	0.69	0.16	0.57	0.11	0.86	0.07
말라리아	훈련 데이터	0.84	0.11	0.76	0.14	0.87	0.11	0.87	0.10
	검증 데이터	0.73	0.21	0.60	0.26	0.92	0.20	0.94	0.19

은색 선(질병 발생 데이터)을 비슷하게 따라가기는 하나 대부분 예측 결과가 좋지 않았다. 하지만 모델이 패턴을 어느 정도 학습하여 미래의 데이터를 예측하려고 노력한 모습이 보인다. 표 3은 앞에서 설명한 2개의 모델을 사용해 질병의 발생을 예측했을 때 각 모델이 얼마나 질병 발생을 잘 예측하는지 평가하기 위해서 피어슨 상관계수(Pearson correlation coefficient)와 RMSE(Root Mean Square Error)를 사용하였다.

여기서 피어슨 상관계수란 두 배열 A = {a1, a2, a3,…}, B = {b1, b2, b3,…} 간의 유사도를 수치로 표현한 값으로 -1 에서 1 사이의 값을 가지며. 이 값은 두 배열 A, B의 각 원소 쌍이 같은 방향으로 증가하는 추세를 보이면 1에 가까워지고 반대 방향으로 증가하는 추세를 가지면 -1에 가까워진다. 또한 두 배열이 서로 연관성이 없으면 0에 가까워진다 [21]. 피어슨 상관계수에 대한 수식은 다음과 같다.

$$PCC(A,B)=\frac{\sum_{i=1}^{n}(a_i-\overline{a_i})(b_i-\overline{b_i})}{\sqrt{\sum_{i=1}^{n}(a_i-\overline{a_i})^2}\sqrt{\sum_{i=1}^{n}(b_i-\overline{b_i})^2}} \quad (1)$$

표 3에서 피어슨 상관계수가 1에 가까울수록 학습모델의 예측 결과가 실제 예측 데이터와의 유사도가 높음을 의미한다. RMSE는 평균 제곱근 오차를 말하며 두 배열 A = {a1, a2, a3,…} , B = {b1, b2, b3,…} 이 있을 때 두 배열의 각 원소 쌍 간의 오차를 제곱하여 모두 더한 후 평균을 메긴 수치이며, 수식으로는 다음과 같이 나타낸다 [22].

$$RMSE(A,B)=\sqrt{\frac{\sum_{i=1}^{n}(a_i-b_i)^2}{n}} \quad (2)$$

표 3에서 RMSE의 값이 적을수록 학습 모델의 예측 결과가 실제 예측 데이터와의 오차가 적음을 의미한다.

표 3에서 NEWS, SNS, 검색 쿼리 는 각각 그림 1의 모델을 사용하여 NEWS, SNS, 검색 쿼리 를 통해 질병을 예측한 결과를 의미하며, Proposed는 우리가 제안하고자 하는 모델의 예측 결과를 의미한다.

표 3에서 유행성이하선염의 훈련 데이터를 그림 1의 모델을 사용하여 NEWS, SNS, 검색 쿼리 데이터로 예측했을 때의 피어슨 상관계수가 각각 0.93, 0.91, 0.82이고, RMSE는 각각 0.04, 0.05, 0.09이다. 우리가 제안한 모델을 사용하여 예측했을 때의 피어슨 상관계수와 RMSE는 각각 0.94, 0.04였으며 나머지 3개의 예측 결과와 비교했을 때 피어슨 상관계수가 가장 높았지만, RMSE는 나머지 3개의 예측 결과와 비교했을 때 더 낮은 값을 보이지는 않았다.

표 3에서 유행성이하선염의 검증 데이터를 그림 1의 모델을 사용하여 NEWS, SNS, 검색 쿼리 데이터로 예측했을 때 피어슨 상관계수는 각각 0.72, 0.69, 0.57이고 RMSE는 0.09, 0.16, 0.11이다. 우리가 제안한 모델의 피어슨 상관계수와 RMSE는 각각 0.86, 0.07으로 피어슨 상관계수가 나머지 3개의 예측 결과보다 더 높았으며, RMSE 또한 나머지 3개의 예측 결과 중 가장 적은 값을 가졌다. 이는 우리가 제안한 모델이 단일 웹 데이터를 사용하는 모델보다 더 정확한 예측을 하였음을 의미한다.

표 3에서 말라리아의 훈련 데이터를 그림 1의 모델을 사용하여 NEWS, SNS, 검색 쿼리 데이터로 예측했을 때의 피어슨 상관계수는 각각 0.84, 0.76, 0.87이고 RMSE는 0.11, 0.14, 0.11이다. 우리가 제안한 모델을 사용하여 예측했을 때의 피어슨 상관계수와 RMSE는 각각 0.87, 0.10이고, 피어슨 상관계수가 나머지 3개의 예측 결과보다 더 높지 않았다. 하지만 RMSE는 나머지 3개의 예측 결과 중 가장 작은 값을 가졌다. 그리고 말라리아의 검증 데이터를 예측했을 때의 피어슨 상관계수는 각각 0.73, 0.60, 0.92이고 RMSE는 0.21, 0.26, 0.20이다. 우리가 제안한 모델의 피어슨 상관계수와 RMSE는 각각 0.94, 0.19로 나머지 3개의 예측 결과보다 피어슨 상관계수가 더 높고 RMSE 또한 가장 적은 값을 가졌다.

6. 결 론

우리가 조사했던 기존의 연구 중 대부분은 단일 웹 데이터를 사용한 연구였으며, 단일 데이터는 “COVID-19”처럼 최근에 등장한 질병의 데이터를 수집할 경우 데이터의 양이 한정되어 있기 때문에 많은 양의 데이터를 수집하기 어렵고, 정확한 질병의 예측이 어려웠다.

본 연구에서는 전염병의 발생을 예측할 때 단일 웹 데이터를 사용하여 예측하는 것보다 두 가지 이상의 웹 데이터를 사용하여 예측하는 것이 전염병 예측에 더 효과적임을 실험을 통해 증명하고 전염병 예측에 적절한 모델을 제안하였다. 제안된 방법은 NEWS, SNS, 검색 쿼리 웹 데이터를 모아서 CSV 파일로 저장하고 그림 2의 모델을 사용하여 3개의 웹 데이터를 모두 사용하여 전염병을 예측하였다. 그 결과 표 3의 검증 데이터에 대하여 우리가 제안한 모델의 유사도가 나머지 3개를 통해 예측했을 때 보다 가장 높았고 RMSE도 가장 낮은 값을 보였다.

결과적으로는 기존의 CDC 데이터에만 의존하던 질병의 발생을 웹 데이터를 통해 실시간으로 전염병의 예측이 가능해졌으며, 기존의 단일 데이터만을 사용하는 모델에 비해 정확도가 더 높아졌다. 또한, 최근에 발생한 전염병의 경우는 한정된 데이터의 양 때문에 정확한 예측을 하기 어려웠던 문제를 해결할 수 있다고 생각한다.

하지만 우리가 수집한 웹 데이터 중 NEWS 나 SNS 같은 경우는 광고성 뉴스나, 홍보 등 질병과 관련이 없는 내용이 일부 포함되어 있어서 모델의 예측 정확도가 떨어진 것으로 판단된다. 따라서 웹 데이터에서 문장을 분석하고 질병 발생과 관련이 없는 데이터를 걸러내어 질병 발생 데이터와 웹 데이터의 연관성을 높이면 모델에 의한 학습 정확도가 더 높아질 것으로 기대된다.

참고문헌(Reference)

[1] Y. G. Song, “전염병의 역사는 진행중,” Korean J. Med., Vol. 68, No. 2, pp. 127, 2005.

[2] S. Chae, S. Kwon, and D. Lee, “Predicting infectious disease using deep learning and big data,” Int. J. Environ. Res. Public Health, Vol. 15, No. 8, Aug. 2018. https://doi.org/10.3390/ijerph15081596.

[3] Muthusami R, Bharathi A, and Saritha K, “COVID-19 Outbreak: Tweet based Analysis and Visualization towards the Influence of Coronavirus in the World,” GEDRAG Organ., Vol. 33, No. 02, pp. 534 – 549.

[4] G. Singh Aujla and S. Grover, “Prediction Model for Influenza Epidemic Based on Twitter Data” Int. J. Adv. Res. Comput. Commun. Eng., Vol. 3, No. 7, pp. 7541 – 7545, 2014.

[5] S. Mandal, M. Rath, Y. Wang, and B. G. Patra, “Predicting zika prevention techniques discussed on twitter: An exploratory study,” in CHIIR 2018 - Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, Feb. 2018, Vol. 2018-March, pp. 269 – 272, 2018. https://doi.org/10.1145/3176349.3176874.

[6] Q. Xu, Y. R. Gel, L. L. R. Ramirez, K. Nezafati, Q. Zhang, and K. L. Tsui, “Forecasting influenza in Hong Kong with Google search queries and statistical model fusion,” PLoS One, Vol. 12, No. 5, May 2017. https://doi.org/10.1371/journal.pone.0176690.

[7] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R Niakan Kalhori, “Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study,” JMIR Public Heal. Surveill., Vol. 6, No. 2, p. e18828, Apr. 2020. https://doi.org/10.2196/18828.

[8] F. Liang, P. Guan, W. Wu, and D. Huang, “Forecasting influenza epidemics by integrating internet search queries and traditional surveillance data with the support vector machine regression model in Liaoning, from 2011 to 2015,” PeerJ, Vol. 2018, No. 6, pp. 02-14, 2018. https://doi.org/10.7717/peerj.5134.

[9] Y. Nan and Y. Gao, “A machine learning method to monitor China’s AIDS epidemics with data from Baidu trends,” PLoS One, Vol. 13, No. 7, pp. 01 – 12, Jul. 2018. https://doi.org/10.1371/journal.pone.0199697.

[10] L. Wang, J. Chen, and M. Marathe, “DEFSI: Deep Learning Based Epidemic Forecasting with Synthetic Information,” pp. 9607 – 9612, 2019. www.aaai.org.

[11] X. Zhu et al., “Attention-based recurrent neural network for influenza epidemic prediction,” BMC Bioinformatics, Vol. 20, Nov. 2019.

https://doi.org/10.1186/s12859-019-3131-8.

[12] D. Liu et al., “A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models.,” ArXiv, 2020.

[13] J. Heo, “Epidemiological Prediction using Deep Learning,” Graduate School of UNIST, 2020.

[14] S. Aich and S. Han, “Malaria Epidemic Prediction Model by Using Twitter Data and Precipitation Volume in Nigeria,” J. Korea Multimed. Soc., Vol. 22, No. 5, pp. 588 – 600, 2019. https://doi.org/10.9717/kmms.2019.22.5.588.

[15] J. Zhang and K. Nawata, “A comparative study on predicting influenza outbreaks,” Biosci. Trends, Vol. 11, No. 5, pp. 533 – 541, 2017. https://doi.org/10.5582/bst.2017.01257.

[16] J. Xu, K. Xu, Z. Li, T. Tu, L. Xu, and Q. Liu, “Developing a dengue forecast model using Long Short Term Memory neural networks method” bioRxiv, 2019. https://doi.org/10.1101/760702.

[17] H. Xue, Y. Bai, H. Hu, and H. Liang, “Influenza Activity Surveillance Based on Multiple Regression Model and Artificial Neural Network,” IEEE Access, Vol. 6, pp. 563 – 575, Nov. 2017. https://doi.org/10.1109/ACCESS.2017.2771798.

[18] Gupta, Rajan, et al. ″Machine learning models for government to predict COVID-19 outbreak.″ Digital Government: Research and Practice” Vol.1, No.4, pp.1-6. 2020.

[19] Siami-Namini, Sima, and Akbar Siami Namin. ″Forecasting economics and financial time series: ARIMA vs. LSTM.″ arXiv preprint arXiv” Vol.1803 No.06386 2018.

[20] Kingma, Diederik P., and Jimmy Ba. ″Adam: A method for stochastic optimization.″ arXiv preprint arXiv” Vol.1412 No.6980 2014.

[21] Benesty, Jacob, et al. ″Pearson correlation coefficient.″ Noise reduction in speech processing. Springer, Berlin, Heidelberg, pp.1-4. 2009.

[22] Chai, Tianfeng, and Roland R. Draxler. ″Root mean square error (RMSE) or mean absolute error (MAE)?.″ Geosci. Model Dev, Vol.7, No.3, pp.1247-1250, 2014.

◐ 저 자 소 개 ◑

김 영 하(Yeongha Kim)9pt
2020년 ~ 현재 상명대학교 컴퓨터과학과(석사과정)
관심분야 : 인공지능, 사물 인터넷
E-mail : sky97613@naver.com

김 인 환(Inhwan Kim)
2020년 ~ 현재 상명대학교 컴퓨터과학과(석사과정)
관심분야 : 인공지능
E-mail : moreih29@gmail.com

장 백 철(Beakcheol Jang)
2009년 North Carolina State University 컴퓨터공학과(공학박사)
2020년~ 현재 상명대학교 컴퓨터과학과 교수
관심분야 : 무선 네트워크, 인공지능
E-mail : bjang@smu.ac.kr