Regular paper | Journal of the KIECS. pp. 389-396, vol. 19, no. 2, Apr. 30. 2024, t. 124, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2024.19.2.389

머신러닝을 이용한 과학기술 문헌에서의 지역명 식별과 분류방법에 대한 성능 평가

이정우* · 권오진**

Performance Assessment of Machine Learning and Deep Learning in Regional Name Identification and Classification in Scientific Documents

Jung-Woo Lee* · Oh-Jin Kwon**

요 약

생성형 AI는 최근 모든 분야에서 활용되고 있으며, 심층 데이터 분석 분야에서도 전문가를 대체할 수준으로 발전하고 있다. 그러나 과학기술 문헌에서의 지역명 식별은 학습 데이터의 부족과 이에 따른 인공지능 모델을 적용한 사례가 전무한 실정이다. 본 연구는 Web of Science에서 한국 기관 소속 저자들의 주소 데이터를 활용해 지역명을 분류하기 위한 데이터셋을 구축하고, 머신러닝 및 딥러닝 모델의 적용을 실험 및 평가했다. 실험 결과 BERT 모델이 가장 우수한 성능을 보였으며, 광역 분류에서는 정밀도 98.41%, 재현율 98.2%, F1 점수 98.31%를 기록하였다. 시군구 분류에서는 정밀도 91.79%, 재현율 88.32%, F1 점수 89.54%를 달성하였다. 이 결과는 향후 지역 R&D 현황, 지역 간 연구자 이동성, 지역 공동 연구 등 다양한 연구의 기반 데이터로 활용이 가능하다.

ABSTRACT

Generative AI has recently been utilized across all fields, achieving expert-level advancements in deep data analysis. However, identifying regional names in scientific literature remains a challenge due to insufficient training data and limited AI application. This study developed a standardized dataset for effectively classifying regional names using address data from Korean institution-affiliated authors listed in the Web of Science. It tested and evaluated the applicability of machine learning and deep learning models in real-world problems. The BERT model showed superior performance, with a precision of 98.41%, recall of 98.2%, and F1 score of 98.31% for metropolitan areas, and a precision of 91.79%, recall of 88.32%, and F1 score of 89.54% for city classifications. These findings offer a valuable data foundation for future research on regional R&D status, researcher mobility, collaboration status, and so on.

키워드

AI Dataset, Multi-class Classification, Publication Database, Regional Name Classification, S&T R&D Monitoring
AI 데이터셋, 다중 클래스 분류, 학술논문 데이터베이스, 지역명 분류, 과학기술 R&D 모니터링

* KISTI 글로벌RnD분석센터(jungwoolee@kisti.re.krl)
** 교신저자 : KISTI 글로벌RnD분석센터
· 접 수 일 : 2024. 02. 27
· 수정완료일 : 2024. 03. 20
· 게재확정일 : 2024. 04. 12
· Received : Feb. 27, 2024, Revised : Mar. 20, 2024, Accepted : Feb. 12, 2024
· Corresponding Author : Oh-Jin Kwon
Center for Global RnD Analysis, KISTI
Email : dbajin@kisti.re.kr

Ⅰ. 서 론

과학기술 연구개발 정책의 효율적인 평가와 결정에 대한 요구가 증대되고 있는 최근의 국면은 증거 기반 의사결정의 중요성을 심화시키고 있다[1]. 특히, 전략적으로 중요한 국가전략기술 분야에 대한 정책결정 과정에서는 과학계량 지표를 활용한 글로벌 R&D 지형 분석결과를 더욱 필요로 하는 현실이다.

이러한 분석에는 연구기관(대학 및 기업), 도시, 국가를 조사 단위로 활용하여 지표 기반의 모니터링 및 평가가 이루어지는데[2], 분석 수행 과정에서 직면하는 중요한 문제 중 하나는 개별 연구 성과를 정확하게 지역 또는 기관에 할당하는 것이다. 그러나 학술논문 데이터베이스에서 저자의 소속기관 주소 정보의 오류와 불일치 등에 의해 발생하는 모호성(ambiguation)은 공동 저자 연구, 국제 협력, 과학 분야의 경계 설정, 기관의 가시성 그리고 연구 기관의 순위 산정 등에 중대한 영향을 미칠 수 있다는 문제점이 제기되어왔다[2-5].

학술논문 데이터베이스 내 수록된 주소 정보는 일반적으로 ① 기관명, ② 하위기관명, ③ 도시(우편번호), ④ 국가의 순으로 구성되며[4], 입력 과정에서 저자에 의해 공식 명칭이 아닌 다양한 변형 명칭이나 오탈자가 포함될 수 있다. 또한, Clarivate, Elsevier 등 데이터베이스 제공 업체의 처리 과정에서도 오류가 발생할 가능성도 상존한다.

실제 데이터 상의 지역명과 행정안전부에서 지정한 표준명 간 비교를 통해 데이터 처리 과정의 어려움을 확인할 수 있다. 예를 들어, 표 1의 첫 번째 레코드에 나타난 'TAEGU CITY'는 광역 단위에서는 'Daegu'로, 시군구 단위에서는 'Daegu Nam-gu'로 정정되어야 한다. 대구광역시와 관련해 자주 나타나는 변형 명칭으로는 'Taegu', 'Taegu city', 'Deagu', 'Dae Gu', 'Daeku', 'Dague', 'Deagu Metropolitan City' 등이 있다. 이러한 변형 명칭들의 표준화가 전제되어야 왜곡 없는 분석을 보장할 수 있다.

또한 시군구 단위 표준화가 필요할 경우 기관 정보의 지역적 맥락을 추가로 고려해야 한다. 예를 들어, "Water Supply Bur, Dept Facil, Taegu City"는 대구시 상수도사업본부를 지칭하며 이는 대구 남구에 위치한다. 따라서 해당 시군구는 Daegu Nam-gu로 지정되어야 한다. 이는 텍스트로만은 파악하기 어려운 지역적 맥락이 고려된 결과이다.

표 1. 실제 데이터와 요구되는 분류 라벨
Table 1. Actual Data & Required Label Class

Publication ID(UT)	Affiliation Information (An Author's)	Required Label Class
WOS:A1977ED68400018	Water Supply Bur, Dept Facil, **Taegu City**, South Korea	Daegu / Daegu Nam-gu
WOS:000251227000023	Pohang Univ Sci & Technol, Commun_Res_Lab, Pohang 790784, South Korea	Gyeongsangbuk-do / Pohang-si
WOS:000383357700031	Seoul Natl Univ, Bundang Hosp, Dept Radiol, **Songnam** 13620, South Korea	Gyeonggi-do / Seongnam-si
...	...	...

상술한 바와 같이 학술논문 데이터에서 저자 주소에 존재하는 모호성을 해결하기 위하여 유사도 기반 문자열 매칭 알고리즘, 확률 규칙 기반 모델, 머신러닝 및 딥러닝 모델을 포함한 다양한 접근방식이 연구되었다[6-9].

먼저, 문자열 매칭 알고리즘은 변형 명칭과 오탈자, 그리고 동일한 지역명이 서로 다른 국가나 지역에서 사용되는 경우에 취약하다는 단점이 있다. 또한, 저자가 입력한 텍스트를 기반으로 하므로 주소 정보에 중요한 정보가 누락되는 경우 지역 분류가 어려워진다. 둘째, 조건부 랜덤 필드(Conditional Random Fields)와 같은 확률 규칙 기반 모델과 머신러닝 및 딥러닝 기반 분류 모델은 훈련 데이터의 품질과 양에 크게 의존한다. 이러한 모델을 효과적으로 학습시키기 위해서는 학술논문 데이터에 기반한 분류된 대량의 고품질 학습 데이터가 필수적이다. 이를 구축하는 데에 필요한 상당한 시간과 노력은 모델의 성능 향상을 넘어서는 별도의 연구 주제가 될 수 있다.

이에 본 연구는 과학기술 R&D 모니터링 및 평가 목적으로 학술논문 데이터베이스인 Web of Science로부터 지역명을 효과적으로 분류하기 위한 표준화된 데이터셋을 구축하고, 이를 딥러닝을 통해 평가하는 것을 목표로 한다. 우선, 본 연구에서는 한국 기관을 소속으로 가진 저자들의 주소 텍스트 데이터를 수집하고 정제하여 지역명을 분류하기 위한 표준화된 데

이터셋을 개발한다. Web of Science에 수록된 한국 기관 소속 저자들의 297만 개의 레코드를 기반으로 광역 및 시군구 행정구역을 라벨링한 학습 데이터셋을 구축한다. 이후 데이터셋의 평가를 위하여 다중 클래스 분류에 적합한 머신러닝 및 딥러닝 모델을 사용하여 지역명의 분류 성능을 평가한다.

본 논문은 서론으로 시작하여 2장에서는 관련 연구를 소개한다. 3장에서는 학습 데이터셋 구축 과정 및 결과와 4장에서는 분류 성능 평가를 다룬다. 마지막 5장에서는 결론을 제시한다.

II. 관련 연구

이전부터 텍스트에서 지역의 모호성을 해결하는 다양한 연구가 진행되어왔다. 이러한 연구의 주요 목적은 'Seoul', 'South Korea'와 같은 도시 또는 국가를 지칭하는 모든 지역 정보를 정확히 식별하는 것이다. 학술논문에서 저자 소속 정보의 모호성을 해결하기 위한 연구는 지역 뿐만 아니라 기관 식별까지 초점을 맞춘 경우가 많지만[6] 본 연구에서는 지역 분류에 초점을 맞추어 선행연구를 검토한다.

지역 정보 탐지 및 추출은 자연어처리 분야의 정보추출(information extraction)에 해당한다. 이는 텍스트 자료에서 지역 정보를 추출하는 작업으로 주로 개체명 인식(named entity recognition) 기술이 적용된다. 2000년 전후에는 형태소 분석을 통해 추출한 명사와 행정구역명을 지역정보 시소러스와 매칭하여 정보를 분류하였다[7].

V. I. Torvik(2015)는 저자 소속 문자열을 지리적 위치로 매핑하는 MapAffil을 제안하였다. PubMed 데이터를 기반으로 도시명(city), 변형명, 역사적 명칭, 오탈자 변형명, 지오코드(geocode)를 포함하는 데이터를 만들고 이를 국가와 결합한 n-gram 사전을 구축했다. 저자 소속 정보 문자열을 쉼표로 분리한 후 이 사전 데이터와 비교하여 최종적인 지역 분류를 수행하였다. 300개의 무작위 레코드에 대한 실험한 결과 도시들을 97.7%의 정확도로 잘 분류했지만, 데이터 필드의 누락에 민감하게 작동하는 한계가 있으며 실험 규모가 작아 일반화하기 어렵다[8].

Saravit et al.(2022)은 해외식품제조업소의 주소 데이터를 표준화된 행정구역명으로 변환하는 방법을 제안하였다. 식품의약품안전처의 수입식품정보포털 사이트에서 수집한 데이터셋을 라벨링하여 RNN-LSTM 모델에 학습시켰다. 제안된 모델의 행정구역 수준별 정확도는 중국 지역 기준으로 광역 레벨은 98.79%, 하위 레벨은 92.85%로 나타났다. 그러나 국가별 실험에서 중국, 일본, 캄보디아에 대한 지역 분류 정확도가 각각 90.18%, 89.44%, 77.81%로 나타났다. 이 결과는 학습 데이터셋의 불충분성과 편향성에 기인한 것이다. 데이터 레코드 기준 중국은 24,401개, 일본은 5,968개, 캄보디아는 76개였기 때문이다[8].

김진실 외(2022)는 Saravit et al.(2022)의 방법을 보완하는 것으로, 동명 지역일 경우에는 중복 분류 방식을 적용하였다. 중복 분류란 여러 개의 클래스 중 하나가 아니라 복수 개의 클래스로 분류하는 방식이다. 식품의약품안전처의 수입식품정보포털 사이트에서 수집한 데이터셋을 라벨링하여 에콰도르, 베트남 레코드를 RNN-LSTM 모델 훈련에 활용하였으며, 정확도는 각각 79.7%, 70%로 나타났다[9].

이러한 연구들은 지역 정보의 정확한 분류를 위한 다양한 접근 방식의 가능성을 보여주지만 동시에 저자주소 데이터 필드 구성요소의 일부 누락시 발생하는 낮은 정확도 문제나 학습 데이터셋의 불충분성 및 편향성과 같은 한계를 드러낸다. 특히 과학계량학 적용을 위한 지역명 학습 데이터셋은 전무한 실정이다.

III. 학습 데이터셋 구축

본 연구는 딥러닝 및 머신러닝 모델 학습을 위한 데이터셋 구축을 목표로 한다. 구축할 데이터셋은 저자의 소속과 주소 정보가 포함된 텍스트로, 각 텍스트는 지역명을 정답 라벨로 가진다. 초기 단계에서는 우편번호 매칭과 영문 지역명의 기계 번역을 활용하여 기계적 처리를 수행한다. 이후 이러한 기계 처리 결과를 사람이 직접 검토하여 자동 매칭되지 않는 텍스트에 대해서 수동으로 라벨링하는 하이브리드 접근법을 사용한다.

학술논문 데이터 소스를 이용하여 학습 데이터셋을 구축하는 과정을 진행했다. 이를 위해, 2023년 1주차 기준으로 Web of Science에서 제공하는 총

87,694,858건의 데이터 중에서, 개별 저자의 소속 기관이 한국('South Korea')인 데이터의 서브셋(subset)을 선별했다. 이렇게 추출된 데이터의 총 레코드 수는 2,974,280개이다. 데이터 추출 및 정제 과정은 XML 형식의 원본 데이터를 기반으로 수행되었다.

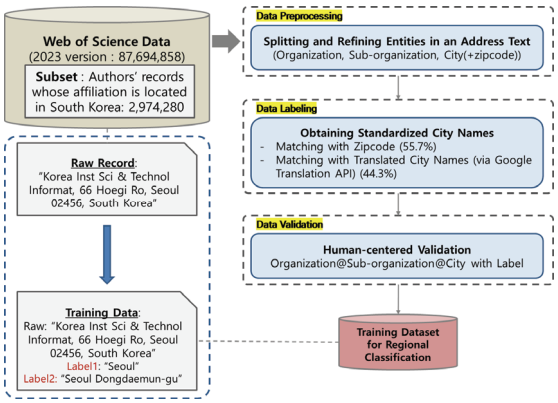

그림 1. 학습 데이터셋 구축 프로세스
Fig. 1 Process of Constructing a Training Dataset

학습 데이터셋 구축 프로세스는 그림 1과 같으며, 크게 세 단계로 데이터 전처리, 라벨링, 그리고 검증 및 라벨링 수정 작업을 수행하였다.

첫 번째 단계에서는 데이터의 표준화를 위해 파이썬의 title() 함수를 사용하여 텍스트의 각 음절 첫 문자를 대문자로 변환하였다. 이후 쉼표 구분자를 사용하여 텍스트를 분할하고 기관(organization), 하위기관(sub-organization), 도시 및 우편번호(city and zipcode), 국가(country)로 구성된 개체를 식별하였다. 도시 및 우편번호 정보는 정규 표현식을 통해 도시명(only city)과 우편번호(zipcode)로 세분화하였다.

두 번째 단계에서는 라벨링 작업을 진행하였다. 우선 Web of Science 데이터에서 식별한 우편번호를 우체국 웹사이트의 우편번호 데이터와 매칭하여 약 55.7%에 해당하는 1,659,535개의 레코드에 표준화된 도시명을 라벨로 할당하였다. 또한, 전처리 과정에서 식별된 영문 도시명(only city)을 Google 기계 번역 API를 사용하여 한국어로 번역하고, 이를 임시 라벨로 지정하였다.

세 번째 단계에서는 사람에 의한 검증 및 라벨링 수정을 진행하였다. 전체 레코드 중 고유 집합을 생성하기 위해 기관, 하위기관, 도시 정보를 결합하여 총 388,627개의 고유 집합을 얻었다. 이를 바탕으로 정답 사전인 시소러스(thesaurus)를 구축하고, 맥락이 정확하다고 판단되는 레코드에 표준화된 도시명을 라벨로 할당하였다. 이 과정에서 부정확한 맥락, 상이한 기관 정보, 한국이 아닌 도시 정보를 포함한 레코드를 제거하였고, 전체 115년 기간동안 지역별 논문 발간 건수가 20개 이하인 시군구는 제외하였다.

마지막 단계에서는 작업자들이 우편번호 매칭, 기계 번역에 의한 임시 라벨, 그리고 시소러스를 사용한 라벨링 결과를 직접 검증하였다. 이 과정을 통해 최종적으로 278,489개의 결합된 레코드를 얻었으며, 이는 초기 원본 레코드 2,795,832개의 약 91.9%를 포함하고 있다. 이렇게 검증된 레코드들은 최종 학습용 데이터셋으로 활용되었다.

IV. 학습 데이터셋 기반 분류 성능 실험

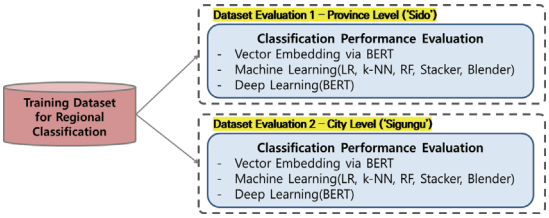

그림 2. 학습 데이터셋 기반 분류 성능 실험 개요
Fig. 2 Overview of Classification Performance Experiments based on Training Dataset

4.1 실험 구성

4.1.1 데이터 구성

학습 데이터셋을 기반으로 한 분류 성능 실험 절차는 그림 2와 같다. 실험은 광역 단위와 시군구 단위로 구분하여 진행되며, 각 실험에서는 주소 텍스트를 BERT 모델[10-12]을 사용하여 벡터로 임베딩한 후, 이 임베딩 벡터와 정답 라벨을 함께 사용하여 학습을 진행함으로써 분류 성능을 평가한다.

클래스 간의 균형을 고려하여 실험용 데이터셋을 구축했다. 총 2,795,832개의 레코드를 기관-하위기관-도시 형태로 재구성한 뒤 20개 이상의 논문 발간 건

수를 가진 210개 시군구를 분류 대상 클래스로 선정했다. 이 시군구들은 상위에 해당하는 광역 단위로도 분류되었다. 최종적으로 실험 데이터셋은 278,489개의 레코드로 구성되었으며, 이 중 80%는 트레이닝 셋으로, 나머지 20%는 테스트 셋으로 무작위 샘플링하여 사용하였다.

트레이닝 셋은 bert-base-nli-mean-tokens 모델을 활용하여 각 텍스트를 768차원의 임베딩 벡터로 변환했다. 텍스트를 임베딩하는 접근법은 고차원의 데이터를 저차원으로 표현함으로써 계산의 효율성을 증가시키고, 모델이 텍스트 간의 의미적 관계를 더 잘 이해할 수 있게 하는 목적을 가진다.

4.1.2 실험 모델 설정

모델의 분류 성능을 검증하기 위해 비교 실험을 하였다. 실험에서는 Python의 AutoML 라이브러리 중 하나인 pycaret과 HuggingFace의 BERT 모델을 활용하여 다양한 머신러닝 모델들을 적용하였다.

AutoML은 특징 공학과 모델 하이퍼 파라미터 설정의 자동화를 통해 머신러닝 모델 학습 과정을 간소화하며, 다양한 모델의 성능을 쉽게 비교할 수 있게 해준다. AutoML 라이브러리인 pycaret을 통해 성능이 우수한 상위 세 모델을 선정하고, 이들을 기반으로 앙상블 기법인 스태킹과 블렌딩을 적용하여 정확도와 일반화 성능을 향상시켰다. 스태킹은 여러 기본 모델로부터 k-fold 교차 검증을 통해 얻은 예측값을 메타 모델의 학습 데이터로 사용하는 기법이며, 블렌딩은 비교적 단순한 방식으로 모델별로 한 번씩 학습한 후 예측된 값으로 메타 모델을 학습하는 방법이다[13].

실험에 사용된 모델은 BERT-base-uncased 버전으로, 총 12개의 Transformer 블록, 768개의 은닉 유닛, 12개의 어텐션 헤드를 포함한다. 'Uncased' 버전은 모든 텍스트를 소문자로 처리하여 대소문자 구분이 결과에 미치는 영향을 줄인다. 분류 작업에는 BERTForSequenceClassification 클래스를 사용했다[10]. 과적합을 방지하기 위해 에포크 수를 10으로 제한하고, 검증 손실이 최소화되는 지점에서 최적의 에포크 값을 찾아냈다. 이에 따라 광역 분류는 4번째 에포크, 시군구 분류는 5번째 에포크에서 최적의 성능을 나타냈다. 모델 훈련 시에는 배치 크기를 32, 학습률을 5e-5로 설정했으며, PyTorch와 Cuda Framework를 사용하여 계산 효율성을 높였다[14].

4.1.3 성능 평가 척도

모델 간의 분류 성능 비교를 위해 정밀도(precision), 재현율(recall), 그리고 F1 점수(F1 score)를 평가 척도로 사용한다. 클래스의 불균형이 있을 경우에 이러한 척도들을 사용함으로써 모델 성능을 보다 정밀하게 평가할 수 있다[15]. 식 (1)의 정밀도는 예측된 샘플 대비 실제 양성 샘플의 비율을 나타낸다. 식 (2)의 재현율은 실제 양성 샘플 중 모델이 양성으로 정확하게 예측한 샘플의 비율을 의미한다. 식 (3)의 F1 점수는 정밀도와 재현율의 조화 평균으로, 각 분류 클래스의 성능을 종합적으로 나타낸다. 이 척도들은 각 클래스별로 계산되며, 모든 클래스에 대한 평균값을 나타내는 매크로(macro) 척도로 최종적으로 제시된다. 식 (1), (2), (3)에서 TP는 True Positive(양성으로 예측된 실제 양성 샘플), FP는 False Positive(양성으로 예측된 실제 음성 샘플), FN은 False Negative(음성으로 예측된 실제 양성 샘플)을 의미한다.

$$\mathrm{Precision}_i = \frac{TP_i}{TP_i + FP_i} \quad \cdots (1)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad \cdots (2)$$

$$F1_i = 2 \times \frac{\mathrm{Precision}_i \times Recall_i}{\mathrm{Precision}_i + Recall_i} \quad \cdots (3)$$

4.2 실험 결과

4.2.1 광역 분류 성능 평가

먼저, 광역 단위 분류 성능을 평가하였으며, 모델별 성능 결과를 표 2에 요약하였다. 머신러닝 모델을 사용한 광역 분류의 성능은 시군구 분류에 비해 전반적으로 우수한 결과를 보였다. 특히, 로지스틱 회귀 모델이 F1 점수 0.9295로 가장 높은 성능을 나타냈으며, 이어 K-Neighbors 분류기와 Random Forest 분류기 순으로 좋은 성능을 보였다. 선정된 세 모델을 기반으로 한 앙상블 모델에서는 블렌딩 기법이 F1 점수 0.9433으로, 스태킹 기법이 F1 점수 0.9477로 나타나 두 기법 모두 개별 모델의 성능을 개선하였다.

BERT 모델을 활용한 광역 분류는 머신러닝 모델

들의 성능을 뛰어넘었으며, F1 점수가 0.9831로 매우 높게 측정되었다.

4.2.2 시군구 분류 성능 평가

다음으로 시군구 단위의 분류 성능을 평가하였고, 각 모델별 성능 결과를 표 3에 정리하였다. 정밀도, 재현율, F1 점수를 포함한 성능 평가 지표를 바탕으로, 상위 세 가지 머신러닝 모델 중 로지스틱 회귀 모델이 F1 점수 0.8607로 가장 높은 성능을 보였다. 그 다음으로 성능이 뛰어난 모델은 K-Neighbors 분류기와 Random Forest 분류기였다. 이 세 모델을 기반으로 한 앙상블 모델에서 블렌딩 분류기는 F1 점수 0.8686으로 기본 모델들보다 더 높은 성능을, 스태킹 분류기는 F1 점수 0.8738로 더욱 나은 성능을 보였다.

광역 분류에 비해 전반적으로 낮은 성능을 보인 상황에서, BERT 모델을 사용한 시군구 분류는 머신러닝 모델들을 초월하는 성능을 보여주었으며, F1 점수는 0.8954로 기록되었다.

표 2. 광역 지역에 대한 분류 성능 결과
Table 2. Classification Performance on the Province Level

Classifiers	Precision	Recall	F1-score
Logistic Regression	0.9295	0.9297	0.9295
k-Neighbors	0.8796	0.8803	0.8793
Random Forest	0.8637	0.8314	0.8302
Blender	0.9434	0.9435	0.9433
Stacker	0.9477	0.9478	0.9477
BERT	0.9841	0.9820	0.9831

* Both the Blender model and the Stacker model combines Logistic Regression, k-Neighbors, and Random Forest models.
** F1-score refers to the Macro F1-score.

4.2.3 행정지역 단위간 분류 성능 비교

표 2와 3을 통해 광역 단위와 시군구 단위의 분류 성능에 차이가 있음을 확인할 수 있다. 이 차이는 주로 광역 지역(17개)과 시군구 지역(210개) 간에 존재하는 분류 클래스 수의 차이에서 기인한다. 분류 클래스의 수가 증가할수록 문제가 복잡해지며, 이는 모델 학습의 어려움과 일반화 성능의 저하로 이어질 수 있다. 특히, 클래스 간 불균형은 이러한 차이를 더욱 심화시킬 수 있다. 광역 지역에서는 클래스 간 구별이 비교적 명확하여 모델이 패턴을 쉽게 학습할 수 있다. 그렇지만 시군구 지역에서는 클래스 간 불균형이 더 심각할 수 있으며 일부 지역은 매우 적은 수의 샘플로 인해 분류가 어려울 수 있다.

표 3. 시군구 지역에 대한 분류 성능 결과
Table 3. Classification Performance on the City Level

Classifiers	Precision	Recall	F1-score
Logistic Regression	0.8619	0.8628	0.8607
k-Neighbors	0.7503	0.7528	0.7473
Random Forest	0.7485	0.7205	0.7080
Blender	0.8706	0.8715	0.8686
Stacker	0.8743	0.8751	0.8738
BERT	0.9179	0.8832	0.8954

* Both the Blender model and the Stacker model combines Logistic Regression, k-Neighbors, and Random Forest models.
** F1-score refers to the Macro F1-score.

분류 클래스의 수가 많고 문제가 복잡한 시군구 분류 작업에서도 BERT와 같은 딥러닝 모델을 사용하여 높은 수준의 성능을 달성할 수 있는지 평가하였다. 표 4에서 볼 수 있듯이 시군구 분류에서의 가중 F1 점수(Weighted F1)가 0.9408로 상당히 높게 나타났다.

보다 확연히 구별되는 결과를 그림 3을 통해 알 수 있다. 가중 F1 점수는 식 (4)와 같이 각 클래스의 샘플 빈도를 고려하여 계산된다. 높은 가중 F1 점수는 소규모 클래스를 포함한 다양한 클래스에서 균형 잡힌 성능을 달성했음을 의미한다.

$$F1_{weighted} = \sum_{i=1}^{N} w_i \times F1_i \quad \cdots (4)$$

$$where,\ w_i = \frac{Number\ of\ samples\ in\ class\ i}{Total\ number\ of\ samples}$$

표 4. 지역 단위별 데이터 요약 및 분류 성능 결과
Table 4. Data Summary and Classification Performance Results by Regional Levels

	Province Level	City Level
No. Classes	17	210
Mean (support)	3276.35	265.22
Std. (support)	3488.93	574.05
Max (support)	11,833	6,554
Min (support)	471	4
Macro F1 (BERT)	0.9831	0.8954
Weighted F1 (BERT)	0.9841	0.9408

* Support refers to the number of actual samples for each class.

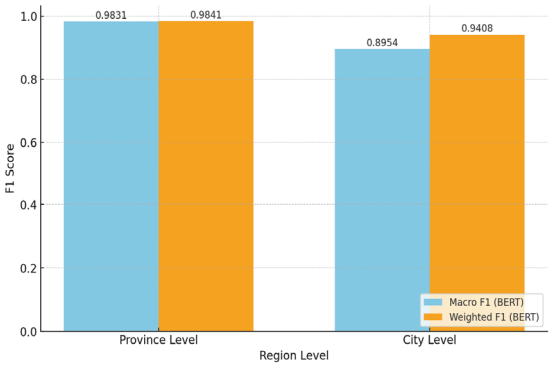

그림 3. 지역 수준별 BERT 모델 분류 성능 비교
Fig. 3 BERT Model Performance by Regional Level

V. 결 론

5.1 연구의 시사점

생성형 AI는 모든 분야에서 활용되고 있으며, 전문가를 대체할 수준으로 데이터 분석 영역에서의 활용도 많이 이루어지고 있다. 그러나 기존 과학기술 문헌에서 지역명을 식별하는 것은 여전히 학습 데이터가 부족하고 생성형 AI의 적용이 미진한 영역이다.

본 연구는 과학기술 R&D 모니터링 및 평가를 위한 지역명 분류의 중요성에 기반하여, 학술논문 데이터의 자연어 문자열을 표준화된 지역명으로 정확하게 분류하는 것을 목표로 한다. 이를 위해, 본 연구는 학술논문 대규모 데이터(Web of Science)에서 한국에 해당하는 서브셋을 추출하고 정제하는 과정을 거쳐 주소 데이터를 기반으로 표준화된 지역명이 라벨링된 학습 데이터셋을 구축하였다. 이 과정은 기계적 방식의 데이터 처리 및 사람에 의한 검증 및 라벨링 수정 작업을 포함하며, 이러한 하이브리드 과정은 데이터의 정확도와 신뢰성을 높이는 데 중점을 두었다.

머신러닝 모델과 비교한 분류 성능 실험 결과, 구축된 데이터셋에서 딥러닝 모델, 특히 BERT가 광역 및 시군구 단위 분류에서 뛰어난 성능을 나타냈다. 시군구 단위에서의 분류는 클래스 수가 많고 복잡성이 높음에도 불구하고, BERT 모델은 가중 F1 점수 기준으로 탁월한 성능을 보였주었다. 따라서, 이 연구 결과는 본 데이터셋에 대한 딥러닝 모델의 활용이 보다 우수한 성과를 낼 수 있음을 시사한다.

결론적으로, 본 연구는 과학기술 R&D 모니터링 및 평가를 위한 정확하고 신뢰할 수 있는 데이터셋의 구축이 가능함을 입증하였다. 고품질 데이터셋의 구축과 우수한 분류 성능 결과는 과학기술 R&D 데이터 분석의 정확도와 신뢰성 향상에 기여할 수 있을 것이다.

5.2 연구의 한계점

본 연구에서는 구축된 데이터셋에 클래스간 편향이 존재함을 확인했다. 이는 현실적으로 특정 지역에서 집중적 연구 활동이 이루어지는 현상이 반영된 결과이다. 현실이 그러할지라도 이러한 편향성을 제거함으로써 강건한 모델 학습을 지원하기 위한 데이터 보완 작업이 요구된다. 인스턴스 수가 적은 클래스에 대한 데이터 증강 기법의 적용 등이 고려될 수 있다.

감사의 글

이 논문은 2024년도 한국과학기술정보연구원(KISTI)의 기본사업으로 수행된 연구입니다.(과제번호: K-24-L03-C01-S01)

References

[1] H. A. Teich, “In Search of Evidence-based Science Policy: From the Endless Frontier to

SciSIP," *Annals of Science and Technology Policy*, vol. 2, no. 2, 2018, pp. 75-199.
[2] W. Glänzel, H. F. Moed, U. Schmoch, and M. Thelwall, *Springer Handbook of Science and Technology Indicators(1st Edition)*. Cham: Springer, 2019.
[3] L. Leydesdorff, "Problems with the 'Measurement' of National Scientific Performance," *Science and Public Policy*, vol. 15, no. 3, June 1988, pp. 149-152.
[4] R. E. De Bruin and H. F. Moed, "Delimitation of Scientific Subfields using Cognitive Words from Corporate Addresses in Scientific Publications," *Scientometrics*, vol. 26, 1993, pp. 65-80.
[5] Z. Taşkın and U. Al, "Standardization Problem of Author Affiliations in Citation Indexes," *Scientometrics*, vol. 98, 2014, pp. 347-368.
[6] J. Kim, S. Hong, and G. R. Thoma, "Labeling Author Affiliations in Biomedical Articles Using Markov Model Classifier," In *Int'l Conf. Data Mining*, Las Vegas, USA, July 2017, pp. 105-110.
[7] K. Min, J. Song, K. Yu, and J. Kim, "A Method for Detecting Location Information using Attention-based Deep Learning Model and Word Embedding," *Journal of Korean Society for Geospatial Information Science*, vol. 27, no. 5, 2019, pp. 33-39.
[8] S. Saravit, J. Bae, K. Lee, and W. Cho, "Global Address Data Quality Verification and Improvement Techniques using Deep Learning," *Journal of Korean Institute of Information Technology*, vol. 20, no. 12, 2022, pp. 15-24.
[9] J. Kim, K. Lee, and W. Cho, "Overseas Address Data Quality Verification Technique using Artificial Intelligence Reflecting the Characteristics of Administrative System," *The Korea Journal of BigData*, vol. 7, no. 2, 2022, pp. 1-9.
[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Arxiv*, arXiv:1810.04805, 2018, pp. 1-16.
[11] J. Lee, "Comparison of Sentiment Classification Performance of for RNN and Transformer-Based Models on Korean Reviews," *Journal of the Korean Institute of Electronic Communication Sciences*, vol. 18, no. 4, 2023, pp. 693-700.
[12] D. Kim, S. Lee, and J. Bong, "Artificial Intelligence for Assistance of Facial Expression Practice Using Emotion Classification," *Journal of the Korean Institute of Electronic Communication Sciences*, vol. 17, no. 6, 2022, pp. 1137-1144.
[13] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras and Tensorflow(2nd Edition)*. CA: O'Reilly, 2019.
[14] Y. Lee and P. Moon, "A Comparison and Analysis of Deep Learning Framework," *Journal of the Korean Institute of Electronic Communication Sciences*, vol. 12, no. 1, 2017, pp. 115-122.
[15] M. Seo, G. Ahn, and H. Sun, "Feature Selection Method from Multiclass Text with Class Imbalance Problem," *Journal of Korean Institute of Industrial Engineers*, vol. 45, no. 2, 2019, pp. 93-100.

저자 소개

이정우(Jung-Woo Lee)

2015년 숭실대학교 소프트웨어특성화대학원 졸업(공학석사)
2019년 한양대학교 대학원 경영컨설팅학과 졸업(박사)
2019년~2021년 한국국토정보공사 공간정보연구원 선임연구원
2021년~2022년 NH농협금융지주 NH금융연구소 부연구위원
2022년~현재 한국과학기술정보연구원 선임연구원
※ 관심분야 : 데이터 과학, 과학계량학, 인공지능

권오진(Oh-Jin Kwon)

2009년 서울시립대학교 대학원 컴퓨터학과 졸업(공학박사)
2001년~현재 한국과학기술정보연구원 책임연구원
※ 관심분야 : 과학계량학, 정보분석 시스템, 기술지능, 정보 구조화