



AI통계분류의 확장 가능성: 산업·직업분류를 넘어서

이규호

국가데이터연구원
데이터과학연구팀

- **(연구 목적)** 본 연구는 건설경기동향조사의 공종(19종), 발주자(27종) 분류코딩 작업의 자동화를 위해 **ELECTRA 기반 한국어 사전학습 모델(KcELECTRA)**을 활용한 인공지능 분류 모델을 시험 구축하고 실무 활용성을 사전 검증하였다.
- **(핵심 결과)** 경인지방데이터청의 10년간 수주자료 약 9만 건을 학습데이터로 활용하고, 2025년 실제 수주 데이터 3,254건으로 모델의 분류 정확도를 검증하였다. 그 결과, 일반적인 분류(Flat) 모델에서는 공종 88%, 발주자 43%의 정확도를 보였으나, 계층형 다중 작업 학습(Hierarchical Multi-Task Learning) 방법론을 적용하여 공종 91%, 발주자 94%로 정확도를 대폭 향상시켰다.
- **(향후 활용)** 본 연구 결과는 지방청 단위의 분류코딩 업무 효율화에 활용될 수 있는 가능성을 보여주었으며, 나아가 **한국표준산업분류·직업분류 외 기타 분류체계**를 가진 통계조사에 대한 **AI통계분류 확산**에 기여할 수 있는 가능성을 보여주었다.

연구 배경 및 목적

- **(연구 배경)** 국가데이터처는 2022년 인공지능 통계분류 자동화 시스템(AI통계분류시스템)을 구축하여 인구총조사, 지역별고용조사, 전국사업체조사, 건설업조사 등 대규모 조사에 활용하고 있고, 산업·직업 분류 외 자체 분류체계를 사용하는 조사에도 실무적용을 확대하고 있음

[그림1] 인공지능기반 통계분류 자동화 시스템



연구 배경 및 목적 (계속)

- (시스템 배경) AI통계분류시스템은 ELECTRA, Roberta 등 사전학습언어모델 기반 지도학습 방법론을 통해 사업체명, 사업체가 하는일, 근무 부서, 직무, 내가 하는일 등 조사표 텍스트로부터 한국표준산업분류·직업분류 코드를 자동 부여하는 체계로, 나라통계시스템과 연계되어 연간 1천만 건 이상의 분류 코딩 업무를 지원함
- (문제 인식) 건설경기동향조사는 종합건설업체의 국내 건설공사 수주액 및 기성액을 공종별·발주자별로 파악하여 건설활동의 단기동향을 신속하게 제공하는 지정통계(승인번호 제101016호)임. 매월 청별로 약 1,500건의 수주 데이터가 입수되며, 공사·발주자명 등 텍스트 자료를 담당자가 수동으로 검토하여 분류코드를 부여하는 작업이 반복적으로 수행되어 담당자 업무 부담과 분류 일관성 확보의 어려움이 상존함
- (연구의 필요성) 2021년 경인지방데이터청에서 KoBERT 모델을 사용하여 자동 분류를 실무에 적용하기 위한 시도를 하였으나, 실제 운용 정확도가 공종 30.6%, 발주자 20.9%에 그쳐 실무 적용에 한계가 있음

[그림2] 경인지방데이터청 건설 공종 발주자 부호 내검 프로그램

3 프로그램 세부 내용

1. 수주 제목과 발주자명 입력하면 AI가 공종부호, 발주자부호를 반환

건설 공종 발주자 부호 내검 프로그램

수주 제목 입력 [101016호 지정통계 제101016호 지정통계 제101016호 지정통계]

(종역)방우-금곡 특선전설 제1공구 노반건설공사 한국철도시설공단

공종부호 : 205 (철도-궤도)

발주자부호 : 1090 (기타공공단체)

공종부호 예측		발주자부호 예측	
205 (철도-궤도)	99.94%	1090 (기타공공단체)	99.99%
109 (기타공공)	0.02%	9999 (기타)	0.06%
203 (도로-교량)	0.01%	1029 (지방자치단체)	0.91%

AI 내검을 위한 수주 자료 엑셀 파일 대량 업로드

파일 업로드 AI 내검

- (연구 목표) 본 연구는 ELECTRA 기반 사전학습 언어모델을 활용하여 건설경기동향조사의 공종(19종) 및 발주자(27종) 분류코딩을 자동화하는 AI 모델을 시험 구축하고, 실무 활용 가능성을 검증하는 것을 목적으로 진행되었음

데이터 및 방법

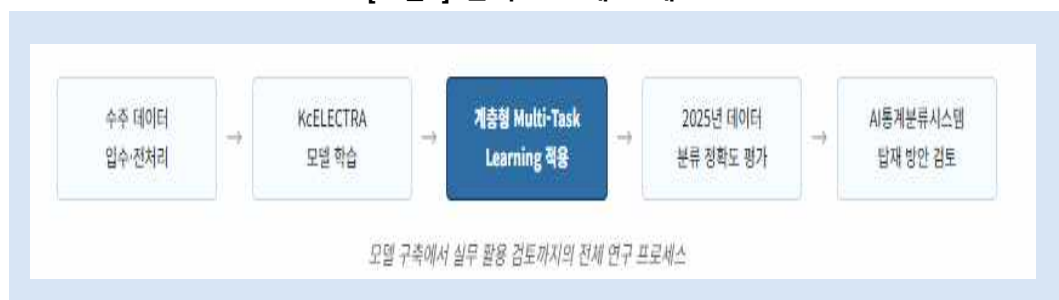
- **(활용 데이터)** 경인지방데이터청에서 입수한 2015년 1월부터 2025년 8월까지의 건설경기동향조사 수주자료 총 90,994건을 활용하였음. 주요 피쳐(Feature)로는 공사명, 기업체명, 공사지역, 발주자명 등의 텍스트 정보를 사용하였고, 정답 라벨(Label)은 공종 부호(3자리, 19종)와 발주자 부호(4자리, 27종)임

<표1> 데이터 구성

구 분	기 간	건 수	용 도
학습셋(Training)	2015.1. ~ 2024.12.	80,155건	모델 학습
검증셋(Validation)	2015.1. ~ 2024.12.	8,907건	학습 중 성능 확인
평가셋(Test)	2025.1. ~ 2025.8.	3,254건	최종 성능 평가
합 계	-	92,316건	-

- **(AI 모델 선정)** 사전학습 언어모델인 KcELECTRA-base(한국어 ELECTRA)를 기반으로 지도학습 방식의 텍스트 분류 모델을 구축하였음. ELECTRA는 구글이 개발한 Transformer 기반 모델로, 기존 BERT 대비 약 1/4 수준의 컴퓨팅 리소스로도 유사한 성능을 달성할 수 있다는 장점이 있음
- **(분석 방법)** 모델 평가에는 정분류율(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score(macro)를 활용하여 분류 단위별 성능을 종합적으로 분석하였음. 과적합 방지를 위해 조기 종료(Early Stopping)와 Checkpoint 기법을 적용하였음

[그림3] 분석 프로세스 개요

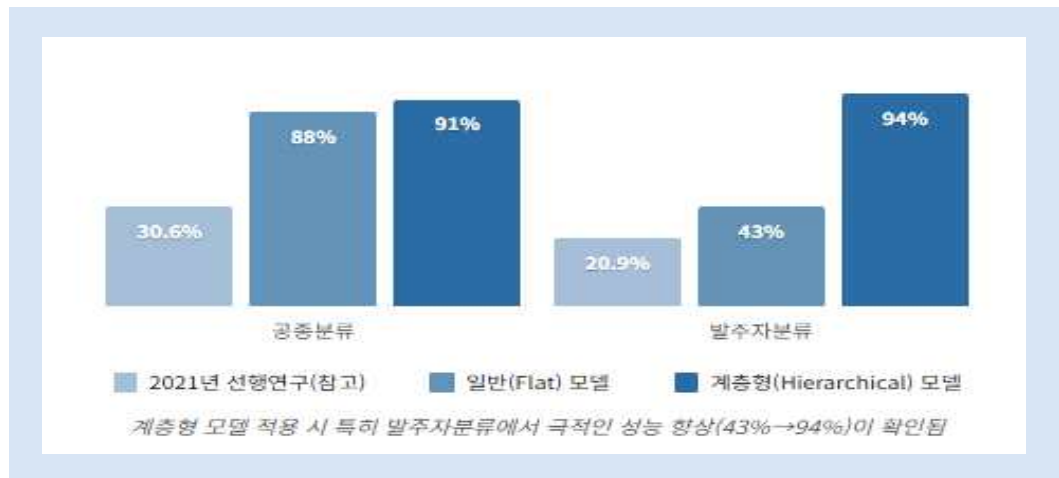


주요 결과

- **(시행 착오)** 먼저 일반(Flat) 구조의 ELECTRA 모델로 시험 구축한 결과, 공종분류는 정분류율 88%(F1-macro 0.86)로 양호한 성능을 보였으나, 발주자 분류는 검증셋에서 93%를 기록했음에도 2025년 실제 평가 데이터에서는 **43% (F1-macro 0.25)로 급격히 저하**되었음. 이는 발주자 분류 체계 27종 중 **민자유치사업 11개 코드의 학습 데이터가 2% 미만**에 불과한 극심한 불균형에 기인했음. 이를 보완하기 위해 공사면 텍스트를 피쳐에 추가하거나(정확도 43%, Precision 감소), 3자리 분류와 키워드 조건 처리를 결합하는 방식(정확도 93%이나 키워드 의존성 과다)을 검토하였으나, 근본적 개선에는 이르지 못하였음

주요 결과 (계속)

[그림4] 모델별 정분류율(%) 비교 - 2025년 평가 데이터(3,254건) 기준



- **(최종 모델)** 이에 분류 코드의 일부 계층적 구조(예:2->21->215->2151)를 반영한 **계층형 다중 작업 학습(Hierarchical Multi-Task Learning)** 방법론을 적용하였음. 하나의 ELECTRA 모델에 분류 단위별 다중 분류기(Head)를 배치하여 계층별 코드를 동시에 학습하도록 설계한 결과, 발주자분류는 정분류율이 **43%에서 94%로** 대폭 향상되었고, 공중분류 또한 **88%에서 91%(F1-macro 0.91)**로 추가 향상되었음

<표2> 최종 모델(계층형) 성능 요약

구 분	학습데이터	평가데이터	정분류율	F1-macro	Precision _{macro}	Recal _{macro}
공중분류(19종)	80,155	3,254	91%	0.91	0.91	0.90
발주자분류(27종)	80,155	3,254	94%	0.79	0.80	0.80

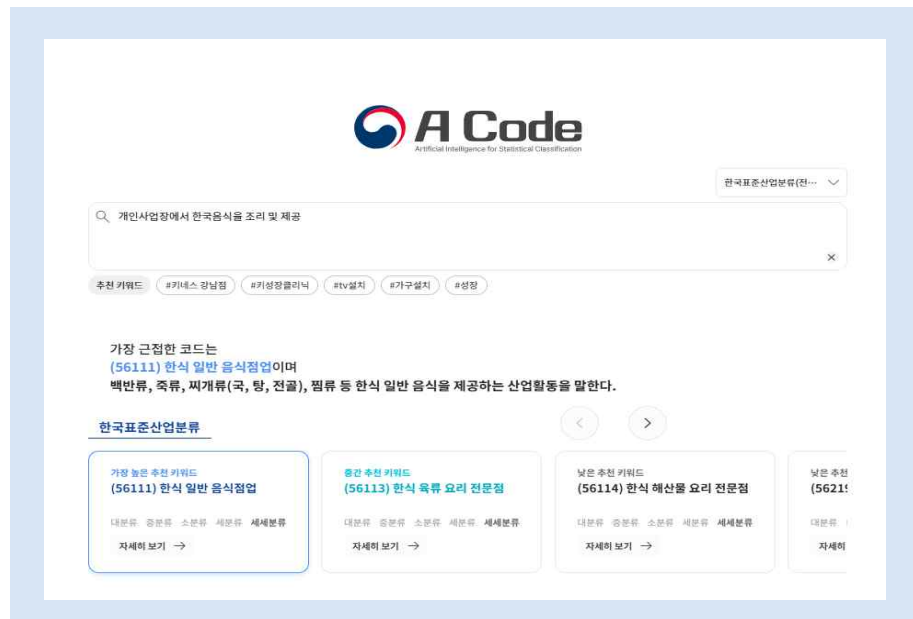
시사점

- **(업무량 감소)** 분류코딩 업무의 실질적 자동화 기능을 확인하였음. 계층형 모델을 통해 **공중 91%, 발주자 94%**의 정분류율을 달성함으로써 지방청 담당자의 수동 코딩 작업 부담을 크게 줄일 수 있을 것으로 기대됨. 특히 **AI 예측 확률이 낮은 건 (예:80% 미만)**에만 담당자가 집중 검토하는 방식으로 자료처리 업무의 효율성을 극대화하고 개선된 데이터를 재학습 해, 모델의 예측 정확도를 높일 수 있음
- **(계층형 방법론)** 계층형 분류 방법론의 유효성을 실증적으로 확인하였음. 발주자 분류와 같이 데이터 불균형이 심하고 계층적 구조를 가진 분류체계에 대해 계층형 다중 작업 학습이 매우 효과적임을 확인하였음. 이는 한국표준산업분류·직업분류뿐 아니라, **자체 분류체계를 사용하는 다양한 통계조사**에 대해서도 **AI 자동분류를 확대 적용할 수 있는 방법론적 기반**을 제공함
- **(실무 활용)** AI통계분류시스템의 적용 범위를 확장할 수 있음. 현재 5종의 대규모 조사에 한정된 시스템을 건설경기동향조사와 같은 월별 조사로 확대함으로써, 통계 생산 전반에 걸친 **AI 기반 자동화 체계(AX)** 구축에 기여할 수 있을 것으로 기대됨

한계 및 향후 연구

- **(학습데이터 한계)** 본 모델은 **경인지방데이터청의 수주 데이터**만으로 학습했다는 점에서 한계가 있음. 특히 발주자분류의 F1-macro(0.79)는 정분류율(94%)에 비해 상대적으로 낮는데, 이는 민자유치사업 등 소수 분류단위의 데이터 부족에 기인함. 향후 **전국 단위 데이터 확보**를 통해 소수 클래스의 학습 데이터를 보강할 필요가 있음
- **(향후 과제)** 전국 5개 지방청의 데이터를 모두 학습할 경우 지역별 공사 발주자 특성을 반영한 표준 모델로의 고도화가 가능할 것으로 기대됨. 실무 활용을 위해서 5개 지방청 전체 수주 데이터를 통합한 **전국 표준 모델 구축**을 통해 **나라통계시스템과의 실시간 연계**를 검토해 볼 수 있음

[그림5] AI통계분류시스템 메인 홈페이지



참고 문헌

- 오교중 외 (2023), 인공지능 통계분류 자동화 확대 적용 연구, 국가데이터연구원
- 임경민 (2023). AI 통계분류 결과분석 및 실무활용성 제고방안 연구, 국가데이터연구원
- Kevin Clark et al. (2020). "ELECTRA: Pre-training text encoders as discriminators rather than generators."
- Wehrmann et al. (2018). "Hierarchical Multi-Label Classification Networks."
- Yinhan Liu et al. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach."

출처를 밝히지 않고 국가데이터연구원 데이터 리서치 브리프의 내용을 무단전재 또는 복제하는 것을 금합니다.
본 리서치 브리프의 내용은 필자들의 개인적 의견이며, 국가데이터연구원의 공식적인 의견이 아님을 밝힙니다.

발 행 일 2026년 7월 31일

발 행 처 국가데이터처 국가데이터연구원

발 행 인 김 진

주 소 35220 대전광역시 서구 한밭대로 713 통계센터 6~8층

전화번호 042-366-7118

홈페이지 mods.go.kr/dsri