발 간 등 록 번 호
11-1241140-100001-10

2025년 연구보고서

AI 기반 건설경기동향 자동분류 코딩 연구

2026. 3.

https://mods.go.kr/dsri

국가데이터처
국가데이터연구원

연구보고서 2025-14

AI 기반 건설경기동향 자동분류 코딩 연구

이규호 · 임경민

Ministry of Data and Statistics
Data and Statistics Research Institute

발간사

“데이터의 가치는 분석과 활용을 통해 의사결정을 지원하고, 혁신과 효율성 향상 등 구체적인 성과를 창출하는 데서 비롯됩니다.”

급변하는 불확실성의 시대에 데이터는 더 이상 단순한 숫자의 기록이 아니라, 미래를 예측하고 사회 문제를 해결하는 핵심 나침반으로 자리매김하고 있습니다. 국가데이터연구원은 이러한 시대적 요구에 부응하여 국민의 삶을 실질적으로 개선하고 AI 기반의 공공 AX 대전환을 뒷받침하기 위한 데이터 기반 연구에 지속적으로 매진해 왔습니다.

2025년 연구보고서에는 우리 사회가 직면한 환경 변화에 능동적으로 대응하고자 첨단 기술을 국가통계에 접목하기 위해 치열하게 고민한 연구 성과를 담았습니다.

첫째, 인공지능(AI) 기반 국가통계 기술혁신을 선도하고자 노력하였습니다.

생성형 AI 기술을 현장조사에 적용하기 위한 기초연구를 통해 조사자료의 내용검토 및 자동분류, 질의응답에 활용 가능성을 모색하였으며, 이는 통계 생산의 신속성과 정확성을 획기적으로 제고하는 토대가 될 것입니다. 아울러 생성형 AI를 활용한 나우캐스트 지표 서비스 제공 방안 연구는 통계서비스의 새로운 가능성을 여는 의미 있는 첫걸음이라 할 수 있습니다.

둘째, 점차 열악해지고 있는 조사환경에 대응하기 위해 새로운 통계방법론 연구와 국가통계 품질제고를 위한 연구를 강화하였습니다.

확률표본과 자원자표본을 통합한 추정 방안 연구는 응답자 조사 부담을 완화하고 비확률표본의 병행 활용 가능성을 제시하였으며, 데이터 과학기술을 활용한 자료수집 개선 연구와 데이터 통합방법 연구는 다양한 데이터의 연계·통합 방법을 보다 체계화하였습니다.

셋째, 사회적 사각지대를 조명하고 지속가능한 미래를 지원하기 위한 데이터 기반 정책 연구에 집중하였습니다.

최근 심각한 사회 문제로 대두된 '고립·은둔 청년'의 실태 파악을 위한 조사 문항 개발 연구를 비롯하여, 돌봄 분야 국가통계 활용 방안과 국내 최초의 기후변화 통계·지표 분석 연구는 데이터가 사회안전망 강화에 기여할 수 있음을 보여줍니다. 또한 소득이동통계 심층 분석 연구와 생애과정 이행에 대한 중·고령기 비교 연구는 관련 정책의 실효성과 활용도를 한층 높일 것으로 기대됩니다.

아울러 가계동향조사의 소비지표 작성 연구와 퇴직연금 적립금 배분 방법 연구는 국민의 체감 경기를 보다 정확히 진단하고 합리적인 경제정책 수립을 지원하는 든든한 기반이 될 것입니다.

2025년 10월부터 새롭게 출발한 국가데이터처 국가데이터연구원은 앞으로도 최신 기술과 사람을 잇는 데이터 연구를 통해 국가통계의 지평을 지속적으로 확장해 나가겠습니다.

본 연구보고서가 통계 생산자와 이용자 모두에게 실질적인 도움이 되고, 각계각층의 의사결정자에게 깊이 있는 통찰을 제공하기를 기대합니다.

많은 관심과 성원을 부탁드립니다.

2026년 3월

국가데이터연구원장

목 차

요 약

본 연구는 건설경기동향조사 수행 과정에서 발생하는 공종 및 발주자분류코딩 작업의 업무량을 완화하고 조사 결과의 정확성을 제고하기 위한 인공지능 기반 통계분류 자동화 모델의 시험 구축을 목표로 한다.

본 연구에서는 먼저 지방데이터청의 건설경기동향조사 자료처리 현황을 파악하고, 현재 국가데이터처에서 사용 중인 유사하지만 더 복잡한 분류체계를 가진 건설업조사의 공종 및 발주자 분류 모델과의 차이점을 비교 및 분석하였다. 이를 통해 건설경기동향조사 자료와 분류체계에 맞는 AI 통계분류 모델의 개발 방향을 설정하였다. 방법론으로는 자연어 처리 기술 중 분류에 탁월한 성능을 보이는 BERT 계열의 인공지능 모델(ELECTRA)을 활용하여 수주 조사표 텍스트 기반으로 공종 및 발주자 부호를 자동으로 분류하는 모델의 시험 구축을 진행하였다.

구축된 모델의 실무 적용 가능성을 검증하기 위해, 2025년 경인지방데이터청에서 입수된 건설경기동향조사의 실제 수주조사 데이터를 사용하여 AI 모델의 분류예측 정확도를 정량적으로 측정하는 실험을 수행하였다. 또한, 모델의 실무 활용성 검토를 위해 국가데이터처에서 현재 운영 중인 AI통계분류시스템에 적용할 수 있는 방안을 사전 검토하여, 향후 지방청 단위에서 활용할 수 있는 방향을 제시하였다.

본 연구를 통해 지방청 단위에서 건설경기동향조사에 대한 업무 효율성 향상뿐만 아니라 한국표준산업분류 및 한국표준직업분류 외에 기타 분류 체계를 가지고 있는 조사들에 대한 인공지능 기반의 통계분류 시스템 도입 및 확산에 기여할 수 있기를 기대한다.

주요 용어 : 인공지능, 건설경기동향조사, 자연어처리, AI통계분류시스템

제 1 장

서 론

최근 다양한 인공지능 및 생성형 AI(Generative AI) 기술의 등장으로 공공 부문에서 AI를 활용한 대국민 서비스 및 내부 행정 업무 효율화 시도가 활발히 진행되고 있다. 공식통계(Official Statistics) 분야 역시 이러한 시대적 흐름에 발맞추어 AI와 머신러닝(ML) 기술을 통계 생산 및 서비스 전반에 적용하려는 노력을 가속화하고 있다.

한편 공식통계의 AI 활용에서 주요 과제 중 하나는 비정형 데이터(Unstructured Data) 처리이다. 공식통계 분야에서 비정형 데이터는 귀중한 정보 자산이지만, 동시에 처리 비용과 품질 관리 측면에서 신중한 접근이 필요한 영역이다. 그 이유는 텍스트로 수집된 정보는 통계분석이 가능하도록 표준화된 코드로 변환하는 코딩(Coding) 작업을 거쳐야 하기 때문이다. 이러한 수동 코딩 과정은 단어의 모호성으로 인해 오류와 편향을 유발할 수 있으며, 대량의 데이터 처리 시 상당한 시간과 인력이 소요된다는 한계를 지닌다. 따라서 비정형 데이터의 처리는 통계 품질 확보와 업무 효율성 제고를 위해 지속적인 개선이 필요한 과제이다.

국가데이터처는 이러한 분류코딩 문제를 해결하기 위해 체계적인 접근을 추진해 왔다. 2020~2021년의 기초연구를 바탕으로 2022년에는 지도학습 기반의 인공지능 통계분류 자동화 시스템 개발 사업[1]을 완료하였다. 이를 통해 인구총조사, 지역별고용조사 등 대규모 5종 조사들에 대해 시범 적용하여 분류예측 결과를 실무에 활용하고 있다. 또한 가계동향조사의 K-COICOP 항목분류에 대한 선행연구를 토대로 2026년에는 실무 적용을 준비하고 있어, AI 기반 자동분류 시스템의 적용 범위가 점차 확대되고 있다.

본 연구에서는 위 시도에 이어 통계법에 의거한 지정통계(승인번호 제101016호)인 건설경기동향조사의 공종 및 발주자 자동분류 모델을 시범 구축하고, 그 타당성과 실무활용 방안을 검증하는 선행 연구를 수행하고자 한다. 이를 통해 다양한 통계 조사에 인공지능 기반 자동분류 방법론을 빠르게 보급 및 확산하여 실무에 활용할 수 있는 방안을 모색하고자 한다.

---

1) 과학기술정보통신부의「디지털 공공서비스 혁신 프로젝트」지원 사업을 통해 개발예산(15.4억)을 확보하고 6개월('22.7~'23.1)의 사업기간을 거쳐 시스템을 개발함

제 2 장

건설경기동향조사 자료처리 현황

제1절 건설경기동향조사 개요[2)]

건설경기동향조사는 통계법에 의거한 지정통계(승인번호 제101016호)로 종합건설업체의 국내 건설공사 수주액 및 기성액을 발주자 및 공사 종류별로 조사하여 국내 건설활동의 단기동향을 신속하게 파악하고 관련 정책수립에 필요한 기초자료를 제공하기 위한 목적으로 실시되고 있는 조사이다.

조사를 통해 월평균 1,500건의 건설수주 자료가 입수되고 있으며 방문조사, 전자조사(Email, CASI, FAX), 전화조사 등 다양한 경로를 통해 매월 자료를 수집하고 있다. 또한 매월 건설업 관련 행정자료를 입수하여 처리하고 있다. 이러한 작업을 거쳐 완성된 통계는 정책 수립이나 기타 연관 통계의 기초자료로 사용된다. 건설 수주자료를 합산하여 국내건설경기동향조사를 공표하고 있고, 경기종합지수, 전산업생산지수 등과 관련이 있으며, 한국은행의 국민계정과도 관련이 있어 정부기관 및 연구소 등에서 국내경기동향 분석자료로 유용하게 활용되고 있다.

<표 2-1> 건설경기동향조사 주요 이용처

주요 이용처	이용자 유형별 용도
대한건설협회	건설수주 원시자료를 활용하여 국내건설경기동향조사 공표
한국은행	건설기성액을 활용하여 국민계정 건설투자지표 산출
기획재정부	각종 경제정책 수립 및 경제동향 분석자료로 활용
국가데이터처	경기종합지수, 전산업생산지수 등의 기초자료로 활용
언론, 연구소, 증권회사	국내 경기동향분석자료

2) 국가데이터처「건설경기동향조사 지침서」및「통계이용자정보 보고서」내용을 토대로 정리함

제2절 건설경기동향조사 공종·발주자 분류

1. 조사항목

건설경기동향조사 조사항목 체계는 크게 건설수주와 건설기성의 2가지로 나뉜다. 이번 연구에서는 건설수주 데이터를 사용하기에 건설기성은 생략하도록 한다. 조사항목으로는 공사명, 공사세분류명, 공사지역, 발주자명, 발주자세분류명, 수주액, 착공예정연월, 완공예정연월 등 8건으로 구성되어 있다. 주요 항목에 대한 조사목적은 <표 2-2>로 갈음한다.

<표 2-2> 주요 항목의 조사목적 정리

주요 항목	조사 목적
공사명	공사내용 파악 및 공종분류를 위한 정보 파악
공종 세분류명	공종세분류를 위한 정보 파악
공사지역	시도별 공표를 위한 정보 파악
발주자명	발주자세분류 및 공종세분류를 위한 정보 파악
수주액	주 공표항목
착공예정연월, 완공예정연월	수주액에 대한 내검, 수주자료와 기성자료 간의 비교·분석 및 향후 건설투자규모 예측에 활용

주요 항목 중 공사명과 발주자명에 대한 텍스트 정보를 토대로 공종 부호(19개), 발주자 부호(27개)를 통해 구분하여 분류하게 된다. 다른 통계조사에서 사용되고 있는 분류체계인 한국표준산업분류[3]와 한국표준직업분류[4]의 개수와 비교하면 분류코드의 복잡성이 크지 않으나 다중클래스 분류의 관점으로는 많은 편이라고 할 수 있다.

3) 한국표준직업분류의 경우 8차 기준 대분류(1자리, 영문대문자) 10개, 중분류(2자리 숫자) 57개, 소분류(3자리 숫자) 167개, 세분류(4자리 숫자) 495개, 세세분류(5자리 숫자) 1,270개로 총 5단계로 구성

4) 한국표준산업분류의 경우 11차 기준 대분류(1자리, 영문대문자) 21개, 중분류(2자리 숫자) 77개, 소분류(3자리 숫자) 234개, 세분류(4자리 숫자) 501개, 세세분류(5자리 숫자) 1,205개로 총 5단계로 구성

2. 조사문항

<표 2-3>의 공종분류 주요 조사 문항으로는 공사명 항목이 있다. 해당 공사가 어떤 공사인지 텍스트 형식으로 기입되며 앞에 해당 공사가 증액되었거나 및 감액되었을 경우 이에 대한 정보가 괄호와 함께 맨 앞에 기입되어 있다. 이러한 비정형 텍스트를 토대로 해당 공사의 분류코드가 무엇인지, 해당 공사가 어느 지역인지를 나타내는 정해진 숫자로 기입하게 된다.

<표 2-3> 건설경기동향조사 공종분류 조사문항 예시

일련번호	공 사 명	공종세분류명 *분 류 부 호	공사지역 *지역부호
1	강원 영동선 동백산-도계 간 철도이설공사	철도·궤도 2 0 5	강원 3 2
2	(증액) 대전 통계아파트 신축공사(1,234세대)	신규주택 1 1 1	대전 2 5
3	(감액) 동북선 도시철도 민간투자사업 건설공사	철도·궤도 2 0 5	서울 1 1

<표 2-4>의 발주자분류 조사문항에 대해 살펴보면 발주자명이 있고 분류 부호가 옆에 기입된다. 그 옆에 수주액 정보와 착공연월 완공연월 등을 기재하게 된다. 발주자분류는 공종분류와 달리 4자리로 되어 있는 것을 볼 수 있다.

<표 2-4> 건설경기동향조사 발주자분류 조사문항 예시

발 주 자 명	발주자세분류명 *분 류 부 호	수 주 액 (백만원)	착공 예정 년 월	완공 예정 년 월
한국철도공사	공기업 1 0 3 0	1 2 3 4 5	2021 년 7 월	2022 년 12 월
한국토지주택공사	공기업 1 0 3 0	6 7 8 9	2021 년 7 월	2023 년 2 월
동북선경전철(주)	운수업 2 2 1 1	- 1 2 3	2021 년 7 월	2024 년 6 월

전체 중 약 80%의 정도는 행정자료를 통해 정보들을 받고 있지만 공사명과 발주자명에 대한 내용을 담당자가 확인하여 분류코드를 부여하거나 해당 코드가 적정한지 확인하는 내용검토 과정 같은 단순 반복 작업이 매월 병행되고 있다.

3. 분류체계

가. 공종분류

<표 2-5>의 공종분류 항목표를 보면 건축(앞자리 1로 시작)과 토목(앞자리 2로 시작)으로 구분되며. 대분류 중 건축의 경우 5개의 공종세분류로 구분되고 토목의 경우 12개의 공종세분류로 나뉘어진다.

<표 2-5> 건설경기동향조사 공종분류 항목표

대분류	건설수주 공종세분류	
건 축	101	주택
	111	· 신규주택
	121	· 재건축
	131	· 재개발
	102	사무실, 점포, 오락·숙박시설
	103	공장·창고
	104	학교·병원, 관공서, 연구소
	109	기타건축
토 목	201	치산·치수
	202	농림·수산
	203	도로·교량
	204	항만·공항
	205	철도·궤도
	206	상·하수도
	207	발전·송전, 옥외 전기·통신
	208	토지조성
	209	댐
	210	기계설치
	211	조경공사
	219	기타

조사를 진행할 때 계약서상의 공사명을 그대로 기입하는 것이 원칙이나 기입된 공사명만으로 공종분류가 어려울 경우, <표 2-6>과 같이 관련 정보를 공사명 옆이나 비고란에 기입하고 있다.

<표 2-6> 건설경기동향조사 공사명 작성 예시

사례	예시
사 례 1	○○면 농촌중심지 활성화 사업 → ○○면 농촌중심지 활성화 사업 (주민복지센터 등)
사 례 2	○○증축공사 → ○○증축공사 (병원 건물)
사 례 3	○○동 ○○2구역 주상복합 신축공사 → ○○동 ○○2구역 주상복합 신축공사 (사무실 위주)

나. 발주자분류

<표 2-7>을 보면 발주자분류의 경우 4자리 숫자로, 크게 공공, 민간, 국내 외국기관, 민자유치 사업으로 나뉜다. 공공 및 국내 외국기관의 경우 비교적 간단하게 나뉘어져 있지만 민간의 경우 제조업과 비제조업 두 개로 나뉜 후 각각의 세분류로 추가 구분되는 형식이다.

<표 2-7> 건설경기동향조사 발주자분류 항목표

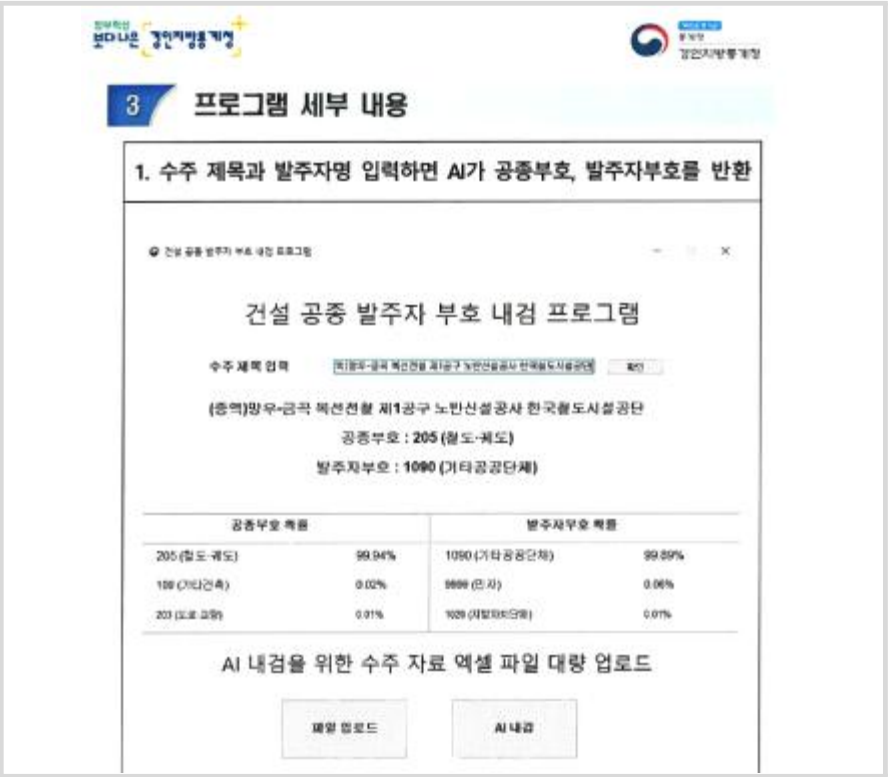
부호	발주자명	부호	발주자명	부호	발주자명
			○ **민간**		
			- 제조업		
		2110	·음식료품 제조업	**3000**	○ **국내외국기관**
	○ **공공**	2120	·섬유·의류 제조업		
1010	- 정부	2130	·석유·화학 제조업	**2******1**	○ **민자유치사업**
1020	- 지방자치단체	2140	·1차 금속 제조업		- 민간의 발주자 분류코드 네 번째에 '0' 대신 '1'로 구분
1030	- 공기업	2150	·기계·장치 제조업		
1090	- 기타공공단체	2190	·기타 제조업		
			- 비제조업		* 외국기업체는 민간으로 분류
		2210	·운수·창고 및 통신업		
		2220	·도소매 금융및사업서비스업		
		2230	·부동산업 및 임대업		
		2250	·건설업		
		2290	·기타 비제조업		

발주자 부호의 분류 항목의 특징 중 하나는 민자유치사업이라는 코드의 존재다. 민자유치사업이란 사회기반시설을 민간부문이 시행 주체가 되어 수행하는 사업으로 항만, 도로, 공항, 철도, 터미널 등 공공재의 재화 성격을 가지면서 소유권이 민간에 있거나 민·관 합동으로 유치한 경우에 해당된다. 민자유치사업 코드의 경우 민간 발주자 분류코드 네 번째 자리에 0 대신 1을 기입하여 구분한다.

공종 및 발주자 부호 모두 눈여겨볼 항목은 각각의 세분류 항목에 해당하지 않는 경우 모두 기타(1090, 2190, 2290)라는 항목 분류로 모여진다는 것이다. 이는 뒤에서 나오는 지도학습 기반 방법론 적용 시 데이터 불균형, 모호성, 문맥 이해의 어려움으로 이어지게 된다.

제3절 건설경기동향조사 인공지능 통계분류 도입 시도

건설경기동향조사의 분류코딩 작업의 효율성을 높이기 위해 지방청 차원에서도 비슷한 시도가 있었다. 경인지방데이터청에서는 2021년 담당 직원의 아이디어를 토대로 사전학습 언어모델인 BERT를 활용해 AI 기반 건설경기동향 공종 및 발주자 자동분류 시스템을 개발하였다. SKT에서 만든 한국어 사전학습 모델인 KoBERT[5]를 기반으로 개인 업무용 PC에서 사용 가능하도록 GUI 환경의 응용프로그램 형태로 개발하였다.



<그림 2-1> 경인지방데이터청 자체 개발 분류코딩 프로그램

모델 학습을 위해 약 6년치 수주자료인 약 4만 6천 건(2015년~2020년)을 사용하여 학습을 진행하였고 학습결과 공종분류는 88%, 발주자분류는 91%의 분류정확도를 확인했다. 이후 실무 활용을 위해 2021년 3월부터 6월까지 시범운용 기간을 거쳐 1,096건에 대해 내검자료와의 일치 여부를 점검했다. 그 결과 공종분류 30.6% 발주자분류 20.9%의 비교적 낮은 정확도를 보였다.

---

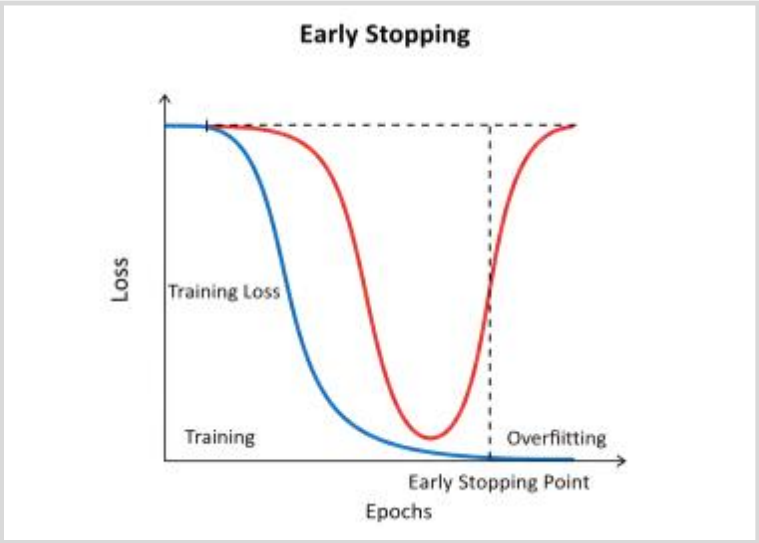
5) SKT Brain에서 개발한 KoBERT는 기존 BERT의 한국어 성능의 한계를 극복하기 위해 수백만 개의 한국어 문장을 학습시킨 모델이다.

<표 2-8> 경인지방데이터청 건설경기동향조사 AI분류 시범운용 결과

구분	정확도	검토 건수	내검 일치 건수		
			본청내검	총괄내검	계
‘21.3.	25.1%	291	159	133	292
‘21.4.	22.8%	224	46	27	73
‘21.5.	26.1%	264	27	24	51
‘21.6	31.2%	317	29	40	69
계	**26.2%**	**1,096**	**57**	**42**	**99**

위 프로그램에 대한 구체적인 코드와 자료가 남아있지 않아 확인하지 못했지만, 검증 데이터와 테스트 데이터를 비교했을 때 성능 차이가 크게 나타난 이유로 학습데이터의 불균형과 반복적인 학습 과정에서 훈련 데이터에만 과도하게 최적화되는 과적합(Overfitting)[6] 현상으로 원인을 추정하였다.

이러한 문제점을 해결하기 위해 이번 연구에서는 각 분류체계별 학습데이터의 분포도를 사전에 파악하고, 모델 학습 시 데이터 과적합에 유의하여 조기 종료(Early Stopping)[7]와 같은 정규화 기법의 도입을 고려하기로 했다.



<그림 2-2> Early Stopping과 Overfitting의 관계

6) 모델이 훈련 데이터에 지나치게 특화되어 학습됨으로써 새로운 데이터에 대한 예측 성능이 저하되는 현상

7) 검증 데이터의 성능이 더 이상 개선되지 않을 때 학습을 중단하여 과적합을 방지하는 기법

제 3 장

인공지능 통계분류 실무활용 현황

제1절 인공지능 통계분류 방법론

통계 작성에 있어 분류 코딩 작업은 여전히 수동 규칙 기반(Rule-based) 시스템[8]과 같은 전통적인 방식에 의존하고 있다. 이러한 방식은 동의어, 다의어 그리고 띄어쓰기 등에 유연하게 대처하기 어려운 단점이 있다. 또한 사전 기반 규칙의 경우 사람이 단어 사전을 주기적으로 업데이트해야만 성능과 커버리지를 지속적으로 유지할 수 있다. 이번 절에서는 이러한 문제를 보완하기 위해 도입된 인공지능 통계분류에 적용된 방법론에 대해 설명하고자 한다.

1. 지도학습 기반 머신러닝

통계 생산 및 자료 처리에 있어 지도학습(Supervised Learning) 접근법은 명확한 정답이 존재하는 경우 매우 효과적이다. 통계자료는 정답이 존재하는 정제된 양질의 대규모 훈련 데이터 그 자체이기 때문에 지도학습 기반 기계학습에는 효과적인 훈련 데이터 세트이다. 인공지능 모델은 학습데이터를 통해 입력값(텍스트)와 정답(분류코드) 간의 패턴을 사전 학습하고 이를 바탕으로 새로운 데이터(신규 입력된 통계조사표)에 대한 정답을 예측한다.

<그림 3-1> Supervised Machine Learning(Nvidia, 2018)

8) 국가데이터처에도 규칙 기반의 분류 시스템인 산업직업자동코딩시스템을 현재까지 운용하고 있으며 이를 대체하기 위해 2023년 AI통계분류시스템을 개발하였다.

가. 데이터전처리

지도 학습은 크게 3가지(데이터전처리, 모델학습, 모델평가) 단계로 진행된다. 그중 데이터는 피처(Feature)와 레이블(Label)로 구성되며 이 둘은 모델의 학습과 예측에 중요한 역할을 한다. 피처는 정제(Cleaning) 등의 전처리 과정을 거치게 되며 기계학습 모델이 이해할 수 있는 형태로 바뀐다. 사전학습 모델 중 하나인 구글의 BERT에서는 임베딩 벡터와 같은 수치형 벡터 형태로 변환된다. 레이블은 모델이 예측해야 할 정답이고 통계분류에서는 주로 코드 형태로 변환된 숫자나 알파벳 형태이다.

<그림 3-2> 데이터 정제(Cleaning) 방법

나. 모델학습

통계분류는 수많은 분류 코드 중 하나를 정답으로 찾아야 하는 다중 클래스 분류(Multi-class Classification) 문제라고 할 수 있다. 일반적인 이진 분류(Binary Classification)와 달리, 다중 클래스 분류는 세 개 이상의 클래스 중에서 하나를 선택해야 하므로 모델의 복잡도와 학습 난이도가 높이진다. 모델이 분류코드를 맞추기 위해 최적의 방법을 찾는 과정이 모델학습이라고 할 수 있으며, 이는 입력 데이터의 특징을 분석하여 각 클래스를 구분하는 패턴을 학습하는 과정을 포함한다.

모델학습 과정을 반복하면서 정확도를 높이기 위해 하이퍼파라미터를 조정하게 되는데, 이러한 하이퍼파라미터에는 학습률(Learning Rate), 배치 크기(Batch Size), 에포크 수(Epochs) 등이 포함된다. 적절한 하이퍼파라미터 조합을 찾기 위해서는 여러 번의 시행착오를 거쳐야 하며, 이 과정을 통해 모델의 성능을 최대화하고 과적합을 방지하는 최적의 학습 방향을 찾게 된다.

다. 모델평가

분류 모델의 성능을 평가하기 위해서는 다양한 지표들이 사용되며, 각 지표는 모델의 서로 다른 측면을 평가한다.

가장 기본적인 지표인 정확도(Accuracy)는 전체 예측 중 올바르게 분류된 샘플의 비

율을 뜻한다. 정확도는 직관적이고 이해하기 쉽지만, 분류 단위별 데이터 불균형이 심할 경우 모델의 실제 성능이 왜곡되는 한계점이 있다.

정밀도(Precision)는 모델이 참(Positive)이라고 예측한 샘플 중 실제로 참인 샘플의 비율을 뜻한다. 즉, 모델이 참으로 판단한 것들이 얼마나 정확한지를 측정하는 지표로, 잘못된 참 예측(False Positive)을 최소화해야 하는 경우 중요한 지표이다.

재현율(Recall)은 실제 참(Postive)인 샘플 중 모델이 올바르게 참으로 예측한 비율을 나타낸다. 또 다른 용어로 TPR(True Positive Rate, 양성률) 또는 통계학에서의 Sensitivity (민감도)와 같다. 재현율은 실제 참인 샘플을 놓치지 않는 것이 중요한 경우 사용한다.

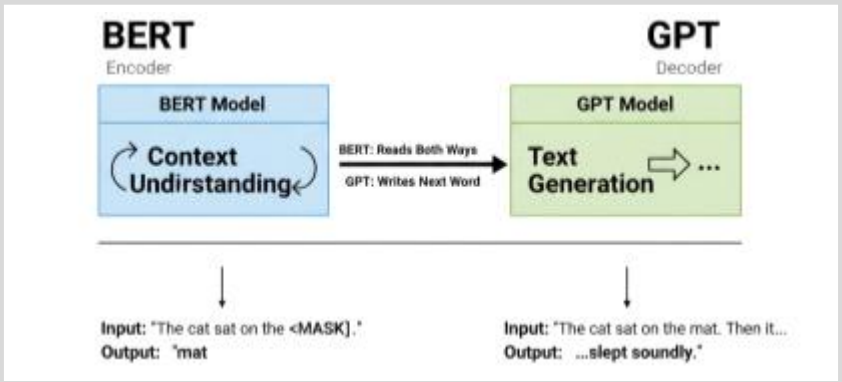
F1-Score는 정밀도와 재현율의 조화평균으로, 두 지표 간의 균형을 고려한 종합적인 성능 지표이다. F1-Score는 0과 1 사이의 값을 가지며, 1에 가까울수록 모델의 성능이 우수함을 의미한다. 통계분류와 같은 다중 분류 문제에서는 F1-score를 확장한 지표들을 사용하는데, 특히 F1-macro는 각 분류 단위별 F1-score를 계산한 후 단순 평균을 취한 값으로, 모든 클래스를 동등하게 취급한다. 따라서 분류 단위별 불균형이 있는 경우에도 소수 분류단위에 대한 성능을 공정하게 평가할 수 있다는 장점을 가진다. 이번 연구에서는 통계분류와 같은 특정 환경에 적합한 모델을 평가하기 위해 <표 3-1>의 다양한 지표들을 사용해 모델을 종합적으로 분석하였다.

<표 3-1> 성능 지표에 대한 설명

성능지표	지표의 의미
Accuracy(정분류율)	전체 예측 중에서 올바르게 예측한 비율 $\frac{TP+TN}{TP+FP+FN+TN}$
Precision(정밀도)	모델이 'A'라고 예측한 것 중에서 실제로 'A'인 비율 $\frac{TP}{TP+FP}$
Recall(재현율)	실제 'A' 중에서 모델이 'A'로 올바르게 찾아낸 비율 $\frac{TP}{TP+FN}$
F1-Score	Precision과 Recall의 조화 평균 $\frac{2\times Precision\times Recall}{\mathrm{Pr}ecision+Recall}$

2. Transformer 기반 언어모델

AI 통계분류시스템에서 사용하고 있는 사전학습 언어모델인 RoBERTa 그리고 ELECTRA는 모두 Transformer라는 아키텍처 기반으로 만들어졌다. 트랜스포머는 2017년 Attention is All You Need라는 논문을 통해 세상에 처음 공개되었는데 문맥 이해에 특화된 인코더 구조와 문장 생성에 특화된 디코더 구조를 사용해 주로 자연어 처리(NLP) 분야에서 강력한 성능을 발휘한다. 앞뒤 문맥의 의미를 파악하는 데 특화된 구글의 BERT와 새로운 문장을 생성하는 작업에 뛰어난 능력을 발휘하는 OpenAI의 GPT가 대표적이다.



<그림 3-3> BERT 모델과 GPT 모델의 비교

가. RoBERTa 언어모델

RoBERTa(Robustly Optimized BERT Approach)는 Meta(구 페이스북)에서 출시한 모델로 BERT 모델을 토대로 사전학습(Pretraining) 방법론과 데이터 학습 부분을 대폭 최적화하여 성능을 개선시킨 모델이다.

성능개선을 위해 시도한 내용 중 첫 번째는 추가 데이터 학습이다. 기존 BERT는 16GB의 훈련 데이터를 사용했지만, RoBERTa의 경우 160GB의 대규모 데이터를 사용해 모델을 학습시켰다. 두 번째 시도는 기존의 BERT 모델보다 더 길고 큰 배치 사이즈로 학습을 진행했다는 점이다. BERT의 경우 배치마다 256시퀀스가 들어가지만 RoBERTa의 경우 8000시퀀스가 들어간다.[9)]

말 그대로 RoBERTa는 BERT를 열심히 튜닝하여 성능을 향상시켰다고 할 수 있다. RoBERTa의 등장은 새로운 모델을 개발하기보다 기존 모델의 추가적인 튜닝 시도만으로 더 높은 성능을 얻을 수 있다는 점을 시사한다.

---

9) RoBERTa:A Robustly Optimized BERT Pretraining Approach(Yinhan Liu at el., 2019) p.5

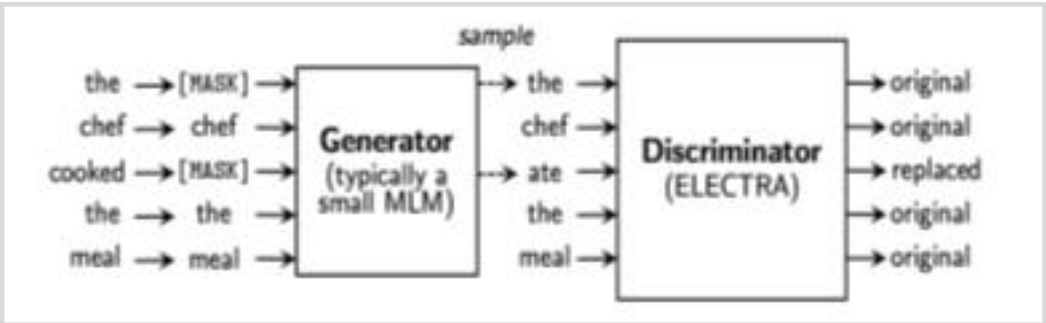
나. ELECTRA 언어모델

ELECTRA는 구글 AI 팀에서 개발한 모델로 자연어처리 모델이 가지고 있는 학습 효율성 문제를 정면으로 겨냥한 모델이다. BERT와 같은 자연어처리 모델들은 MLM (Masked Language Model)[10] 방식을 통해 사전 학습을 진행하게 되는데 마스킹된 약 15% 단어에 대해서만 사전학습이 이루어지고 이는 학습 효율성을 떨어뜨리는 문제를 유발한다.

이러한 문제점을 해결하기 위해 ELECTRA는 RTD(Replaced Token Detection)라는 새로운 사전학습 방법론을 추가해 더 적은 리소스를 가지고 좋은 성능을 보여 주었다. 여기서 예측 인코더 구조를 가진 생성자(Generator)라는 모델과 판별자(Discriminator) 모델이 등장한다.

생성자는 기존의 MLM 방식과 같이 단어를 가리고 예측한다. 이후 판별자가 예측한 각 단어가 원래 단어인지 생성된 단어인지 판별하는 이진분류 문제를 수행한다. 여기서 기존과 다르게 마스킹된 부분만 학습하는 게 아니라 모든 입력에 대해 사전학습을 진행하여 학습 효율성을 높혔다.

ELECTRA는 이 방법론을 통해 4분의 1 수준의 더 적은 컴퓨팅 리소스와 적은 계산량으로 RoBERTa와 유사한 성능 결과를 보였다.[11]



<그림 3-4> Genrator와 Discriminator의 동작 다이어그램(Kevin Clark at el., 2020)

10) 주어진 텍스트 내에서 일부 단어를 가리고(마스킹) 모델이 이 단어를 앞뒤 주변 문맥을 통해 예측한 후 해당 단어 맞췄는지를 훈련하는 자기 지도 학습 방법론

11) ELECTRA:Pre-training text encoders as discriminators rather than generators (Kevin Clark at el., 2020) p.7

3. 계층형 분류 방법론

일반적인 텍스트 분류(Flat Classification)는 정답라벨 간의 관계를 고려하지 않는다. 하지만 한국표준산업분류나 직업분류 같은 통계분류는 대분류(2) → 중분류(25) → 소분류(251) → 세분류(2511)와 같이 명확한 계층 구조를 가지고 있다. 계층형 분류 알고리즘은 이러한 구조적인 정보를 활용하여 복잡한 다중 분류의 정확도를 높이는 하나의 방법론이라고 할 수 있다.

<표 3-2> 플랫 분류와 계층형 분류의 차이

종류	의미
플랫 분류 (Flat Classification)	각각의 정답 라벨이 모두 동일한 계층에 있는 것으로 가정하고 모든 세부 분류 코드(예: K, 64, 614)를 서로 독립적인 별개의 클래스로 취급함
계층형 분류 (Hierarchical Classification)	정답 라벨 간의 계층적 관계를 활용함. 특히 복잡한 계층으로 나누어진 분류 구조에 특화된 성능을 보임

국가데이터처의 AI통계분류시스템에는 이러한 계층형 분류 구조를 지닌 통계분류의 특성을 반영하기 위해 계층형 모델(HiBERT)[12]을 개발 후 실무에 적용하였다. 텍스트 분류 작업의 미세조정(Fine-tuning)을 위한 통계자료 처리 전용 지도학습 모델로 각각의 분류기(Classifier) 4개를 배치해 계층별로 각각의 분류 결과(대-중-소-세분류)를 참고하여 다음 계층의 분류코드를 학습 및 예측하는 방식이다.

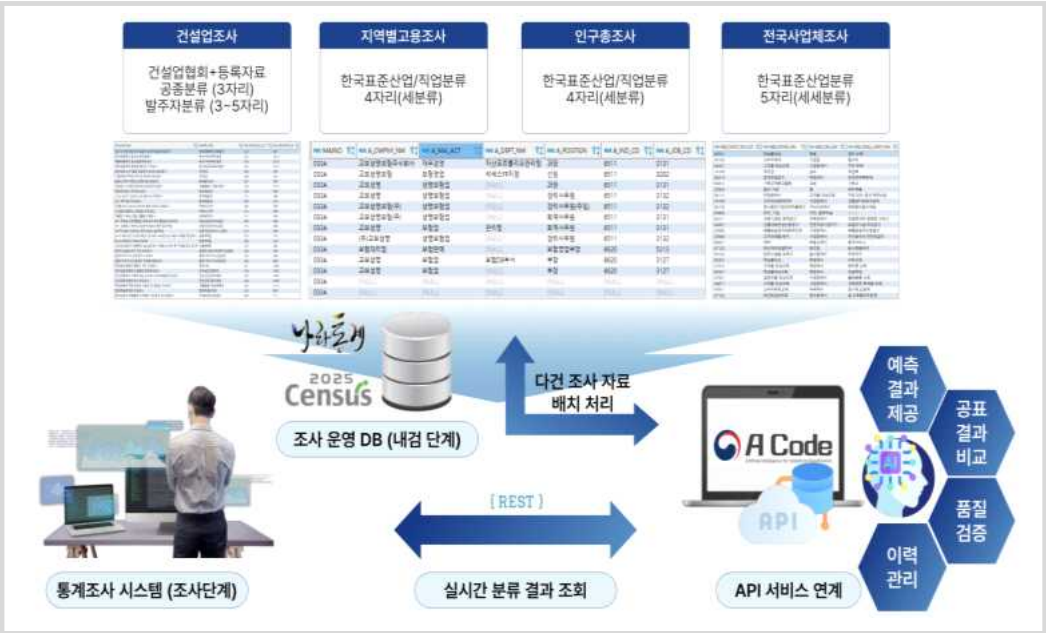
이러한 방식은 여러 계층으로 구성되는 분류체계를 사용하는 경우 매우 효과적이다. 또한 대분류나 중분류 단위만 사용하는 조사들에 대한 분류 지원도 가능하다는 장점을 가진다.

12) 2022년 통계청(현 국가데이터처) 데이터 인공지능 활용대회 수상작 통계분류 특화 모델 하이버트(HiBERT) 모델을 인공지능 통계분류 자동화 시스템 구축 및 개발에 적용

제2절 인공지능 통계분류 자동화 시스템

1. 시스템 현황

국가데이터처는 2023년 지원사업을 통해 인공지능 통계분류 자동화 시스템(AI통계분류시스템)을 개발하여 2024년부터 정식 운영 중에 있다. 본 시스템은 국가데이터처 통계 생산 시스템인 나라통계시스템[13]과 연동되어 분류 결과에 대한 정보를 조사원 및 내검원 그리고 조사 담당 직원들에게 실시간 API를 통해 제공하고 있다.



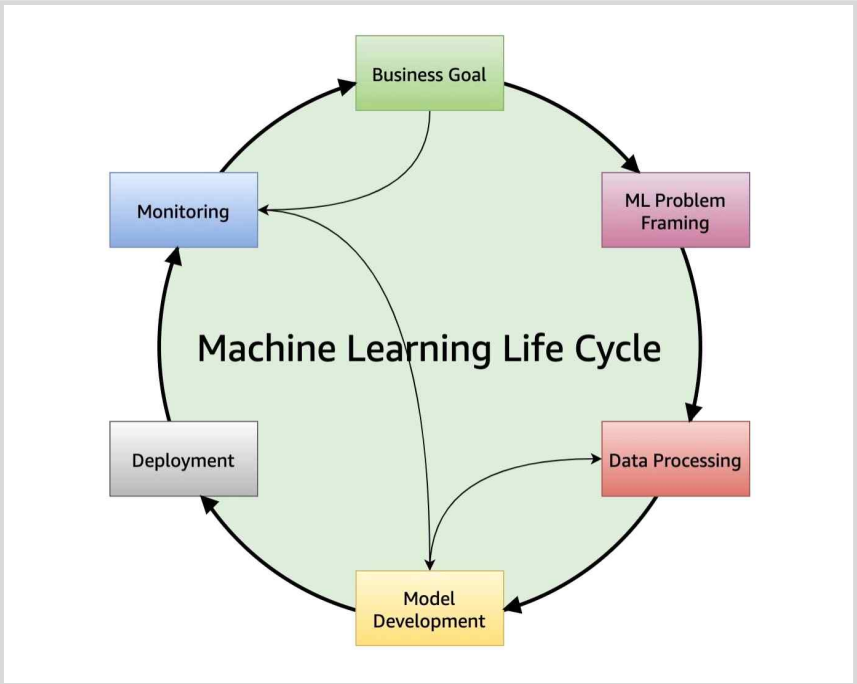
<그림 3-5> AI통계분류시스템 개념도(시스템 유지관리사업 착수 보고서, 2025)

2. 시스템의 특징

일반적인 시스템의 경우 시스템을 잘 구축하고 배포하면 단순 운영 및 유지보수 단계에서 끝나지만 본 시스템은 그렇지 않다. 분류 모델의 정확도 향상을 위해 새로운 학습데이터(신규 공표자료)를 사용하여 주기적으로 AI 모델의 재학습 및 배포가 필요하기 때문이다.[14]

---

13) 국가데이터처에서 운영하는 정보 시스템이며 통계 작성 기관이 공동으로 활용하는 범용 통계정보 시스템이다.



<그림 3-6> Machine Learning Life Cycle(AWS, 2023)

본 시스템을 운영 및 유지 관리하면서 직면하게 되는 가장 큰 위험 요소는 분류체계의 개편[15]이다. 산업 구조의 변화와 기술 발전에 따라 새로운 직업이 지속적으로 등장하고, 기존 직업과 산업의 성격과 범위가 변화하면서 분류체계 역시 이를 반영하기 위해 개편될 수밖에 없다. 이렇게 분류체계가 변경되면 코드 자체가 바뀔 수 있고 기존 산업 및 직업 분류 항목이 추가, 삭제되거나 하나로 통폐합되기도 한다.

문제는 이러한 변화가 생기면 과거 데이터로 학습된 통계분류 모델의 예측 정확도가 낮아진다는 점이다. 모델이 학습한 패턴과 실제 개편된 분류 기준이 불일치하게 되면서, 새로운 분류체계를 반영할 수 있는 모델 재학습이 필요하게 된다. 따라서 본 시스템은 일회성 모델 개발 및 배포로 끝나는 게 아닌, 프레임워크 측면에서 분류 정확도의 지속적인 모니터링 및 재학습 파이프라인 구축을 포함한 머신러닝 생명주기[16]에 따른 관리 체계가 필요하다는 특징을 가지고 있다.

---

14) 주로 통계 공표 시점에 맞추어 연간조사는 1년에 한 번, 반기 조사는 1년에 두 번 재학습을 진행하고 성능 향상폭을 모니터링하고 있다.

15) 한국표준산업분류의 경우 2024.1. 제11차 개정·고시가 있었으며 다음 12차 개정고시는 5년 뒤인 2029.1.로 예정되어 있다

16) 머신러닝 생명주기는 목표에 따른 데이터처리, 모델 개발 튜닝, 모델 배포, 모니터링 단계로 반복된다.

제3절 건설업조사 공종발주자 분류 사례 검토

AI통계분류시스템이 지원하는 5개 통계조사 중 하나는 건설업조사이다. 건설업조사는 우리나라 건설업 부문의 종사자수, 급여액, 매출액, 부가가치, 공사실적 등에 관한 사항을 조사하고, 건설업 부문의 구조 및 활동실태를 파악하여 정책수립에 필요한 기초자료를 제공하기 위한 목적으로 1973년부터 실시되어 온 조사이다.

건설업체에서는 응답자 기입방식으로 각 협회 시스템을 통해 건설경영실태 및 건설공사실적 등의 자료를 입력하고 국가데이터처는 건설 관련 협회 대상 인터넷 조사를 통해 전수조사 방식으로 자료를 입수하고 있다.[17)]

<표 3-3> 건설업조사 공종발주자 분류 조사문항 예시

(2) 일련번호	(3) 공사명	(4) 공사유형	(5) ※ 공종세분류부호	(5-1) 공사지역명	(5-2) ※ 공사지역부호	(6) 발주자명 (7) 하도급자는 아래 칸 괄호 안에 원도급자명을 기입	(8)도급종류 1. 원 도 급 2. 하 도 급 3. 자기공사	(9) ※ 발주자분류부호
				시도 / 시군구 / 동읍면		( )		

연간 약 200만 건의 건설공사 실적자료가 입수되며, 공사명 자료를 기초로 공종 세분류 코드가 결정되고 발주자명 자료를 기초로 발주자분류 코드가 결정된다. 이후 다른 조사들처럼 해당 분류코드가 적정한지 확인하는 내용검토 과정을 거치게 된다.

건설업조사는 건설경기동향조사와 같이 공종 및 발주자에 대한 분류코딩 작업을 진행한다는 점이 같으나 실제로 같은 분류코드를 사용하고 있지 않다. 건설업조사의 경우 훨씬 더 세부적이고 많은 분류코드 체계를 가지고 있다.

또 다른 점은 그 자료처리에 필요한 업무량이다. 건설경기동향조사와 달리 연간조사로 진행되며 200만 건의 행정자료를 처리하고 있다. 이와 같은 자료처리 작업은 많은 인력이 필요한 작업이기에 인공지능 분류 기술을 우선 적용한 이유 중 하나이다.

이와 같은 분류코딩 작업량을 완화하고자 2023년부터 공종발주자 분류 예측결과를 건설업조사 자료처리 과정에 활용하기 시작하여 2025년 현재도 AI 분류 코드 예측확률

17) 임경민(2023), AI 통계분류 결과분석 및 실무활용성 제고방안 연구, p.13

값 정보를 내용검토 우선순위 판단에 활용하는 방식으로 실무에 활용하고 있다. 또한 연간 조사가 끝나면 자료처리가 완료된 데이터를 재학습하여 분류 예측 결과가 떨어지는 분류 라벨의 예측 정확도를 끌어올리고 있다. 2023년에는 자료처리가 끝난 약 70만 건의 학습데이터를 추가 학습해 분류 성능을 지속적으로 향상시켰다.

모델 구축에 앞서 기존에 구축한 건설업조사 분류 모델을 건설경기동향 공종발주자 분류에 사용할 수 있을지 분석해 보았다. 건설업조사도 공사명, 발주자명 같은 텍스트를 공종 발주자 코드로 분류한다. 그러나 양 조사 간의 공식 연계표가 없었기 때문에 직접 연계표를 작성하는 방법을 검토해 보았다.

우선 각 조사의 분류체계를 비교해 보았다. 그 결과 <표 3-4>와 같이 건설업조사 분류체계의 경우 건설경기동향조사처럼 민간 발주자에 대해 민자사업 분류가 별도 코드로 나뉘어 있지 않았다. 건설경기동향의 민자사업 분류 관련 코드는 발주자 분류코드 27개 중 11개로 약 40%의 비중을 차지하고 있다. 이처럼 전체 분류코드 중 민자사업 관련 분류 코드의 비중이 높을 경우 연계표만으로 기존 모델을 사용하는 데 제약이 있어, 민자사업 관련 분류를 수행할 수 있는 별도의 모델을 구축해야만 했다.

<표 3-4> 건설업조사 및 건설경기동향조사 발주자분류 비교표

비교	건설업조사	건설경기동향조사
민간부문 분류코드	전체 96종 중 43개	전체 27종 중 22개
민자사업 분류코드	없음	11개

이 경우 기존의 건설업조사 모델을 재활용하는 장점이 퇴색된다. 건설업조사 분류 모델을 재활용하는 방법 대신 건설경기동향 전용 모델을 구축하는 것이 더 빠르고 효과적인 방법이라고 생각하여 신규 모델을 구축하는 방향으로 연구를 진행하였다.

<그림 3-7> 건설업조사 공종분류(64종) 및 발주자분류(96종) 예시(임경민, 2023)

제 4 장

모델 구축 및 시험 분석

제1절 학습데이터 분석 및 전처리

1. 학습데이터 분석

건설경기동향조사 공종발주자 분류모델 구축에 앞서 모델학습에 필요한 데이터를 경인지방데이터청 경제조사과를 통해 입수하였다. 입수한 데이터는 2015년 1월부터 2025년 5월까지의 수주자료 입력 데이터이며, 총 90,994건을 활용하였다.

가. 학습데이터

입수한 데이터 중 2015년 1월부터 2024년 12월까지의 데이터 89,063건을 학습셋(training set)과 검증셋(validation set)으로 나누어 사용하였다. 각각의 비율은 9대 1로 나누었으며 학습셋 80,155건 검증셋 8,907건이다. 2025년 5월까지의 데이터 1,931건은 모델학습에 대한 평가를 위해 학습에 사용하지 않고 별도로 분리하였다. 연구를 진행하면서 추가로 2025년 6월부터 8월까지 3개월 분량의 수주자료 1,323건을 추가로 입수하여 2025년 1월부터 8월까지 총 3,254건을 모델 평가에 활용하였다.

	NO	조사년도	조사월	기업체고유번호	기업체명	본사코드	분소코드	수주일련번호	공사명	공종부호	...	조사방법	집계구분	응답여부	담당자ID	담당자	비고	수정자	입력일시	수정일시
0	0	1	2015	1	1018[redacted]52	기업(주)	11.0	A00	1	이전 Parcel2B	104	...	E-Mail조사	1. 수주+기성	응답	[redacted]	NaN	중액공사만발생	NaN	2015-02-13 13:40:32	2017-02-15 09:18:24
1	1	2	2015	1	1018[redacted]52	기업(주)	11.0	A00	2	0[redacted]	111	...	E-Mail조사	1. 수주+기성	응답	[redacted]	NaN	중액공사만발생	NaN	2015-02-13 13:40:32	2017-02-15 09:18:24

<그림 4-1> 건설경기동향조사 수주자료 데이터 예시

나. 학습데이터 분포

학습데이터의 분포를 파악하기 위해 트리맵(TreeMap) 시각화 툴을 사용해 분류 단위별 데이터 비중을 분석하였다. 먼저 공종 부호의 분포를 살펴보면 사무실·점포, 오락·숙박시설(102)이 14%의 비중으로 가장 높은 비율을 보였다. 이어서 학교·병원·관공서연구소(104), 신규주택(111), 공장·창고(103) 순으로 비율을 차지하고 있음을 알 수 있다.

<그림 4-2> 건설경기동향조사 수주자료 공종 데이터 트리맵

발주자 부호의 트리맵 분포를 살펴보면 지방자치단체(1020), 공기업(1030), 부동산업 및 임대업(2230) 순의 데이터 분포를 볼 수 있었다. 또한 발주자 부호 중 끝자리가 1로 끝나는 민자사업의 경우 총 27종의 분류코드 중 11개의 높은 비중을 차지하고 있지만 실제 데이터 개수는 2% 미만의 비중을 차지하고 있다. 이러한 민자사업 분류코드의 데이터 불균형은 발주자분류 모델 구축 시 분류 단위별 성능 확보에 영향을 크게 미칠 것으로 예상되었다.

<그림 4-3> 건설경기동향조사 수주자료 발주자 데이터 트리맵

다. 분류라벨

데이터를 입수하고 분리한 후에는 각각 학습에 사용될 데이터의 고유한 라벨을 분석하였다. 학습데이터 안에 공종분류의 경우 고유 라벨이 19개가 존재했고, 발주자분류의 경우 고유 라벨이 22개 존재하였다. 다만 발주자 부호 중 음식료품-민자(2111)와 기계장치-민자(2151) 부호의 경우 데이터 개수가 1개뿐이어서 제외 처리하였다. 또한 stratified sampling[18] 옵션을 통해 훈련데이터의 분류 라벨 단위 클래스의 불균형을 최소화하였다.

```
y_train의 고유한 라벨: [210 102 111 103 211 208 104 207 206 203 109 219 204 205 201 209 131 202
 121]
y_train의 고유한 라벨 개수: 19

y_train의 고유한 라벨: [2220 2230 1090 2150 1020 1030 2130 1010 2110 2210 2290 2250 2140 2211
 2190 2120 2221 2191 2231 3000 2291 2251]
y_train의 고유한 라벨 개수: 22
```

<그림 4-4> 2015년 ~ 2024년 학습데이터의 고유 라벨 분석

2. 피처 값 선정

총 두 개의 모델 구축을 목표로 학습에 사용한 주요 피처 값은 기업체명, 공사명, 공종 부호, 공사지역, 발주자명, 발주자 부호로 총 6개이다. 이 중 공종 부호와 발주자 부호가 정답 라벨이며, 공종 분류 모델의 학습에 사용한 주요 피처 값은 기업체명, 공사명, 공종 부호, 공사지역으로 총 4개이고, 발주자분류 모델의 학습에 사용한 주요 피처 값은 발주자명, 발주자 부호이다. 연구를 진행하면서 공사명 컬럼도 추가하였으나 이 부분은 2절의 모델 개선 부분에서 설명하도록 한다.

3. 데이터전처리

학습에 앞서 데이터전처리를 진행하였는데 전체 데이터 중 공사명 데이터가 누락된 케이스가 발견되어 해당 데이터 로우는 삭제 처리하였다. 그리고 공사명 텍스트 중 (증액), (감액)이 맨 앞에 기입되어 있는 경우가 있는데 해당 텍스트 데이터는 모델의 학습에 불필요하다 생각되어 제거하였다. 위와 같은 전처리 과정을 거친 이후 학습에 사용할 피처 값들을 별도의 컬럼에 하나로 합친 후 모델 학습에 사용하였다.

---

18) 각 클래스의 비율을 유지하면서 훈련 테스트 데이터를 나누는 방법

제2절 학습 및 분류정확도 분석

1. 모델 선정 및 학습

앞서 데이터전처리를 완료한 후 <표 4-1> 환경에서 모델 학습을 진행하였다.

<표 4-1> 연구 수행 환경

항목	시스템 환경
CPU	AMD Ryzen 7800x3d 8 core 16 thread 4.2Ghz
메모리	DDR5 32GB
GPU	Nvidia RTX4070super VRAM 12GB
실행환경	Linux Ubuntu 20.04 LTS cuda version 11.8

학습에 사용한 언어모델은 ELECTRA모델 기반의 한국어 사전학습 모델인 kc-electra-base[19]이다. 현재 AI통계분류시스템에서 사용 중인 RoBERTa-Large와 비교했을 때 적은 자원과 빠른 속도로 학습이 가능하여 모델 실험 및 분석에 유리하다고 판단하였다.

모델학습은 공종 발주자 모델 모두 10회(10 epoch) 이내 반복 학습을 기준으로 진행하였으며, 조기 종료(Early Stopping)기법을 적용하여 모델의 과적합(Overfitting)을 방지하였다. 또한 Checkpoint 기능을 통해 학습 과정을 모니터링하며 각 Epoch마다 모델 성능을 기록해 두어 가장 최적의 분류성능 단계를 저장하였다. 이를 통해 최고 성능 시점의 모델을 바로 사용하거나 이어서 학습할 수 있도록 설정하였다.

A	B	C	D	E	F	G	H
원본 발주자명	원본라벨	1순위 예측라벨	1순위확률	2순위 예측라벨	2순위확률	3순위 예측라벨	3순위확률
국가철도공단	1090	1090	0.999423981	1030	0.000255904	2210	0.0001985
더다솔	2250	2230	0.946101904	2220	0.050333973	2221	0.0009933
에이치엠지에스	2230	2220	0.705727816	2210	0.181127891	2140	0.0464569
한국도로공사	1030	1030	0.999871492	1090	8.42248E-05	1020	1.32E-05
한국도로공사	1030	1030	0.999871492	1090	8.42248E-05	1020	1.32E-05
한국전력공사	1030	1030	0.999874353	1090	7.85355E-05	1020	1.244E-05
국가철도공단	1090	1090	0.999423981	1030	0.000255904	2210	0.0001985
한국토지주택공	1030	1030	0.999870896	1090	7.42423E-05	2250	1.671E-05
한국토지주택공	1030	1030	0.999870896	1090	7.42423E-05	2250	1.671E-05
(주)그린도시개	2230	2230	0.908958972	2220	0.088633396	2221	0.0009194
(주)디씨알이	2230	2220	0.995893121	2230	0.002072237	2210	0.0016005
아이비리얼티공	2230	2220	0.947389662	2210	0.048218705	2230	0.001553

<그림 4-5> 예측 순위별 분류 결과 제공 예시

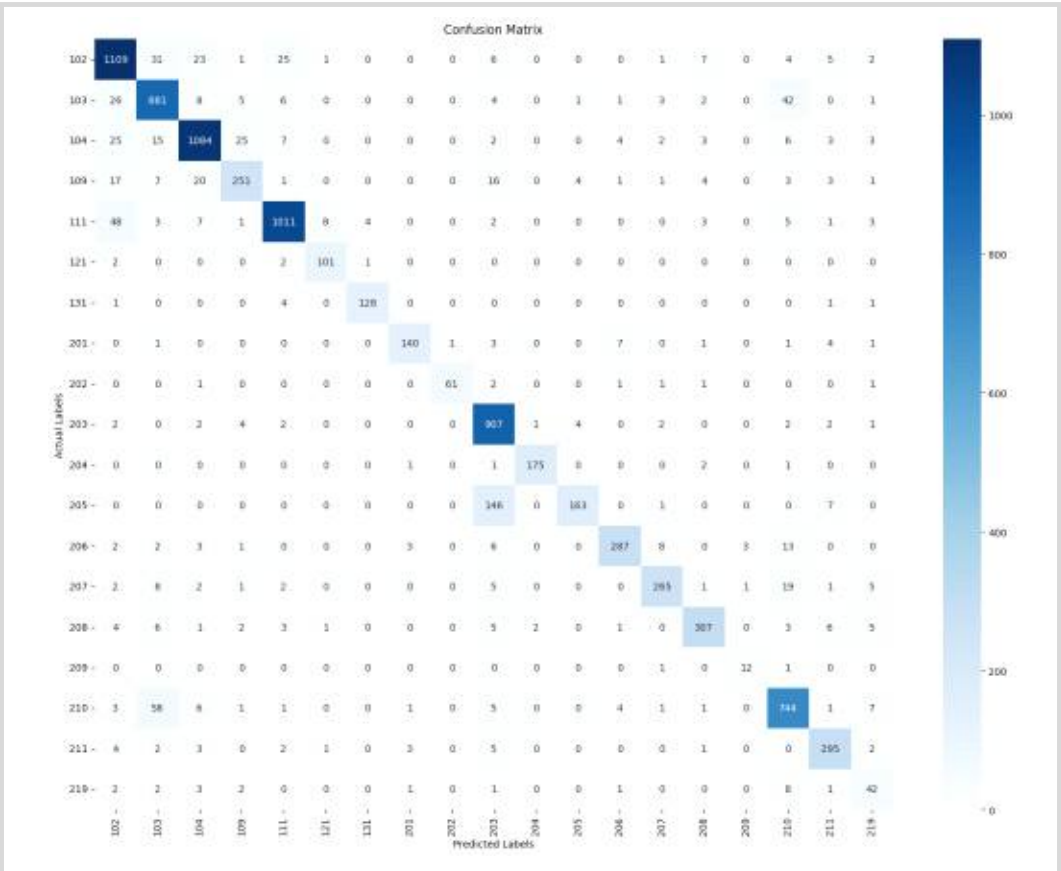
19) HuggingFace의 'beomi/KcELECTRA-base' 오픈소스 공개 모델을 활용하였다.

2. 공종분류 모델 학습 결과

우선 공종분류 모델 학습 결과를 살펴보았다. 성능이 더 이상 오르지 않는 7 epoch 단계에서 학습을 종료하였으며 약 한 시간이 소요되었다. 검증셋 8,907건을 활용해 평가한 결과 공종 모델의 정분류율은 <표 4-2>와 같이 약 89%의 성능을 보였다. F1-macro 역시 0.87로 양호한 결과를 보였다.

<표 4-2> 공종분류 모델 학습 결과

항목	Precision	Recall	F1-score	support
macro avg	0.89	0.87	0.87	8,906
weighted avg	0.90	0.89	0.89	
Total accuracy	0.89 (89%)			

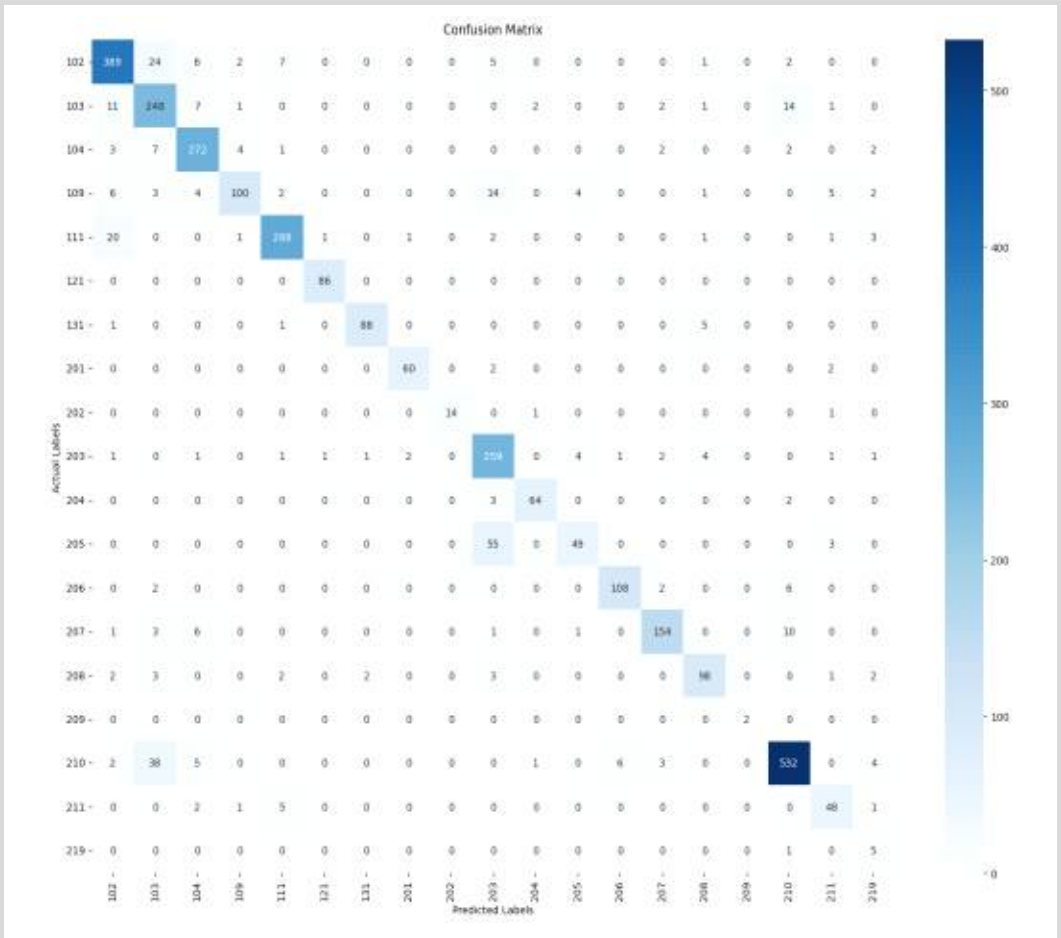


<그림 4-6> 검증셋 평가 결과 혼동행렬(공종분류)

추가로 모델 학습에 사용하지 않은 2025년 1월부터 8월까지의 수주 조사표 3,254건을 활용하여 모델 평가를 추가로 진행하였다. 그 결과 공종분류 모델의 경우 <표 4-3>과 같이 정확도는 약 88%, F1-macro 0.86으로 검증 평가와 유사한 결과를 보였고 기타(219) 분류를 제외한 나머지 결과에서 전체적으로 양호한 결과를 보였다.

<표 4-3> 공종분류 모델 평가 결과

항목	Precision	Recall	F1-score	support
macro avg	0.87	0.87	0.86	3,254
weighted avg	0.89	0.88	0.88	
Total accuracy	0.88 (88%)			



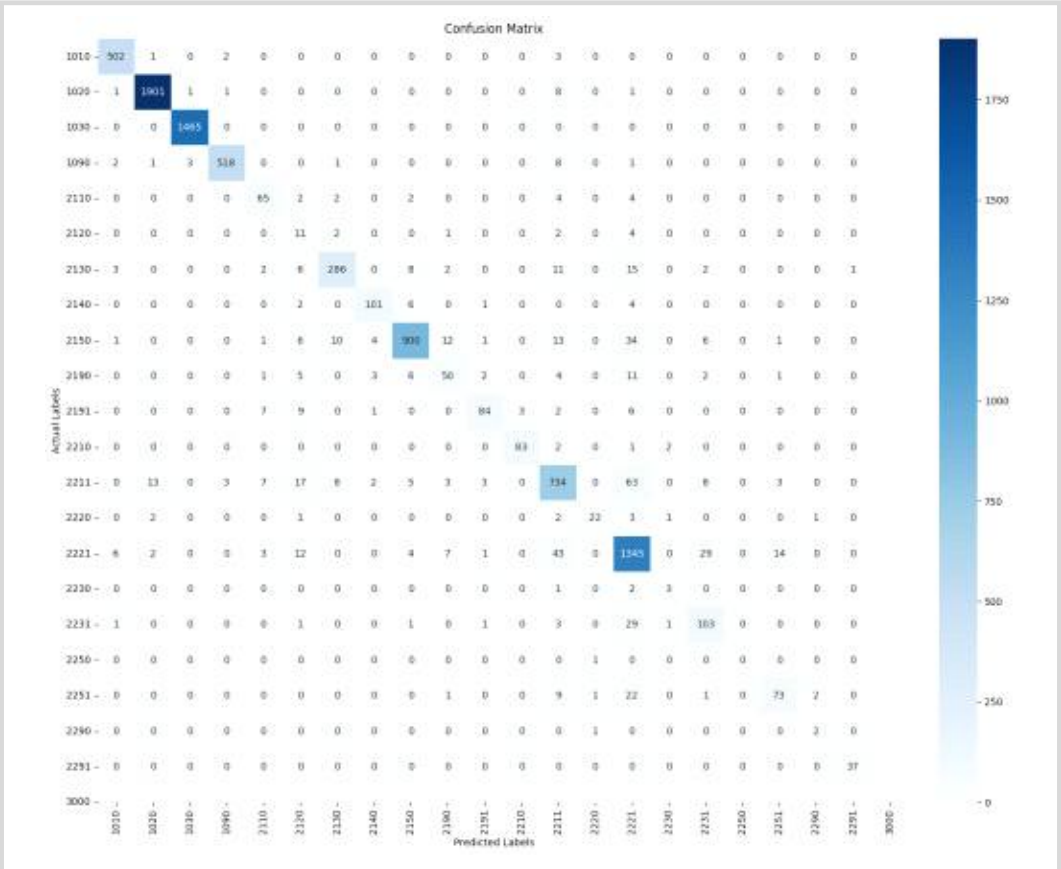
<그림 4-7> 평가셋 결과 혼동행렬(공종분류)

3. 발주자분류 모델 학습 결과

발주자분류 모델의 정분류율을 살펴본 결과 정확도는 <표 4-4>와 같이 약 93%의 성능으로 공종보다 높은 성능을 보였다. F1-macro는 약 0.76으로 공종과 비교하였을 때 조금 떨어지는 결과를 보였다.

<표 4-4> 발주자분류 모델 학습 결과

항목	Precision	Recall	F1-score	support
macro avg	0.77	0.73	0.76	8,906
weighted avg	0.94	0.93	0.93	
Total accuracy	0.93 (93%)			

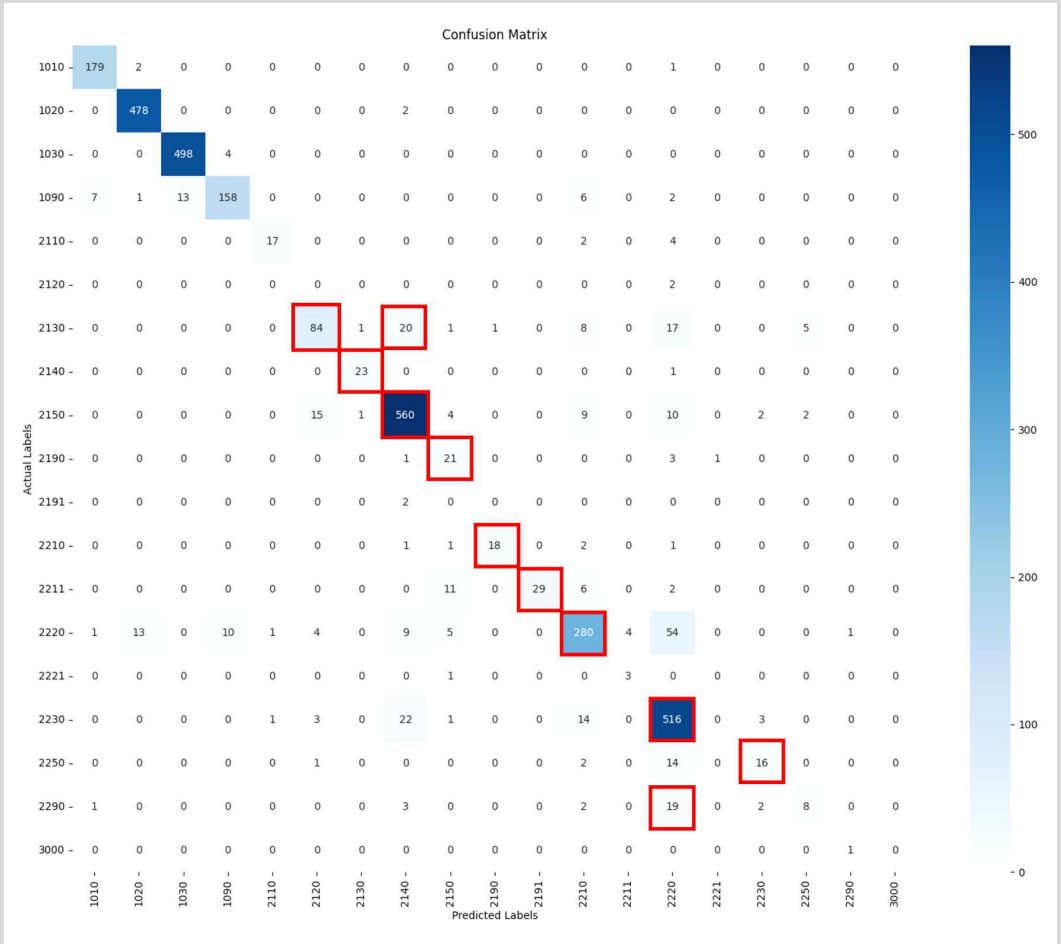


<그림 4-8> 검증셋 결과 혼동행렬(발주자분류)

발주자 모델도 공종과 동일하게 학습에 사용하지 않은 2025년 1월부터 8월까지의 수주조사표 3,254건을 활용하여 평가를 진행하였다. 그 결과 공종분류 결과와 다르게 발주자분류 모델의 경우 정확도는 <표 4-5>와 같이 약 43%, F1-macro 0.25로 검증평가 대비 매우 낮은 결과를 보였다.

<표 4-5> 발주자분류 모델 평가 결과

항목	Precision	Recall	F1-score	support
macro avg	0.27	0.25	0.25	3,254
weighted avg	0.46	0.43	0.42	
Total accuracy	0.43 (43%)			



<그림 4-9> 평가셋 결과 혼동행렬(발주자분류)

평가셋 결과를 토대로 분류단위별 오분류 사례를 분석한 결과 2120, 2130, 2140, 2190, 2210, 2220, 2230, 2250, 2290 등 전체적인 분류 단위에서 다수의 오분류 사례가 나타났고, 끝자리가 1로 끝나는 민자유치사업 분류 코드에서도 성능이 떨어졌다. 특히 학습에 사용된 짧은 발주자명 텍스트만으로 해당 발주자가 민자사업인지 아닌지 여부를 구분할 수 없어 정분류율이 떨어지는 결과가 나타났다. 또한 분류 단위별 학습데이터의 불균형 역시 발주자분류 성능 저하의 원인으로 보였다. 데이터의 분포를 세부적으로 살펴본 결과 끝자리가 1로 끝나는 민자코드는 전체 데이터 89,063건 중 1,302건으로 나타났다.

이처럼 분류 단위별 소수클래스에 대한 모델 성능 개선을 위해서는 절대적인 학습데이터의 추가 확보가 필요했다. 하지만 현재 5개 지방청 전체 데이터를 활용할 수 없어, 다른 방법론을 통해 추가 모델 개선이 필요했다.

4. 모델 개선 시도

가. 피처 값 추가

첫 번째로 시도한 방법은 발주자 모델 학습에 사용되는 피처 값을 추가하였다. 기존 모델은 발주자명 텍스트 데이터만 학습에 사용하였지만 공사명 데이터를 추가 학습 피처로 선정하였다. 공사명 데이터에 민자, 민간과 같이 학습에 도움이 될만한 단어들을 포함하여 추가 학습을 진행한 후 그 결과를 <표 4-6>과 같이 확인하였다.

그러나 모델 평가 결과 공사명 텍스트를 추가한 후, 전체적인 성능 향상을 얻을 수 없었고 오히려 Precision-macro는 0.01, Precision-weighted는 0.02가량 낮아지는 결과를 보였다. 추가 학습된 공사명 텍스트 데이터가 모델 학습에 노이즈 데이터로 작용한 것으로 추측되어 다른 방법으로 모델 개선을 시도하였다.

<표 4-6> 피처 값 추가 발주자분류 모델 평가 결과

항목	Precision	Recall	F1-score	support
macro avg	0.26	0.25	0.25	3,254
weighted avg	0.44	0.43	0.42	
Total accuracy	0.43 (43%)			

나. 키워드 검색 방법 적용

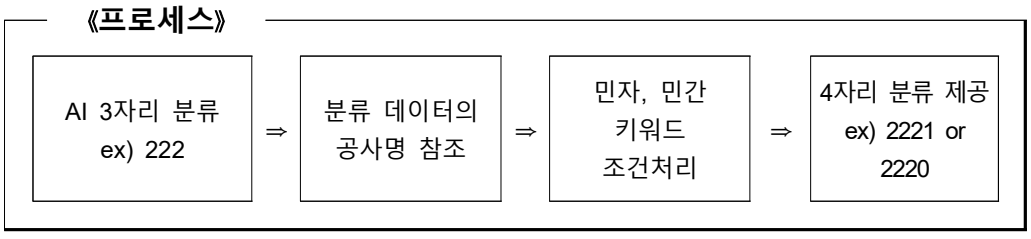
다음으로 시도한 방법은 끝자리가 1로 끝나는 민자사업 라벨을 모두 3자리 분류 라벨로 변환시켜 3자리 예측 모델을 별도로 구축하였다. 이 경우 모델이 분류해야 할 단위가 기존 27개에서 16개로 줄어들게 된다. 3자리 분류 모델 평가 결과 정확도 93%, F1-macro는 0.83으로 높은 점수를 보였다. 기존 모델과 비교했을 때 소규모 분류단위에 대한 문제점이 해소되면서 <표 4-7>과 같이 결과를 얻을 수 있었다.

<표 4-7> 발주자분류 모델 평가 결과 (3자리 코드)

항목	Precision	Recall	F1-score	support
macro avg	0.84	0.82	0.83	3,254
weighted avg	0.93	0.93	0.93	
Total accuracy	0.93 (93%)			

하지만 3자리 분류결과를 현장에서 활용하기에는 어려울 것으로 보인다. 결국 사람이 민자사업 여부를 판단해 4자리 분류결과를 최종적으로 부여하기 위해 추가 수작업이 필요하다. 이는 자료처리 실무활용 단계에서 완벽한 자동화라고 볼 수 없다. 따라서 이를 해결하기 위해 추가 방법론을 적용하여 이를 해결하고자 하였다.

추가 적용한 방법은 키워드 조건 처리 방법이다. 기존에 공사명 데이터를 모델학습 단계에서 사용하지 않고 AI를 통한 3자리 분류 이후, 공사명 텍스트 데이터에서 별도의 키워드 조건 처리를 수행하여 민자와 관련된 키워드 검색을 통해 4자리 코드를 최종 분류하도록 수정하였다.



<그림 4-10> 민자사업 분류를 위한 추가 프로세스 흐름도

동일한 데이터로 평가해 본 결과 <표 4-8>처럼 3자리 분류 모델과 비슷한 성능을 보였으나 성능 향상 결과는 크게 보이지 않았다. 위 방법론은 공사명 데이터의 키워드에 의존성이 높다는 명확한 한계점이 있다. 이 때문에 새로운 데이터가 들어왔을 때

분류예측의 일관성과 정확도를 기대할 수 없을 것으로 보였다. 따라서 키워드를 활용한 방법론을 사용하지 않고 모델 자체를 계층형 분류에 특화되도록 수정하는 방법론을 시도해 보기로 하였다.

<표 4-8> 발주자분류 모델 평가 결과 (3자리 코드+ 키워드)

항목	Precision	Recall	F1-score	support
macro avg	0.81	0.81	0.80	3,254
weighted avg	0.93	0.93	0.93	
Total accuracy	0.93 (93%)			

다. 계층형 방법론

공종분류와 발주자분류 모두 일부 계층적 구조를 가진다는 점에 착안하여 계층형 다중 작업 학습(Hierarchical Multi-Task Learnning) 방법론을 적용하여 계층 분류를 하도록 모델을 수정하였다. Wehrmann 등(2018)의 연구와 유사하게, 분류코드의 모든 계층(L1, L2, L3, L4)을 개별적인 작업(Task)으로 정의하고, 하나의 모델이 이를 동시에 학습하도록 설계하였다.

<표 4-9> 각 모델의 구조 및 방법론

항목	계층형 모델 1 (공종분류)	계층형 모델 2 (발주자분류)
목표	공종명 ⇒ 3자리 공종 부호	발주자명 ⇒ 4자리 발주자 부호
기반 모델	beomi/KcELECTRA-base	beomi/KcELECTRA-base
구조	1 Body + 3 Heads	1 Body + 4 Heads
데이터 라벨	L1, L2, L3	L1, L2, L3, L4
Loss 가중치	L3 (3자리)에 최고 가중치(1.5) 부여	L4 (4자리)에 최고 가중치(2.0) 부여
최종 예측	Head 3 (L3)의 예측 값 사용	Head 4 (L4)의 예측 값 사용

이를 위해서는 학습데이터 준비 단계에서 추가적인 라벨링 작업이 필요했다. 기존의 최종 코드(예: 2151)를 각 계층별로 분리하여 L1, L2, L3, L4라는 새로운 정답 컬럼

을 생성했다(예: L1 = 2, L2 = 21. L3 = 215, L4 = 2151).

이후 ELECTRA 모델을 기반으로 맞춤형 AI 모델을 설계했다. Body(언어 모델)가 먼저 학습에 사용할 텍스트를 입력 받아 텍스트 전체의 문맥적 의미를 우선 파악한다. 그다음 nn.Linear 함수를 사용하여 만든 여러 개의 분류기(Head)[20]가 Body가 파악한 문맥정보를 전달받아 각 계층별로 코드를 예측한다.

이 경우 모델이 여러 개의 정답을 동시에 예측하기 때문에 어떤 계층에 대한 정답을 더 중요하게 학습시킬지에 대한 전략이 필요했다. 이를 위해 계층별로 예측한 오차(CrossEntropyLoss)들을 단순 합산하지 않고 <표 4-10>처럼 별도의 가중치를 부여해 최종 오차(Total Loss)를 계산하였다. 이를 통해 모델이 최종 목표(L4)에 집중하도록 유도하고 앞선 계층의 분류 결과(L1, L2)를 보조로 이용하도록 하였다.

<표 4-10> 손실 가중치 계산식

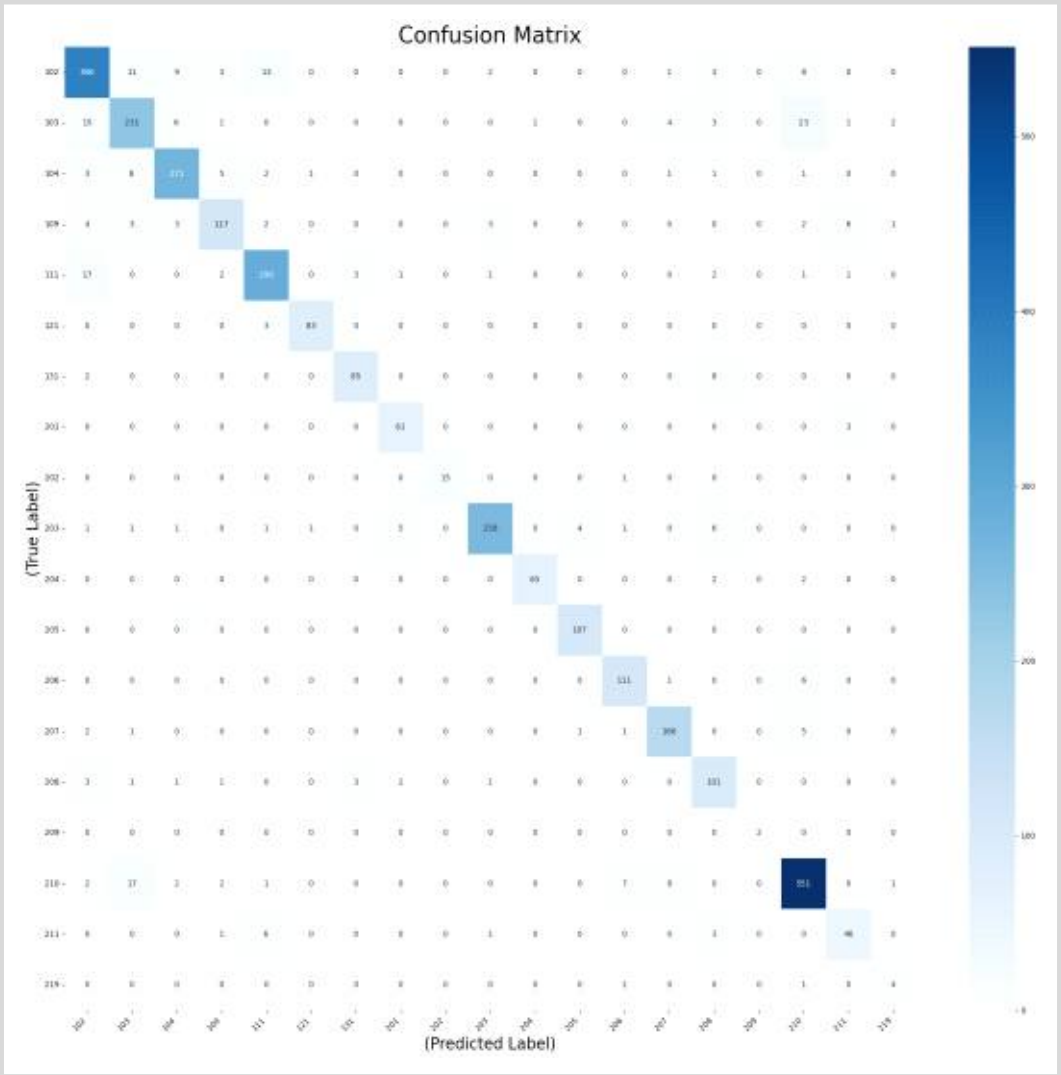
모델	손실 가중치 (예)
공종	$Total\,Loss = (w \times L1\,Loss) + (w \times L2\,Loss) + (w \times L3\,Loss)$
발주자	$Total\,Loss = (w \times L1\,Loss) + (w \times L2\,Loss) + (w \times L3\,Loss) + (w \times L4\,Loss)$

계층형 방법론을 적용한 모델의 성능을 기존과 동일한 데이터셋으로 비교 평가해 보았다. 공종 모델의 경우 정분류율은 91%, F1-macro 0.91로 기존 모델보다 정분류율과 F1-macro 모두 높아졌다. 발주자 모델 역시 정분류율 94% F1-macro 0.79 수준으로 기존의 정분류율 43%, F1-macro 0.25 대비 활용 가능한 높은 수준까지 성능을 끌어올릴 수 있었다. 상세 평가 결과는 <표 4-11>, <표 4-12>와 같다.

<표 4-11> 공종 분류 모델 평가 결과 (계층형)

항목	Precision	Recall	F1-score	support
macro avg	0.91	0.90	0.91	3,254
weighted avg	0.91	0.91	0.91	
Total accuracy	0.91 (91%)			

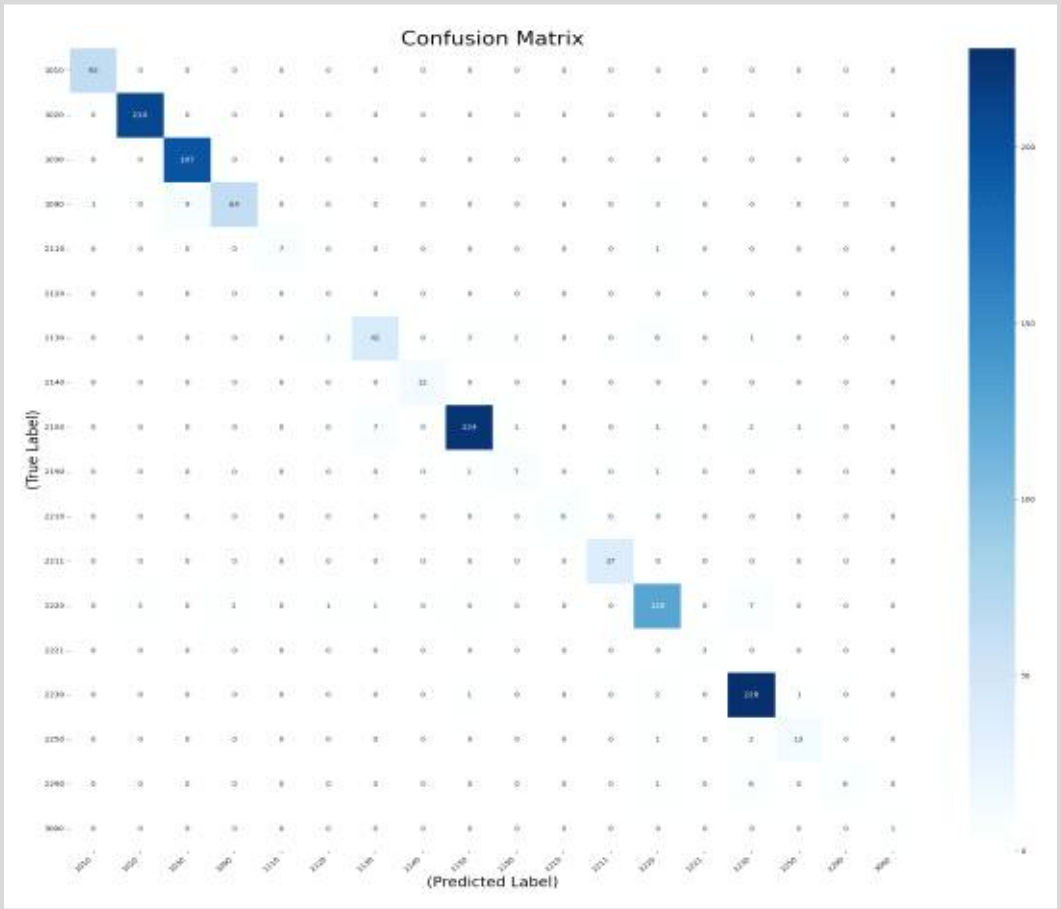
20) 공종분류는 3개의 분류기를 사용하고 발주자분류는 4개의 분류기를 사용했다.



<그림 4-11> 계층형 모델 평가 결과 혼동행렬(공종분류)

<표 4-12> 발주자분류 모델 평가 결과(계층형)

항목	Precision	Recall	F1-score	support
macro avg	0.80	0.80	0.79	3,254
weighted avg	0.94	0.94	0.94	
Total accuracy	0.94 (94%)			



<그림 4-12> 계층형 모델 평가 결과 혼동행렬(발주자분류)

계층형 방법론을 적용한 모델 설계를 통해 현재 보유한 학습데이터 기준 최적의 모델을 만들 수 있었다. 처음 구축한 일반 모델(Flat-model)과 계층형 모델(Hierarchy-model)의 성능을 <표 4-13>와 같이 비교하였다.

<표 4-13> 각 모델별 학습 결과 및 성능 비교표

항목	일반 모델 (공종분류)	계층형 모델 (공종분류)	일반 모델 (발주자분류)	계층형 모델 (발주자분류)
학습데이터	80,155	80,155	80,155	80,155
검증데이터	8,907	8,907	8,906	8,906
평가데이터	3,254	3,254	3,254	3,254
Epoch	7	5	10	10
정분류율	88%	91%	43%	94%
F1-macro	0.86	0.91	0.25	0.79

제3절 현행 시스템을 통한 실무활용 방안

1. 시스템 적용 방안

3절에서는 실무 활용을 위해 현행 시스템을 통해 서비스를 제공할 수 있는 방법을 제안하고자 한다. 현재 시스템에서 지원하고 있는 5종 조사들은 실시간 API를 통해 국가데이터처의 조사시스템(나라통계시스템)과 연계하여 실시간으로 추론 결과를 제공할 수 있도록 개발되었다.

또한 추론 결과를 별도의 파일이나 DB 테이블 형태로 제공하여 충분히 검증할 수 있도록 관련 부서에 자료를 제공하고 있다. 실제로 2025년에 생활시간조사, 가계동향조사, 사회조사 등에 AI 분류결과를 시범 제공하였다.

가. AI통계분류 홈페이지 활용

첫 번째 방법으로 AI통계분류 홈페이지의 검색 기능을 활용하는 방법이 있다. 홈페이지 메인화면에는 검색 기능이 탑재되어 있다. 현재 우측의 선택 박스를 통해 한국표준산업분류, 한국표준직업분류, 가계동향 항목분류 모델을 선택한 뒤 검색창에 텍스트를 입력하면 하단에 추론결과를 확인할 수 있다.[21] 지방청 직원들이 업무수행 과정에서 분류 코딩 작업에 참고하기 위한 지원 도구로서 도입을 고려해 볼 수 있다.

추가로 파일이나 데이터셋 업로드 기능을 활용하여 엑셀 형태로 분류결과를 받을 수 있도록 UI를 구현해 놓았다. 향후 다양한 조사들을 지원하는 것을 염두에 두어 국가데이터처 내 모든 조사들에 대해 리스트가 만들어져 있고 건설경기동향조사에 대해서도 페이지를 만들어 놓았다.

인공지능 분류 추천 값을 1순위부터 최대 10순위까지 받아볼 수 있도록 추천 값 출력 옵션도 존재한다. 향후 시범 도입이 필요한 조사들에 대해 파일럿 테스트용으로 사용할 수 있을 것으로 기대한다.

---

21) 가계동향 항목분류의 경우 2026년 K-COICOP AI 항목분류 도입 전 시험분석을 위해 2025년 9월 실험적으로 배포 적용하였다.

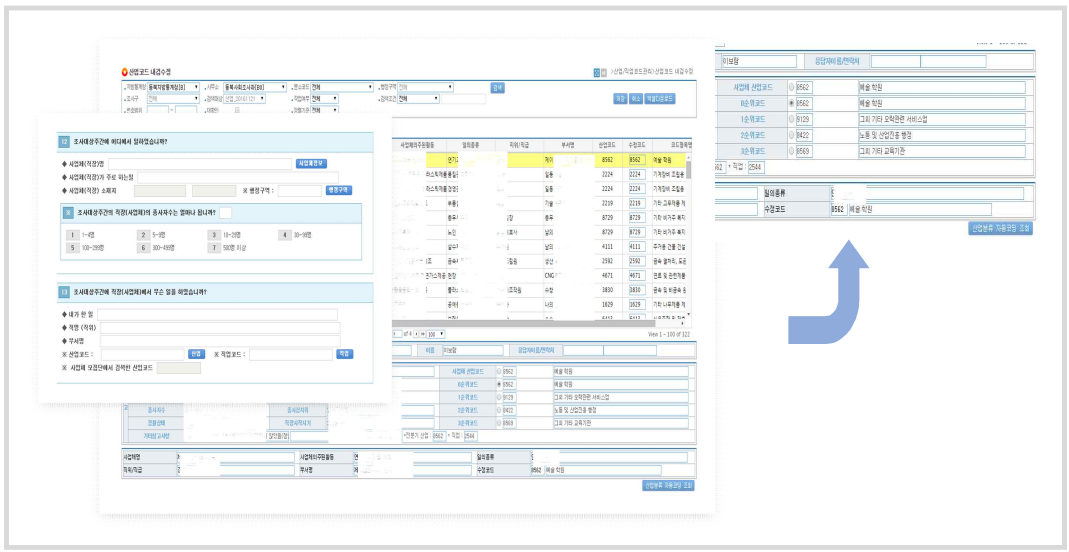
<그림 4-13> AI통계분류시스템 메인 페이지 검색 결과 활용 예시

<그림 4-14> AI통계분류시스템 건설경기동향조사 파일 업로드 기능 페이지

나. 나라통계를 통한 API 결과 제공

국가데이터처의 조사시스템인 나라통계시스템을 통해 실시간 API 결과를 제공하는 방법이 최종 실무 활용 단계로 볼 수 있다. 조사표 입력이나 내검 작업 시 편리하게 사용이 가능하고, 시스템을 운영하는 입장에서도 월별 사용량 집계를 통한 서비스 이용 실적 등을 실시간으로 모니터링할 수 있어 관리에 용이하다.

향후 실무 활용을 위한 파일럿 테스트를 거친 후 AI 분류 정확도에 대한 충분한 검증을 거친 후에는 시스템 정식 탑재를 통해 AI 분류결과를 활용하는 방안을 고려해 볼 수 있다.



<그림 4-15> 나라통계 시스템 화면 예시

제 5 장

결론 및 시사점

제1절 연구 요약

이번 연구는 통계 생산 과정에서 발생하는 분류코딩 업무 자동화의 일환으로 건설경기동향조사의 공종 및 발주자분류에 특화된 인공지능 모델을 시험 구축하는 것을 목표로 하였다.

건설경기동향조사는 매월 발생하는 수주 정보를 바탕으로 담당자가 공사명과 발주자명 등의 비정형 텍스트를 수동으로 검토하여 코드로 분류하는 작업을 수행한다. 이 과정은 단순 반복적이며 오분류 가능성이 있어 업무 효율성 저하의 주요 원인이 되어 왔다. 특히 2021년 경인지방데이터청의 선행 연구에서 시범 운용 정확도가 20~30%대에 그친 사례는, 본 연구에서 반드시 해결해야 할 중요한 과제였다.

이에 본 연구에서는 경인지방데이터청에서 10년간 수행한 약 9만여 건의 수주 데이터를 확보하고, 사전학습 언어모델인 ELECTRA(kc-electra-base)를 기반으로 분류 모델을 구축하였다. 일반적인 분류(Flat Classification) 모델을 적용했을 때, 공종분류는 88%의 양호한 테스트 정확도를 보였으나, 발주자분류 모델은 검증 데이터에서는 93%의 높은 성능을 보이고도 실제 2025년 테스트 데이터에서는 43%라는 매우 낮은 정확도를 기록하는 결과를 보였다.

원인 분석 결과, 발주자분류 체계의 데이터 불균형 문제가 주요 문제로 예측되었다. 특히 전체 발주자분류 중 약 40%의 비중을 차지하고 있는 민자유치사업 분류 종수에 비해 실제 조사 데이터는 전체의 2% 미만이였기에, 모델이 해당 분류에 대한 정보를 충분히 학습하지 못하는 결과를 확인할 수 있었다.

이를 해소하기 위해 계층형 다중 작업 학습(Hierachical Multi-Task Learning) 방법론을 적용하여 문제를 해결하였다. 발주자 분류코드(4자리)가 일부 계층적 구조를 가진다는 점에 착안하여, 모델이 발주자명 텍스트를 입력받아 4개의 분류기(Head)를 통해 각 계층별 코드를 동시에 예측하도록 설계하였다. 이를 통해 모델이 계층별 분류 결과를 보조 정보로 활용하도록 유도하였다. 이와 같은 방법론을 적용한 결과, 공종분류 모델과 발주자분류 모델의 최종 정확도는 각각 91%, 94%로 크게 향상되었다.

제2절 실무 활용을 위한 향후 과제

본 연구를 통해 개발된 모델은 분류코딩 업무의 효율성을 향상시킬 잠재력을 가지고 있다. 그러나 기관 전체 차원의 실무 적용을 위해서는 학습데이터의 편중성이라는 한계점을 보완해야 한다. 본 모델은 경인지방청의 데이터만으로 학습된 만큼, 향후 5개 지방청 전체의 수주 데이터를 통합하여 학습함으로써, 지역별로 다르게 나타날 수 있는 공사 또는 발주자의 특성을 반영해 전국 단위의 표준 모델로 고도화할 필요가 있다.

이와 같은 한계점을 해결한 이후에 국가데이터처의 핵심 조사 시스템인 나라통계시스템과의 실시간 API 연계를 추진해야 할 것이다. 조사표 입력 및 내검 단계에서 AI가 실시간으로 공종 및 발주자 코드를 추천하고, 특히 AI 예측 확률이 낮은 건(예: 80% 미만)이나 신규 키워드에 대해서만 담당자가 집중적으로 검토하는 방식으로 인공지능 통계분류를 활용한 자료처리 업무 체계를 구축해야 할 것이다.

참고문헌

국가데이터처. 건설경기동향조사「조사지침서」,「직무편람」,「이용자용 통계정보보고서」,
국가데이터처. 건설업조사「직무편람」,「이용자용 통계정보보고서」,
국가데이터처. (2023). “인공지능 통계분류 자동화시스템 구축사업”, 사업계획서.
국가데이터처. (2024). “AI통계분류시스템 유지관리 사업 완료보고회”, 발표자료.
국가데이터처. (2025). “AI통계분류시스템 유지관리 사업 착수보고회”, 발표자료.
오교중 외. (2023). **인공지능 통계분류 자동화 확대 적용 연구**. 통계개발원(현 국가데이터연구원).
임경민. (2023). **AI 통계분류 결과분석 및 실무활용성 제고방안 연구**. 통계개발원.

AWS. (2023). *Machine Learning Lens – AWS Well-Architectured Framework*.
Cagia Aksoy at el. (2020). *Hierarchical Multitask Learning Approach for BERT*.
Kevin Clark at el. (2020). *ELECTRA:Pre-training text encoders as discriminators rather than generators*, 7p.
Nvidia. (2018). SuperVize Me: What’s the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning?. blogs.nivida.com/blog/supervised-unsupervised-learning
Wehrmann at el. (2018). *Hierarchical Multi-Label Classification Networks*.
Yinhan Liu at el. (2019). *RoBERTa:A Robustly Optimized BERT Pretraining Approach*, 5p.

부 록

공종 및 발주자 AI 분류예측 정확도

□ 공종분류 결과

코드	항목명	학습 데이터	평가 데이터	precision	recall	f1-score
102	사무실, 점포, 오락·숙박시설	12,148	436	0.8922	0.8922	0.8922
103	공장·창고	9,799	287	0.7561	0.8641	0.8065
104	학교·병원, 관공서, 연구소	11,785	293	0.8977	0.9283	0.9128
109	기타건축	3,288	141	0.9174	0.7092	0.8000
111	신규 주택	10,958	318	0.9381	0.9057	0.9216
121	재건축	1,057	86	0.9773	1.0000	0.9885
131	재개발	1,353	95	0.9670	0.9263	0.9462
201	치산·치수	1,586	64	0.9524	0.9375	0.9449
202	농림·수산	680	16	1.0000	0.8750	0.9333
203	도로·교량	9,286	279	0.7529	0.9283	0.8315
204	항만·공항	1,798	69	0.9412	0.9275	0.9343
205	철도·궤도	3,169	107	0.8448	0.4579	0.5939
206	상·하수도	3,283	118	0.9391	0.9153	0.9270
207	발전·송전, 옥외 전기·통신	3,124	176	0.9333	0.8750	0.9032
208	토지조성	3,462	113	0.8829	0.8673	0.8750
209	댐	141	2	1.0000	1.0000	1.0000
210	기계설치	8,327	591	0.9350	0.9002	0.9172
211	조경공사	3,184	57	0.7619	0.8421	0.8000
219	기타	635	6	0.2500	0.8333	0.3846
accuracy		89,063	3,254	0.8801	0.8801	0.8801
macro avg		89,063	3,254	0.8705	0.8729	0.8586
weighted avg		89,063	3,254	0.8879	0.8801	0.8799

□ 발주자분류 결과

코드	항목명	학습 데이터	평가 데이터	precision	recall	f1-score
1010	정부	5,084	182	0.9836	0.9890	0.9863
1020	지방자치단체	19,125	480	0.9856	1.0000	0.9928
1030	공기업	14,652	502	0.9748	1.0000	0.9872
1090	기타공공단체	5,343	187	0.9765	0.8877	0.9300
2110	음식료품 제조업	793	23	0.8077	0.9130	0.8571
2111	음식료품 제조업(민자)	1	0	-	-	-
2120	섬유·의류 제조업	200	2	0.0000	0.0000	0.0000
2130	석유·화학 제조업	3,361	137	0.8248	0.8248	0.8248
2140	1차 금속 제조업	1,139	24	0.8214	0.9583	0.8846
2150	기계·장치 제조업	9,889	603	0.9595	0.9436	0.9515
2151	기계·장치 제조업(민자)	1	0	-	-	-
2190	기타 제조업	850	26	0.7097	0.8462	0.7719
2191	기타 제조업(민자)	3	2	0.0000	0.0000	0.0000
2210	운수·창고 및 통신업	1,121	23	0.9091	0.8696	0.8889
2211	운수창고 및 통신업 (민자)	878	48	1.0000	1.0000	1.0000
2220	도소매, 금융 및 사업서비스업	8,687	382	0.8923	0.8455	0.8683
2221	도소매, 금융 및 사업서비스업(민자)	315	4	0.8000	1.0000	0.8889
2230	부동산업 및 임대업	14,661	560	0.9042	0.9607	0.9316
2231	부동산업 및 임대업(민자)	63	0	-	-	-
2250	건설업	1,396	33	0.8077	0.6364	0.7119
2251	건설업(민자)	12	0	-	-	-
2290	기타 비제조업	1,090	35	0.8095	0.4857	0.6071
2291	기타 비제조업(민자)	29	0	-	-	-
3000	국내외국기관	370	1	1.0000	1.0000	1.0000
accuracy		89,063	3,254	0.9367	0.9367	0.9367
macro avg		89,063	3,254	0.7982	0.7979	0.7938
weighted avg		89,063	3,254	0.9366	0.9367	0.9356

Abstract

A Study on AI-based Automatic Classification Coding for Construction Cycle Trend Survey

KyuHo Lee, GyeongMin Im

This study presents a pilot development of an AI-based automatic classification model to enhence the efficiency and accuracy of coding type of work and ordering entity for the Construction Cycle Trend Survey

This research first analyzes current data processing procedures at Regional Statistics Offices and compares them with the more complex classification model used for the Construction Industry Survey at the Minister of Data and Statistics (MODS). This comparative analysis informed the development strategy for a model tailored to the Construction Cycle Trend Survey's specific data and classification schema.

Methodologically, a piolot model developed using ELECTRA, a BERT-based architecture recognized for high performance in NLP classification. The model automatically classifies codes from text in the order survey forms.

The model's practical applicability was validated by quantitatively measuring its classification accuracy using 2025 real-world data from the Gyeongin Regional Statistics Office. Furthermore, a preliminary review assessed its integration potential into the existing AI-based Statistical Classification Automation Systems in MODS, proposing a roadmap for adoption at the regional level.

This Study is anticipated to enhence operational efficiency for the Construction Cycle Trend Survey at the regional level. Moverover, it is expected to contribute to the broader adoption of AI-based classification systems for surveys with non-standard classification frameworks, such as those outside the KSIC and KSCO.

Key words: Artificial Intelligence, Construction Cycle Trend Survey, Natural Language Processing, AI-based Statistical Classification Automation

연구진

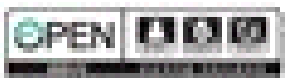
- 이규호(국가데이터처 국가데이터연구원 통계방법연구실 주무관)
- 임경민(국가데이터처 국가데이터연구원 통계방법연구실 사무관)

* 연구진의 소속 및 직급은 연구과제 완료 시 기준임을 알려드립니다.

연구보고서 2025-14

AI 기반 건설경기동향 자동분류 코딩 연구

인 쇄	2026년 3월
발 행	2026년 3월
발 행 인	김 진
발 행 처	국가데이터처 국가데이터연구원 35220 대전광역시 서구 한밭대로 713 TEL.(042)366-7100 FAX.(042)366-7123
홈페이지	https://mods.go.kr/dsri/
ISSN(Online)	2733-4120



국가데이터처
국가데이터연구원