# (1) Introduction to Data Mining

## A1. Summarize your research questions with the related data set

## (1) Research Question

웹 및 실시간 소셜 미디어 데이터를 활용한 데이터마이닝으로 공공보건 체계를 향상할 수 있을까?

## (2) Related Paper

Zhang, Yiming et al. “An intelligent early warning system of analyzing Twitter data using machine learning on COVID-19 surveillance in the US.” *Expert systems with applications*
vol. 198 (2022): 116882. doi:10.1016/j.eswa.2022.116882

(https://www.sciencedirect.com/science/article/pii/S0957417422003268)

- Citations in Scopus : 94th percentile (q1)

- FWCI : 3.78

- Journal : Expert System With Applications

- **논문 선정 이유**

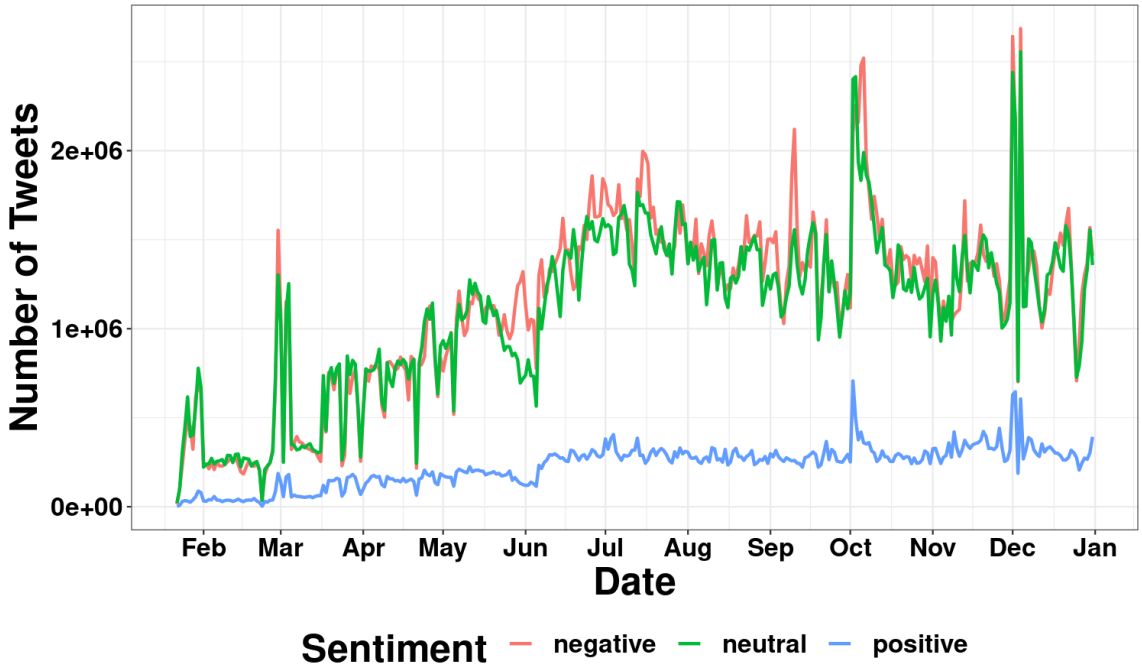
  데이터마이닝, 기계 학습, 딥러닝 등 빅데이터 분석 및 활용 기법이 최근 보건 의료 분야에서도 적용되고 있다. 특히 전 세계적으로 COVID-19가 유행하면서 전염병의 확산을 예측하고 대응하기 위한 연구가 활발하게 논의되었다.  그 중에서 특히 Twitter 데이터를 활용하여 기존 시스템과 비교해 전염병의 확산을 조기에 예측할 수 있다는 본 논문의 아이디어와 내용이 흥미로웠고, 우리 팀의 연구 방향을 설정하는 과정에서 이러한 아이디어가 도움이 될 수 있을 것으로 생각하여 해당 논문을 선정하였다.

- **논문의 배경 및 주제**

  세계보건기구는 COVID-19의 확산을 팬데믹으로 선언했다. 기존의 전염병 감시는 공공보건 당국이 적시에 개입하여 COVID-19가 대유행이 되기 전에 이를 완화하고 통제하

도록 경고하지 못했다. 전통적인 공공보건 감시와 비교할 때 Twitter를 포함한 소셜 미디어의 풍부한 데이터를 활용하는 것은 유용한 도구로 간주되어 왔으며 기존 감시 시스템의 한계를 극복할 수 있다.

본 논문은 Twitter 데이터와 새로운 머신러닝 기법을 활용한 지능형 COVID-19 조기경보 시스템을 제안한다. 자연어 처리(NLP) 모델 중 하나인 BERT를 미세 조정하여 학습시켜 Twitter 데이터를 분류하였다. 또한 Twitter 데이터 기반 선형 회귀 모델을 통해 예측 모델을 구현하여 COVID-19 발생 초기 징후를 감지하고, 이를 기반으로_조기 경보 웹 애플리케이션을 제공했다.




- **논문의 Research Gap**

  [기존 연구]

  웹 및 소셜 미디어 데이터를 사용한 전염병의 발생률을 분석 및 감지한 선행 연구들이 있다.

  웹 데이터의 경우, Google에서 검색 쿼리 데이터를 사용하여 독감 트렌드를 파악하고, 선형 회귀를 통해 주간 인플루엔자 상태를 예측한 사례가 있다. 그들은 웹 검색 데이터와 인플루엔자 활동 사이에 강한 상관관계가 있음을 확인했다. 또한 Google 검색 로그, Twitter 데이터, 병원 방문 기록을 활용하여 스택 선형 회귀, SVM 회귀 등으로 인플루엔자 감시를 수행한 연구 결과가 존재한다.

  소셜 미디어 데이터의 경우, 공공보건 감시에 Twitter 데이터를 사용한 연구가 존재한다. Twitter는 전 세계적으로 5억 명 이상의 사용자를 보유하고 있으며, 사용자는 건강 상태

및 기타 건강 관련 조건을 포함하여 Twitter에 자신의 상태와 생각을 실시간으로 게시할 수 있다. 실제로 Twitter 데이터에 시계열 분석을 적용하여 나이지리아의 에볼라, 그리고 플로리다의 지카 바이러스의 공식 발표보다 선제적인 질병 감시를 수행한 사례가 있다.

[Research Gap]

이 논문에서는 기존 연구와 두 가지 차별점이 있다.

첫째, COVID-19 감시에 사용할 수 있는 지리적 위치가 있는 Twitter 데이터를 사용했다. 이때 트윗 분류 방법으로 미세 조정된 Bert를 사용했다.

둘째, Twitter 기반 선형 회귀 모델을 제안하여 발생에 대한 위험을 조기에 감지한다. 또한 이러한 결과를 시각화하고 사용자가 웹에서 상호 작용할 수 있는 조기 경고 시스템을 구축했다. 이를 통해 건강 부서와 공공보건 공무원에게 신속한 결정을 내릴 수 있는 도움을 제공했다.

- **사용된 방법론**

  [데이터셋]

  Twitter에서 corona virus 키워드로 2020년 1월 ~ 3월, 3달 간 수집된 공개 데이터셋을 사용했다. 1130만 개의 원본 데이터에서 미국의 특정 주 단위로 위치 정보를 특정할 수 있는 110만 개의 데이터를 선별했다. 각 데이터는 문장 부호를 제거한 다음 모든 단어를 원형으로 바꾸어 토큰화 후 사용했다. (복수형, 과거형 등의 단어를 원형으로 사용)

  [텍스트 분류]

  각각의 데이터를 COVID-19와 관련이 있는 데이터 / 관련이 없는 데이터로 분류했다. 누군가가 양성 판정을 받았다는 내용이나 의심 환자가 있다는 내용, 코로나 증상을 설명하는 내용이 포함된 데이터를 1, 포함되지 않는 데이터를 0으로 라벨링했다. 110만 개의 트윗 데이터 중 7064개의 데이터를 수작업으로 라벨링한 다음 텍스트 분류 모델을 학습시켜 나머지 데이터를 분류했다.

  텍스트 분류 모델은 KNN과 SVM, DPCNN, BERT의 4가지 모델을 사용 후 비교하였다.

  [코로나 경향성 예측]

  당일 확진자 수, 일일 트윗 수, 일일 코로나 관련 트윗 수의 세 값을 독립변수로 사용하는 다중회귀분석을 활용했다.

$$y = \sum_{k=1}^{m} \alpha_k Conf_k + \sum_{k=1}^{m} \beta_k Tweet_k + \sum_{k=1}^{m} \gamma_k ClsTweet_k + \delta$$

선형회귀모델의 성능 평가를 위해 R2 상관계수를 사용했다.

- **결론 (도출할 수 있는 의미)**

  텍스트 분류 작업에서는 BERT 모델이 가장 좋은 정확도와 F1 score를 보여주었다.

**Table 4**
Twitter classification results (The best performance is marked in bold font).

| Algorithm | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|
| KNN (Modu et al., 2017) | 0.83 | 0.61 | 0.66 | 0.95 |
| SVM (Modu et al., 2017) | 0.82 | 0.67 | 0.72 | 0.95 |
| DPCNN | 0.20 | 0.75 | 0.32 | 0.94 |
| **Fine-tuning BERT** | **0.96** | **0.99** | **0.98** | **0.99** |

선형 회귀 모델에서는 다음과 같은 정확도를 보여주었다.

**Table 7**
US level prediction model with Twitter data results.

| Prediction Model | Pseudo $R^2$ |
|---|---|
| US Linear Regression Model (predict one day prior) | 0.977 |
| US Linear Regression Model (predict two days prior) | 0.979 |
| US Linear Regression Model (predict three days prior) | 0.977 |
| US Linear Regression Model (predict four days prior) | 0.979 |
| US Linear Regression Model (predict five days prior) | 0.789 |
| US Linear Regression Model (predict six days prior) | 0.909 |
| US Linear Regression Model (predict seven days prior) | 0.621 |

위의 결과에 따라 코로나의 경향성을 약 6일 전에 높은 정확도로 조기 경보할 수 있다는 결론이 도출되었다.

이를 바탕으로 미국에서의 COVID-19를 Twitter 데이터를 통해 분석하고 예측하여 보건 당국에서 사용하는 감시 시스템보다 빠르게 현재 상황에 대한 추정치를 웹 애플리케

이션을 통해 배포하였다. 이를 통해 사람들은 빠르게 현 상황을 파악하고 대처할 수 있다.

- **논문을 개선시킬 수 있는 점 / 아쉬운 점**

  논문에서는 2020년 3월까지 수집된 데이터셋을 이용했는데, 이는 COVID-19가 발발한 지 시간이 얼마 경과되지 않은 시점이다. 이에 최근 데이터셋을 활용함으로써 논문을 개선할 수 있다.

  또한, Twitter 데이터를 사용한 실험만 수행하지만 사전 연구처럼  웹 검색 쿼리 데이터 등의 데이터를 통해 성능을 향상을 기대할 수 있다.

  더불어, 수작업으로 라벨링한 데이터셋(7064개)의 크기가 수집된 데이터셋(110만 개)에 비해 비교적 작다. 향후 연구를 위해서는 더 많은 데이터가 필요하겠다.

# (3) Related Dataset

위 논문에서 사용한 데이터셋은 다음 논문에 제공되어 있다.

- Christian, Lopez E., et al. “Understanding the Perception of COVID-19 Policies by Mining a Multilanguage Twitter Dataset.” *PREPRINT (Version 1) Available at Research Square [Https://Doi.Org/10.21203/Rs.3.Rs-95721/V1]*, Oct. 2020.

| Year | Month | Daily Avg. Original | Daily Avg. Retweets | Daily Avg. Tweets | Total of Orignal | Total of Retweets | Total of Tweets | Total with Geolocation |
|---|---|---|---|---|---|---|---|---|
| 2020 | 1 | 5,947 | 30,576 | 35,501 | 1,958,346 | 7,852,504 | 9,810,850 | 1,773 |
| 2020 | 2 | 10,978 | 29,918 | 40,604 | 7,624,648 | 21,944,443 | 29,568,948 | 8,103 |
| 2020 | 3 | 13,095 | 44,714 | 56,283 | 12,610,824 | 46,659,589 | 59,270,412 | 19,952 |

논문에서는 2020년 3월까지의 데이터만 사용하였지만, 데이터셋을 만든 사람의 github에 2022년 12월까지의 데이터가 추가로 제공되어 있다. 이 최신 데이터를 활용할 수 있겠다.

- https://github.com/lopezbec/COVID19_Tweets_Dataset_2020

논문에서는 COVID-19에 초점을 맞추어 진행했지만, COVID-19 이외에도 매년 반복되는 독감이나 조류 인플루엔자 등의 전염병은 지속적으로 사회 문제가 되고 있다. 해당 연구에서 사용한 방법론을 이러한 전염병에 적용하여 전염 추이를 예측할 수 있다면 공공 의료체계에 도움이 될 것이다. 특히 독감 백신 미접종자를 대상으로 유행을 공식 보고서보다 선제적으로 감지하여 그 전에 미리 백신 접종을 하도록 장려하거나, 의료 시설이 취약한 지역에 시설을 확충할 수 있도록 미리 대비를 할 수 있을 것이다.

- 예를 들어 미국 질병통제예방센터(CDC)에서는 미국 전역의 조류 독감 발병 현황을 일일 단위로 제공하고 있다. (https://www.cdc.gov/flu/avianflu/data-map-commercial.html)
- 한국의 농림축산검역본부에서도 유사한 데이터를 제공하고 있다. (https://www.qia.go.kr/viewwebQiaCom.do?id=58647&type=2_12qlgzls)

---

## A2. Comment on social impact aspect of applications of your interests

데이터마이닝을 활용한다면 공공보건 측면에서 ‘건강한 삶 보장’과 ‘복지 증진’이라는 크게 두 가지의 긍정적인 영향을 끼칠 수 있다.

데이터마이닝을 이용하여 수집한 데이터는 예방 접종 및 감염병 감시 등의 분야에서 중요한 정보를 제공한다. 이를 통해 공공보건 기관은 보다 정확하고 효과적인 의사결정을 내리며, 질병 예방, 감시, 진단 및 치료 등 다양한 분야에서 활용할 수 있다. 예를 들어, 질병 감염 위험성이 높은 지역을 미리 파악하여 적극적인 예방 조치를 취할 수 있다. 특히, 실시간 소셜 미디어 데이터를 활용함으로써 전염병 유행의 특징과 패턴을 파악하여 적시에 대응하는 등의 방식으로 데이터마이닝이 공공보건 분야에서 기여할 수 있다.

복지 면에서는, 데이터 수집 대상을 질병이 아닌 일반 건강 분야로 바꿈으로써 다양한 사회적 긍정적 영향을 이끌어낼 수 있다. 예를 들어, 소셜 미디어 데이터에 포함된 지역 정보를 통해 지역별 건강 현황을 파악할 수 있다. 이로부터 지역 간 의료격차를 파악하여 의료 시설을 개선한다면 의료 접근성 향상에 기여할 수 있다. 특히, 의학적 장애 또는 질병을 소프트웨어 프로그램이나 온라인 서비스를 통해 예방, 관리 및 치료하는 디지털 치료 분야와 연계하여 건강 복지를 제공하는 데에 크게 이바지할 수 있다.

즉, 웹 및 실시간 소셜 미디어 데이터를 활용한 데이터 마이닝은 공공보건 분야에서 미래를 예측하고 위험을 예방함으로써 공공보건 체계를 향상할 수 있다.