발 간 등 록 번 호
11-1240245-000057-14

OPEN 공공누리 출처표시 상업용금지 변경금지 공공저작물 자유이용허락

2023년 연구보고서

AI 통계분류 결과분석 및 실무활용성 제고방안 연구

- 건설업 공종·발주자분류 사례 검토-

2024. 4.

http://kostat.go.kr/sri

ISSN 2288-1166(Print)
ISSN 2733-4120(Online)

통계청
통계개발원

연구보고서 2023-16

AI 통계분류 결과분석 및 실무활용성 제고방안 연구

- 건설업 공종·발주자분류 사례 검토-

임경민

Statistics Korea
Statistics Research Institute

발간사

기업 경영, 개인의 일상에 이르기까지 합리적 의사결정의 근간인 통계에 대한 중요성이 점점 커지고 활용범위도 넓어지고 있으며, 특히 국가통계는 정책결정에 필수적으로 활용되면서 그 중요성이 더욱 증대되고 있습니다.

이러한 시대의 변화에 따라 통계청은 빅데이터의 활용, 조사자료와 행정자료 간의 연계 등과 같은 통계생산방식의 혁신을 통해서 응답자 부담은 최소화하면서 동시에 보다 정확하고 사용자 친화적인 통계를 만들고자 끊임없이 노력하고 있습니다.

통계개발원은 국가통계의 중추를 담당하는 통계청의 싱크탱크로써 전략적인 연구를 수행하고 있는 국내의 유일한「국가통계 전문연구기관」입니다. 2006년에 설립된 이래 기존의 조사통계를 보다 효율적으로 작성하기 위한 각종 기법과 관련된 통계방법론적 연구는 물론 데이터에 기반한 국가정책이 수립될 수 있도록 경제・사회현상에 대한 심층 분석 연구를 강화하고 있습니다.

또한 저출산・고령사회 현상 등으로 인해 대내외적으로 관심이 높아지고 있는 인구집단 및 인구동향에 관한 분석연구 및 인구동태 관련 방법론 연구를 밀도 깊게 수행하고 있습니다. 이러한 연구의 구체적인 결과를 중심으로 통계개발원은「2023년도 연구보고서」를 발간하게 되었습니다.

이번「2023년도 연구보고서」에는 AI 통계분류 결과분석 및 실무활용성 제고방안 연구 등 데이터과학 연구, 2025년도 인구주택총조사 등 조사표 개선 연구, 경제·사회·환경 변화를 반영한 인구통계, 격자통계를 활용한 도시화 현상 분석 등 인구통계 연구, 인구감소지역과 생활밀접업종 관계 분석 등 경제통계 연구, 비확률표본을 위한 통계적 추론 등 국가통계 방법론 연구, 위성영상을 활용한 국토그린지표 개발 기초연구 등 SDG 지표 관련 연구 등을 수록하고 있습니다.

본 연구보고서는 통계개발원이 전년에 국가통계 개선·개발을 위해 수행한 연구과제로서 국가통계 생산자의 통계개발 및 개선에 유용한 자료로 활용되고 될 수 있기를 기대합니다. 앞으로도 통계개발원이 "국가통계 전문연구기관"으로서 대내외적으로 선도적인 역할을 할 수 있도록 독자 여러분의 지속적인 관심을 부탁드립니다.

통계개발원은 본 연구보고서가 데이터 이용자의 통계 활용에 도움이 되고, 통계 작성자의 통계 개발 및 개선에 유용한 자료로 활용될 수 있기를 기대합니다. 앞으로도 국가통계의 통계연구에 대한 독자 여러분의 지속적인 관심을 부탁드립니다. 아울러 실용적이고 품질 높은 연구 결과를 도출하기 위해 최선을 다한 연구진에게 따스한 감사를 전합니다.

2024년 4월

통계개발원장

목 차

제1장 서론 1

제2장 데이터과학 연구 추진경과 2

제1절 데이터과학 연구팀 신설 2

제2절 데이터과학 기법 국가통계 활용 해외사례 검토 3

제3절 인공지능 통계분류 시험분석 4

제4절 인공지능 통계분류 자동화 시스템 개발 8

제5절 인공지능 통계분류 결과 시험적용 및 고도화 검토(2023) 12

제3장 인공지능 통계분류 결과분석 14

제1절 건설업조사 공종·발주자분류 개요 14

제2절 인공지능 통계분류 방법론 17

제3절 인공지능 통계분류 예측결과 및 실무활용 21

제4절 학습데이터 갱신 및 재학습을 통한 분류정확도 향상 23

제4장 인공지능 통계분류 실무활용성 제고방안 24

제1절 분류체계 개선 24

제2절 조사단계에서 인공지능 통계분류 예측결과 활용 26

제3절 기존 시스템을 인공지능 통계분류 시스템으로 대체 27

제4절 인공지능 통계분류 예측 결과를 활용한 내검 효율화 29

제5장 결론 및 시사점 30

참고문헌 31

부 록 32

Abstract 48

요 약

4차 산업혁명 시대에 빅데이터를 IT기술과 연계하여 신속하고 정확하게 정보를 창출하고 활용하는 방법론으로 데이터과학(Data Science)의 중요성이 강조되고 있으며, 최근에는 인공지능(AI)과 같은 데이터과학 기법을 국가통계에 적용해야 한다는 요청이 증대되고 있다.

국가통계 영역에서 AI 기술 활용 분야로 특히 '통계분류 텍스트 정보를 이용한 분류코딩 자동화'가 주목받고 있으며, 미국 등 많은 국가에서 자연어처리 기술 등을 활용한 비정형 텍스트 분류 자동화 기술을 통계 생산 프로세스에 적용하기 위한 노력을 경주하고 있다.

통계청에서도 미래 통계환경 변화 대응을 위한 데이터과학 연구 일환으로 2020년부터 자연어처리 및 지도학습 방식의 기계학습 분류모델을 활용한 통계분류 코딩 자동화 연구를 수행하였고, 이를 토대로 2022년에는 '인공지능 통계분류 자동화 시스템'을 개발하여 현재 일부 조사를 대상으로 시범적용하고 있다.

본 연구에서는 통계개발원에서 진행한 데이터과학 연구 추진경과를 살펴본 후 인공지능 기반 통계분류 방법론 및 인공지능 통계분류 자동화 시스템 구축 결과를 소개하였다. 그리고 건설업조사에서의 공종 및 발주자분류 사례를 중심으로 AI 통계분류 적용 결과를 분석하고 향후 통계분류 자료처리 과정에서 AI 통계분류 예측 결과를 더욱 효과적으로 활용하기 위한 실무활용성 제공방안을 모색하였다.

이러한 분석 결과는 통계청 통계분류 자료처리의 정확성, 일관성 및 시의성을 제고하는 한편 반복적 업무의 자동화, 지능화를 통한 행정 효율성 향상에도 이바지할 것으로 생각한다. 또한 자연어처리 기술을 적용 가능한 국가통계 전문영역을 개척하고 관련된 기술을 검증함으로써 향후 '비정형 텍스트 분석을 통한 분류자동화' 기술의 보급 및 확산에 기여할 것으로 기대한다.

주요 용어 : 인공지능과 국가통계, 자연어처리, 통계분류 자동화

제 1 장

서 론

4차 산업혁명 시대의 도래 및 조사환경 악화[1)]에 따라 빅데이터 및 행정자료 등을 최신 ICT 기술과 연계하여 정보를 창출하는 데이터과학(Data Science) 방법론을 국가통계에 적용해야 한다는 요청이 증대되고 있다.

특히 최근에는 데이터과학 기법의 국가통계 적용 분야로 '인공지능(AI) 기술을 활용한 통계분류 자동화'에 대해 논의가 집중되고 있는데, 이미 미국·캐나다 등 많은 국가에서 AI 기반 비정형 텍스트 분류 자동화 기술을 통계 생산 프로세스에 적용하기 위한 노력을 경주하고 있다.[2)]

이에 통계개발원에서도 미래 통계환경 변화 대응을 위한 데이터과학 연구를 2020년부터 진행하여 왔는데, 통계분류 자료처리 과정에 자연어처리 및 지도학습 방식의 기계학습 분류 모델 활용 가능성에 대한 기초연구를 수행하였고 이를 토대로 2022년에는 '인공지능 기반 통계분류 자동화 시스템'을 개발하였으며[3)] 현재는 일부 조사의 통계분류 자료처리 과정에 AI 분류 예측결과를 시험적용하는 단계에 이르렀다.

본 연구에서는 지금까지 통계개발원에서 진행된 데이터과학 연구 추진경과를 간략히 살펴본 후 인공지능 기반 통계분류 방법론 및 인공지능 통계분류 자동화 시스템 구축 결과를 소개하도록 하겠다. 그리고 건설업조사에서의 공종·발주자분류 사례를 중심으로 AI 통계분류 적용 결과를 분석하고 향후 통계분류 자료처리 과정에서 AI 통계분류 예측결과를 더욱 효과적으로 활용하기 위한 실무활용성 제고 방안을 모색해 보고자 한다.

---

1) 현재 국가통계 생산은 지속적인 조사환경 변화의 시대에 직면하고 있다. 과학적인 조사 설계의 발전, 데이터 수집 및 분석기술의 고도화, 인터넷 및 웹 기반 조사환경 확산 등은 긍정적 변화라고 볼 수 있다. 반면 사생활 보호의식 확산, 1인가구 증가 등으로 인한 응답률 감소와 조사비용 증가에 따른 조사환경 악화에 대한 문제점도 동시에 발생하고 있다.

2) 예를 들어 미국 센서스국(Census Bureau)에서는 기계학습 알고리즘을 경제총조사 북미산업분류코드 코딩작업에 활용하고 있으며, 노동통계국(BLS)에서는 업무상 상해·질병분류코드 코딩작업에 AI 기법을 적용하고 있다.

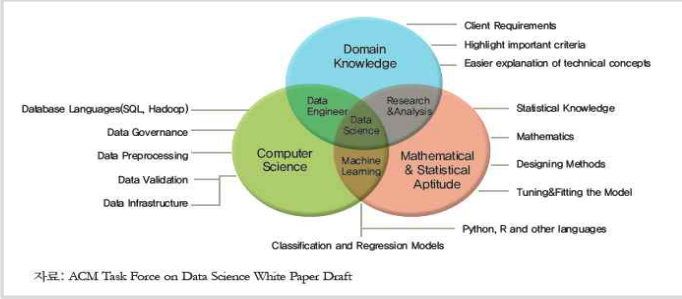
3) AI 등 첨단 ICT 기술을 공공분야에 선도적으로 적용하기 위해 중앙행정기관 및 지자체, 공공기관을 대상으로 사업공모를 통해 개발예산을 지원하는 과학기술정보통신부「디지털 공공서비스 혁신 프로젝트」'22년 지원사업으로 선정되어 개발예산(15.4억)을 확보하여 구축하였다.

제 2 장

데이터과학 연구 추진경과

제1절 데이터과학 연구팀 신설

데이터과학(Data Science)은 대규모 데이터를 소프트웨어와 연계하여 정보를 신속·정확하게 창출하고 활용하는 것을 연구하는 융합 학문이라고 볼 수 있다.[4)]



<그림 2-1> 데이터 사이언스의 구조

데이터과학의 융합 과학(convergence science)적 특성으로 인해 IT 분야, 분석적 영역, 비즈니스 컨설팅 영역을 포괄하는 다양한 주제들이 광의의 측면에서의 데이터과학 방법론으로 논의되고 있으나, 최근에는 협의의 측면에서 인공지능(AI)을 중심으로 한 데이터과학 방법론 논의가 집중되고 있는 추세이다.

한편 4차 산업혁명 시대의 도래를 맞아 국가통계 분야에도 데이터과학 기법을 적용해야 한다는 요청이 증대하고 있으며, 이에 통계개발원에서는 미래 통계환경 변화 대응을 위한 데이터과학 연구를 2020년부터 추진하게 되었다.[5)]

---

4) 데이터과학이라는 용어는 1974년 코펜하겐 대학의 전산학교수 피터 나우어(Peter Naur)가 <Concise Survey of Computer Method> (Studentlitteratur, Sweden)에서 처음으로 사용했다고 하며, 2002년 학회지로 <Data Science Jounal>이 발행되면서 보편화되기 시작하였다.

- 「4차 산업혁명 시대의 데이터 사이언스의 역할과 미래비전」(박성현, 2018.06.14. 미래한국 기고문) 中

제2절 데이터과학 기법 국가통계 활용 해외사례 검토

국가통계에 AI 기술을 도입하는 사례는 빅데이터에 기반해 '신규통계를 생산'하는 방식과 전통적인 조사통계의 효율성을 향상시키기 위한 '통계생산 프로세스 현대화' 두 가지 범주로 나누어 살펴볼 수 있는데, AI 기반 신규통계 생산은 기존 조사통계와 비교하여 신뢰성 및 설명 가능성 확보 측면에서의 한계를 보여 국가통계에서의 AI 기술 적용은 통계생산 프로세스 현대화 측면에서 주로 논의되고 있다.6)

그리고 인공지능(AI) 기술을 국가통계에 활용하고 있는 해외사례 대다수는 '비정형 텍스트 분석을 통한 분류 자동화'에 초점을 맞추고 있다.

<표 2-1> 비정형 텍스트 분석을 통한 분류 자동화 적용 사례

국가	활용 내용
미국	- (BLS) 업무상 상해 및 질병 설문조사 응답결과 자동 코드 분류
	- (Census) 미국 지역사회 설문조사의 산업 및 직업에 대한 코드 분류
독일	- 국내 법인에 대한 유럽국민계정 체계상 기관 분류 할당
	- 온라인 구인 공고 분류
캐나다	- 월별 소매거래 및 분기별 소매상품 조사에 스캔 데이터 활용
핀란드	- 경찰청 교통사고 보고 문건에 기반해 사고 내역 자동 분류
벨기에	- 구인 공고별 알맞은 경제 활동 통계 분류 코드 예측
룩셈부르크	- 민간 기업의 스캐너 데이터를 활용해 상품 분류 코드 추천
스위스	- 항공 이미지에 기반한 토지 피복 및 토지 사용 코드 분류 자동화

한편 행정자료 등이 보편적으로 조사통계에 활용됨에 따라 '다출처 자료의 연계 및 보완' 과정에 AI 기술을 활용하기 위한 다양한 시도도 진행되고 있다.

<표 2-2> 다출처 자료의 연계 및 보완 적용 사례

국가	활용 내용
독일	- 데이터 연계를 통한 연방 고용청의 패널 데이터 보강
캐나다	- 인구 조사를 구성하는 이민 허가 변수에 대한 결측 값 대체
네덜란드	- 보건 실태조사를 대상으로 한 혼합형 설문조사의 무응답 대체
오스트리아	- 행정 데이터와 SILC 융합을 통한 ICT 조사의 소득 문항 대체

5) 2020년 2월 통계방법연구실에 데이터과학 연구팀 신설(5급 1명, 6급 1명)

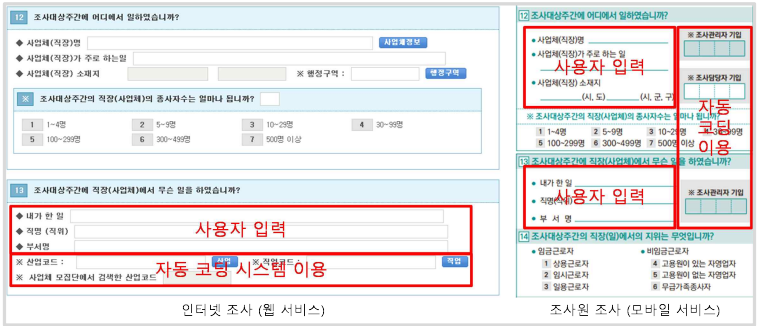
6) 이 절의 내용은 김정민 외(2020), "AI 기술의 국가통계 활용 사례 및 국내 도입 촉진 방안"에서 국가별로 제시된 AI·빅데이터 통계 주요 도입 사례를 범주별로 나누어 재구성하였다.

제3절 인공지능 통계분류 시험분석

앞서 살펴본 AI 기술 국가통계 적용 해외사례 검토 결과를 토대로, 통계방법연구실에서는 2020년~2021년 2년에 걸쳐 통계조사의 통계분류 관련 입력 텍스트를 '자연어처리' 및 '지도학습 기반 머신러닝' 기술을 이용하여 분석하고 분류 코딩을 자동화하는 기초연구를 수행하였다.7)

1. AI 통계분류 가능성 검토(2020)

통계청 지역별고용조사 2회차 약 520,000건의 산업분류 문항 텍스트 자료에 대한 AI 통계분류 예측 정확도에 대한 시험분석을 진행하였다.

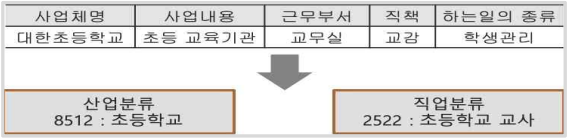


<그림 2-2> 지역별고용조사 산업·직업분류 조사화면 예시8)

지역별고용조사에서는 조사에 응한 16세 이상 경제활동인구가 근무하는 '사업체명, 사업내용, 근무부서, 직책, 하는 일의 종류'에 대한 텍스트 입력을 기초로 그에 해당하는 산업·직업 분류코드를 내검과정을 거쳐 확정하게 된다.

7) 이 절은 임경민과 김혜란(2020)「데이터과학 기법의 국가통계 시험적용 분석」통계개발원 연구보고서 및 오교중(2021)「AI 기반 통계분류 자동코딩시스템 연구」통계개발원 연구용역 보고서 내용을 요약하였다.

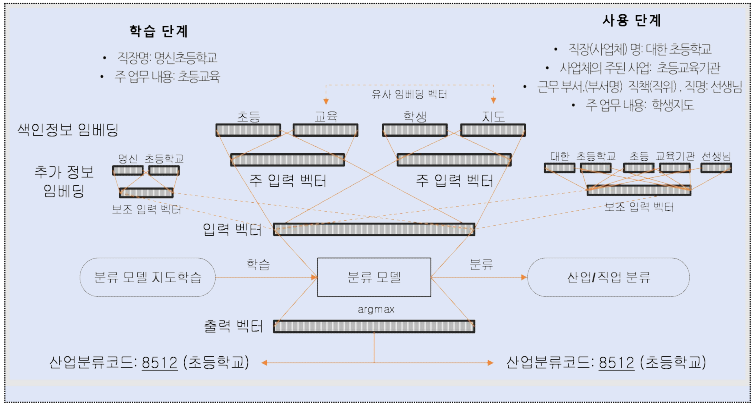
8) 통계청 지역별고용조사 - 나라통계(http://www.narastat.kr/eregion) 입력시스템 화면



<그림 2-3> 지역별고용조사 산업직업분류 입력과 코딩 예시

분석에 사용된 지역별고용조사 2회차 약 520,000건 자료 중 한 회차 자료는 학습용으로 다른 회차 자료는 평가용 데이터로 사용하였으며, 최종 공표결과 분류코드를 정답으로 가정하고 학습 및 평가를 수행하였다.9)

자연어처리 및 분류 기본 모델로는 아래 그림과 같은 단문 분류를 목적으로 개발된 CBoWFCNN(Contiunous Bag-of-Word Fully Connected Neural Network) 모델을 사용하였다.



<그림 2-4> 지도학습 기반 분류 학습을 위한 초기 기준 모델

지도학습 기반 머신러닝 방법론을 적용하여 시산한 결과 대분류 수준에서는 95.9%, 중분류 수준에서는 93.8%, 소분류 수준에서는 91.4%, 세분류 수준에서는 88.9%의 분류 예측 정확도 결과가 도출되었다. 제한된 일부 자료에 대한 분석이지만 비교적 높은 수준의 분류 정확도와 빠른 수행속도를 보여 향후 AI 통계분류 자동화의 가능성을 확인할 수 있었다.

9) 학습 및 평가용으로 사용한 총 526,391개 데이터에서 띄어쓰기와 중복을 제거한 115,081개의 텍스트를 분석하였고, 일부 oo, ** 등 비식별 데이터는 ##와 같은 글자로 통일하여 사용하였다.

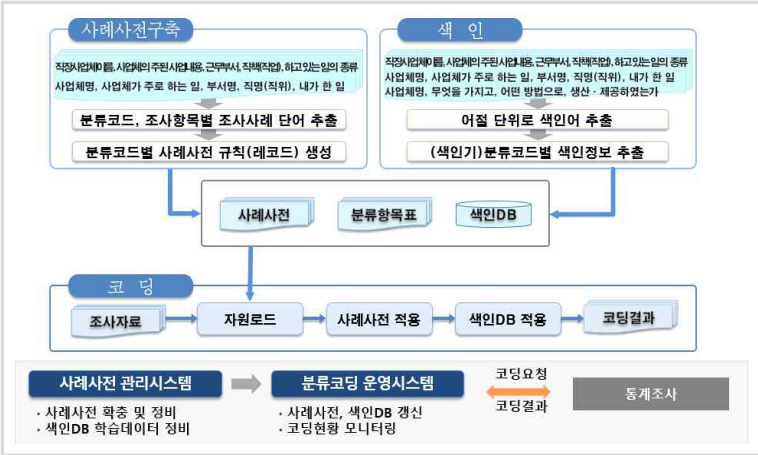
<표 2-3> 머신러닝 기반 산업분류 예측 성능 평가

모델	분류	대분류 (21)	중분류 (77)	소분류 (232)	세분류 (495)
기본 모델 (FastTest + FCNN, epoch=200, lr=0.05)	정확도(F1)	0.959	0.938	0.914	0.889
	학습시간(s)	67.0	143.0	271.2	393.0
	분류수행속도(iter/s)	79691.9	81395.6	39805.3	21773.8

2. 기존 시스템과의 성능 비교(2021)

2020년 기초연구를 통해 AI 통계분류 가능성을 확인한 후 2021년에는 기존 통계분류 시스템과 AI 통계분류를 비교하는 연구를 수행하였다.

일반적으로 통계청의 통계분류 관련 조사항목의 텍스트 입력값을 토대로 분류코드를 확정하는 자료처리 과정은 ① 사례사전(단어들의 조합으로 이루어진 규칙체계)에 의한 분류코드 자동코딩과 ② 색인DB를 활용하여 색인어 빈도 정보로 추천 분류코드를 제공하는 분류시스템을 참고하여 ③ 최종적으로 내검원들이 분류코드를 확정하는 방식으로 이루어지고 있다.



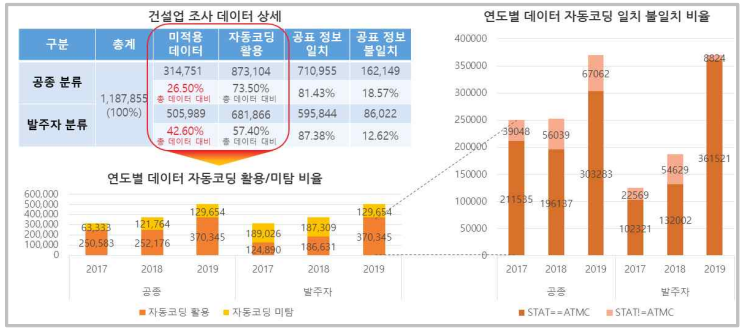
<그림 2-5> 통계청 통계분류 자료처리 시스템 개요10)

10) 통계청「산업·직업분류 코딩시스템 & 사례사전관리시스템 업무설명자료」中

사례사전에 의한 자동코딩은 규칙 베이스 기반 분류모델로 볼 수 있는데, 비교적 매우 높은 정확도에도 불구하고 낮은 적용범위(coverage)를 가진다는 한계를 지니고 있으며, 정확도 제고 및 적용범위 확대를 위한 규칙 정비 및 관리에 많은 노력이 지속적으로 필요하다.[11)]

그에 따라 많은 양의 데이터는 색인DB 추천정보를 참고하여 내검원의 수작업을 거쳐 분류코드를 정하게 되는데 이에 많은 시간과 자원이 소요되고 있는 상황이다.

2017~2019년 건설업조사의 공종·발주자분류 일부 자료 약 1,200,000건의 데이터에 대한 분석결과 사례사전에 의한 자동코딩 정확도는 80% 중반 수준이고 커버리지는 대략 60% 중반 수준을 보이고 있다.



<그림 2-6> 건설업조사 사례사전 자동코딩 활용 현황

이 자료에 대해 XLM-RoBERTa를 사전학습 언어모델로 사용하여 기계학습 기반 분류예측을 수행한 결과는 아래와 같다.

전체 1,187,855건 자료에 대한 예측정확도는 88%이며, 사례사전 자동코딩 결과가 최종 공표결과와 같은 710,955건에 대해서는 95%의 정확도를 그리고 사례사전 결과가 없는 314,751건에 대해서는 82%의 분류예측 정확도를 보였다.

즉, AI 분류예측이 기존 규칙기반 사례사전 자동코딩에 비해 정확도 및 커버리지에서 더 우월할 결과를 보임을 확인할 수 있었다.

11) 가구부문 산업분류의 경우 세분류(495개) 기준 약 34,000개의 사례사전 규칙체계를 관리하고 있음에도 불구하고, 2015년 인총의 경우 사례사전에 의한 자동코딩 커버리지는 50%에 불과하다.

<표 2-4> 건설업조사 공종·발주자분류 AI 분류예측 성능 평가

평가 데이터	분류	정밀도 (Prec.)	재현율 (Recall)	F-1 Score	정확도 Accuracy	비고
전수 데이터 (전체)	공종	0.88	0.88	0.88	0.88	전체적인 정확도 성능
	발주자	0.88	0.88	0.88	0.88	
공표자료 일치	공종	0.96	0.95	0.96	0.95	자동코딩 시스템과 유사성
	발주자	0.95	0.95	0.95	0.95	
자동코딩 미적용	공종	0.81	0.82	0.81	0.82	내검 작업 감소 정도
	발주자	0.83	0.82	0.82	0.82	
내검 및 수정	공종	0.67	0.62	0.63	0.62	오답률 개선 정도
	발주자	0.77	0.74	0.73	0.74	

제4절 인공지능 통계분류 자동화 시스템 개발(2022)

2020년~2021년의 기초연구를 통해 AI 기반 통계분류 자동화의 가능성을 확인한 후 통계청은 추가 후속연구 및 실무활용을 위해 2022년 AI 통계분류 자동화 시스템 개발사업을 추진하였다.[12)]

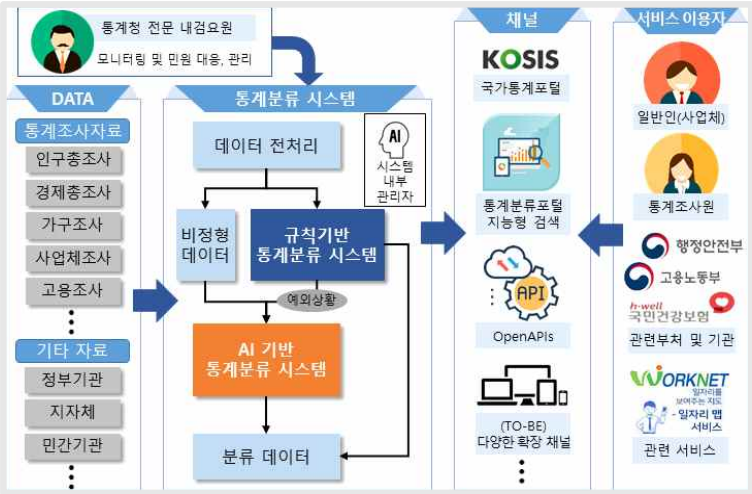
해외사례 및 선행연구를 기초로 "인공지능 기반 통계분류 자동화 시스템 구축"이 '22년 과학기술정보통신부「디지털 공공서비스 혁신 프로젝트」지원 사업[13)]으로 선정되어 개발예산(15.4억)을 확보하였고, '22.7.~'23.1.(6개월) 사업기간을 거쳐 기계학습기반 범용 통계분류 자동화 시스템을 개발하였다.

AI 통계분류 자동화 시스템은 대규모 데이터를 처리하는 5개 조사(인주주택총조사, 전국사업체조사, 지역별고용조사, 건설업조사, 온라인쇼핑조사)의 5종 분류(산업·직업분류, 공종·발주자분류, 상품군별분류)를 대상으로 개발되었다.[14)]

12) 대용량 학습데이터를 활용하는 AI 기술의 특성으로 인해 이를 처리하기 위한 대규모 컴퓨팅 자원이 필수적인 상황에서, 공공기관에서 수행하는 대부분의 AI 관련 연구들은 제한된 일부 데이터에 대한 분석이나 외부자원을 활용한 연구용역의 형태로 진행할 수밖에 없는 한계가 있었다.

13) AI 등 첨단 ICT 기술을 공공분야에 선도적으로 적용하기 위해 중앙행정기관 및 지자체, 공공기관을 대상으로 사업공모를 통해 총 11개 과제를 선정하여 197억 원을 지원('22년 기준)

14) 시스템 개발범위: 산업/직업분류[인총/경총/지역별고용], 공종/발주자분류[건설업조사], 상품군별분류[온라인쇼핑조사]



<그림 2-7> 목표시스템 개념도[15)]

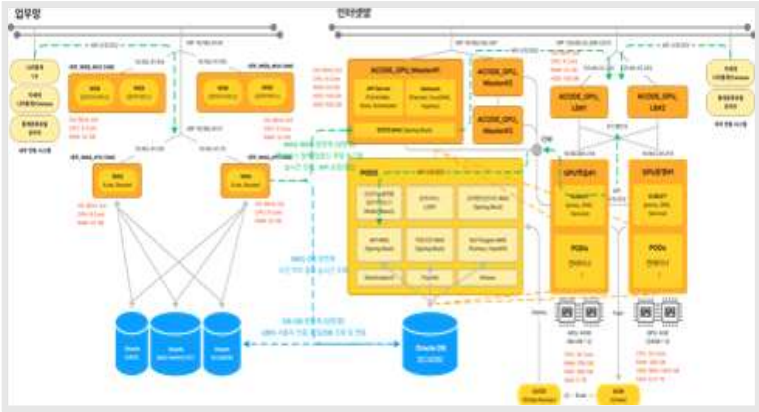
시스템에 적용된 사전학습 언어모델은 RoBERTa[Yinhan Liu, et al, 2019] 모델을 사용[16)]하였고, 계층형 통계분류 구조를 고려하여 지도학습 방법론으로는 Hierarchical Text Classification(Hi-TC) 모델을 적용하였다(상세내용은 3장 참고).

시스템 구성은 나라통계시스템 등 조사시스템의 자료를 DB Link를 통해 제공받아 별도 독립 서버로 구성된 AI 통계분류 시스템에서 데이터학습 및 분류예측 작업을 수행하고 그 결과를 다시 조사시스템에 API 방식으로 서비스하는 형태로 개발하였다. 학습 및 서비스를 위한 HW는 신규 도입한 두 개의 GPU서버(A100과 A30)[17)]와 국가정보자원관리원 통계청 클라우드 환경의 WEB, WAS 장비를 활용하였다.

15) 통계청 '인공지능 기반 통계분류 자동화 시스템 구축' 사업계획서 中

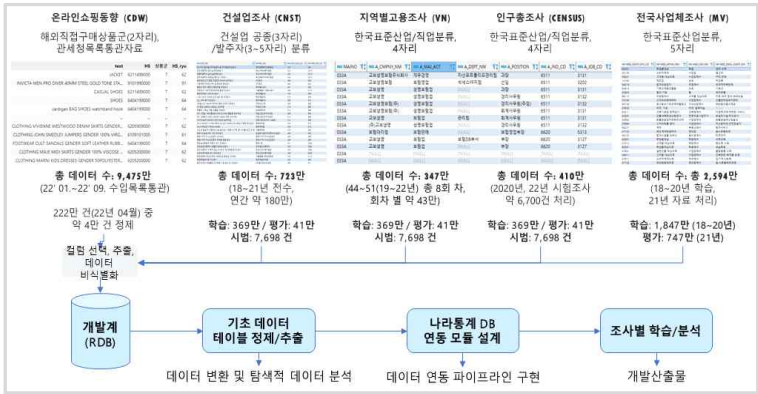
16) KcBERT-Large를 기반으로 Klue-RoBERTa-Large와 xlm-RoBERTa-Large 모델을 활용함

17) 모델학습은 A100(GPU:80GB*2, CPU:16Core, RAM:256GB, HDD:960GB, SAN:5TB)에서, 서비스 운영은 A30(GPU:24GB*2, CPU:24Core, RAM:384GB, HDD:960+1920GB, SAN: 5+5TB)에서 주로 수행하는 방식으로 설정하였다.



<그림 2-8> AI 통계분류 자동화 시스템 구성도[18)]

AI 분류예측을 위해 5개 조사의 5종 분류 관련 총 1억 6,700만 건의 텍스트를 중복제거 등의 전처리 정제작업을 거쳐 학습데이터로 사용하였다.



<그림 2-9> AI 통계분류를 위한 학습데이터 구성

자연어처리 및 지도학습 방식의 학습을 거쳐 평가데이터셋을 대상으로 AI 분류예측 정확도를 평가한 결과는 다음과 같다.

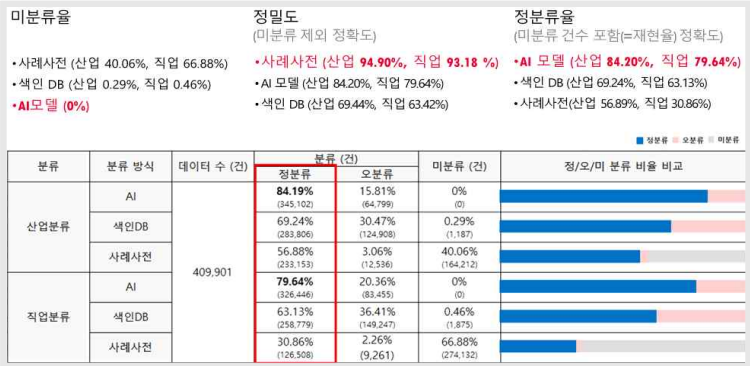
18) 통계청 '인공지능 기반 통계분류 자동화 시스템 구축사업' 완료보고회 발표자료 中

<표 2-5> 분류모델 구축(5종 분류, 5개 조사) 및 성능평가 결과

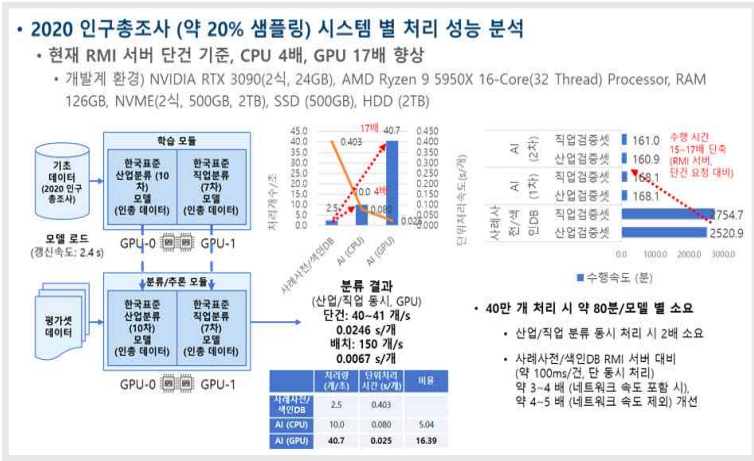
분류 구분	관련 부서	적용조사 (주기)	데이터양 (추이/예측)	기초 데이터 이수 건수	데이터 전처리	데이터 정제	데이터 분리	학습/평가 품질 (Accuracy / F-1)	구축횟수 (최종학습일)
한국표준 산업 (10차)	고용 통계과	지역별 고용조사 (연 2회)	약 43만 (6개월 간격)	총 347만 (44~51회 총 8회차 19~22 4년 자료)	결측: 109만 중복: 32만	122만	학습: 110만 검증: 12만 평가: 43만	검증: **92.74% / 0.9271** 평가: **94.17% / 0.9414**	4회 (12/14)
	경제 총조사과	전국사업체 조사 (연 1회)	약 800만 (1년 간격)	총 2,594만 (18~21년)	결측: 484만 중복: 553만	834만	학습: 751만 검증: 83만 평가: 723만	검증: **85.04% / 0.8508** 평가: **86.48% / 0.8623**	9회 (01/11)
	인구 총조사과	인구총조사 (5년)	약 1,000만 (5년 간격, 전가구원 종합)	총 981만 (2020 인구총조사 기초데이터)	처리: 571만 중복: 40만 결측: 65 개	410만	학습: 369만 검증: 41만	검증: **86.01% / 0.8610** 평가: **85.60% / 0.8551**	6회 (11/06)
한국표준 직업 (7차)	고용 통계과	지역별 고용조사 (연 2회)	약 43만 (6개월 간격)	총 347만 (44~51회 총 8회차 19~22' 4개년 자료)	결측: 109만 중복: 32만	122만	학습: 110만 검증: 12만 평가: 43만	검증: **91.44% / 0.9141** 평가: **92.67% / 0.9259**	5회 (12/14)
	인구 총조사과	인구총조사 (5년)	약 1,000만 (5년 간격, 전가구원 종합)	총 981만 (2020 인구총조사 기초데이터)	처리: 571만 중복: 40만 결측: 65	410만	학습: 369만 검증: 41만	검증: **79.52% / 0.7915** 평가: **79.25% / 0.7896**	6회 (11/06)
건설업 공종	산업 통계과	건설업 조사 (연 1회)	평균 180만 (1년 간격)	총 883만 (17~21년 전수)	중복: 250만	633만	학습: 192.1만 검증: 21.3만 평가: 92.1만 (21년, 내검 대상)	검증: **84.22% / 0.8411** 평가: **85.03% / 0.8482**	8회 (01/04)
건설업 발주자	산업 통계과	건설업 조사 (연 1회)	평균 180만 (1년 간격)	총 883만 (17~21년 전수)	중복: 294만 결측: 5만	589만	학습: 192.6만 검증: 21.4만 평가: 92.2만 (21년, 내검 대상 Y)	검증: **91.81% / 0.9177** 평가: **89.32% / 0.8937**	6회 (01/04)
물품통관 분류	서비스업 동향과	온라인쇼핑 동향 (월 1회)	평균 230만 (1월 간격)	총 9,730만 (21 ~ 22년 2년 자료)	중복: 8,315만 (21' 01.~22' 08.)	상품군: 93만 (22년 04월) HS코드: 1,182만 (21' 01.~22' 08.)	학습: 26만 검증: 2.9만 평가: 9.3만 (상품군, 22년 04월)	검증: **96.54% / 0.9653** 평가: **96.86% / 0.9685**	4회 (12/02)

규칙베이스 기반 사례사전 및 색인DB를 활용하는 기존 시스템과의 비교 분석 결과 AI 통계분류 자동화시스템이 분류예측 정확도 및 커버리지 측면에서 더 우월한 성능을 보임을 확인할 수 있었다.

<표 2-6> 기존 시스템과의 분류예측 성능비교 ('20년 인총데이터 분석결과)



또한 텍스트 분석에 특화된 최신 GPU서버 활용 등으로 인해 AI 통계분류 자동화 시스템 처리속도 또한 기존 시스템 대비 대폭 향상되었다.



<그림 2-10> 기존 시스템과의 처리속도 성능비교 ('20년 인총데이터 분석결과)

제5절 인공지능 통계분류 결과 시험적용 및 고도화 검토(2023)

2020년~2021년	2022년	2023년	2024년~
[기초연구]	**[시스템 구축]**	**[시스템 시험 적용]**	**[시스템 확산]**
AI 통계분류 가능성 분석 (지고,건설업조사)	AI 통계분류 자동화 시스템 개발 (과기부 지원 15.4억)	AI 통계분류 실무적용 (5대분류, 5종조사) AI 시스템 고도화 검토	AI 분류결과 적용대상 확대 (나라시스템 적용)

2023년부터는 5개 조사(인주주택총조사, 전국사업체조사, 지역별고용조사, 건설업조사, 온라인쇼핑조사)의 5종 분류(산업·직업분류, 공종·발주자분류, 상품군별분류)에 AI 통계분류 예측결과를 통계청 내검과정에 시범 적용하여 실무에 활용하고 있다.

통계분류 관련 텍스트 자료에 해당하는 AI 통계분류 예측코드와 함께 제공된 예측 정확도에 대한 확률정보를 토대로 내검원의 내용검토 우선순위를 결정하는 방식의 선택적 에디팅을 통해 자료처리의 정확성 및 효율성 제고를 도모하였다.[19]

19) 또한 GPU서버 활용에 따른 시스템 속도 향상으로 대규모 텍스트자료의 처리속도가 대폭 단축되었는데, 특히 월별 500만 건 이상의 데이터(관세청 목록통관자료)를 처리해야 되는 온라인쇼핑조사 상

한편 2023년에는 5개 조사 5종 분류를 대상으로 개발된 AI 통계분류 자동화시스템 적용대상 확대를 위한 기초연구를 진행하고 있다. 가계동향조사의 소비목적별 지출분류코드, 생활시간조사의 행동분류코드, 사망원인통계의 질병사인코드 등 3개 조사 3종 분류에 대한 AI 통계분류 자동화 가능성에 대한 시험분석을 추진하였다.[20]

<그림 2-11> AI 통계분류 자동화 확대 적용 검토 대상 조사

품군별분류에서는 (분류예측 정확도 이외에) 처리속도 문제가 중요하게 검토되었다.

20) 분석결과는 오교중(2023)「AI 통계분류 자동화 확대 적용」통계개발원 연구용역 보고서 참고

제 3 장

인공지능 통계분류 결과분석[21]

제1절 건설업조사 공종·발주자분류 개요[22]

건설업조사는 우리나라 건설업부문의 종사자수, 급여액, 매출액, 부가가치, 공사실적 등에 관한 사항을 조사하고, 건설업부문의 구조 및 활동실태를 파악하여 정책수립에 필요한 기초자료를 제공하기 위한 목적으로 1973년부터 실시되어 온 조사이다(2022년 기준 건설업조사: 제50회 조사).

건설업체에서는 응답자 기입방식으로 각 협회 시스템을 통해 조사 대상 연도의 건설경영실태 및 건설공사실적 등의 자료를 입력하고 통계청은 건설 관련 협회 대상 인터넷조사를 통해 전수조사 방식으로 자료를 입수하고 있다.[23]

<표 3-1> 건설업조사 공종·발주자분류 조사문항 예시

(2) 일련번호	(3) 공사명	(4) 공사유형	(5) ※ 공 종 세분류부호	(5-1) 공사 지역명	(5-2)※ 공사지역 부 호	(6) 발주자명 / (7) 하도급자는 아래 칸 괄호 안에 원도급자명을 기입	(8) 도급 종류 1. 원 도 급 2. 하 도 급 3. 자기공사	(9) ※ 발주자 분 류 부 호
				시도 / 시군구 / 동읍면		( )		

연간 약 200만 건의 건설공사 실적자료가 입수되며 (3)공사명 자료를 기초로 공종세분류 코드가 결정되고 (6)발주자명 자료를 기초로 발주자분류 코드가 결정되며 해당 분류코드의 적정성을 통계청 내검원들이 확인하는 내용검토 과정을 거치게 된다.

21) 이 장에서는 AI 통계분류 자동화 시스템을 활용한 5개 조사 5종 분류 중 건설업조사의 공종 및 발주자분류를 중심으로 AI 통계분류 분석내용을 살펴보도록 하겠다.

22) 통계청「건설업조사 지침서」내용을 정리함

23) 건설업조사는 대한건설협회, 대한전문건설협회, 대한기계설비건설협회, 대한시설물유지관리협회, 한국전기공사협회, 한국정보통신공사협회 및 한국소방시설협회, 건설산업정보원에서 관련 회원사 및 등록업체를 대상으로 조사한 내용을 통계청에 송부하는 방식으로 조사가 이루어지고 있다.

<표 3-2> 건설업조사 공종분류코드(64종) 및 발주자분류코드(96종) 예시

토목:
210: 일반도로
211: 고속도로
212: 고속화도로
220: 도로교량
221: 철도교량
230: 댐
240: 간척
250: 항만
251: 공항
260: 도로터널
261: 철도터널
262: 기타터널
270: 일반철도
271: 고속철도
272: 지하철
280: 택지조성
281: 공업용지 조성
282: 치수·하천
283: 치산·사방
290: 상수도 1천mm이상
291: 상수도 1천mm미만
292: 하수도
293: 관개수로
294: 정수장
295: 농지정리
299: 기타 토목시설

건축:
410: 단독주택 및 연립주택
411: 저층아파트(5층이하)
412: 고층아파트(6-15층이하)
413: 초고층아파트(16층이상)
414: 주거·상업용 겸용건물
420: 시장·백화점·쇼핑센터
421: 사무실빌딩
422: 오피스텔
423: 인텔리전트빌딩
430: 관공서건물(11층이하)
431: 관공서건물(12층이상)
432: 호텔·숙박시설
433: 학교
434: 병원
440: 교회 등 종교용건물
441: 전통양식건축
442: 기타문화재·유적건물
450: 공연·집회장소
451: 경기장·운동장
452: 전시시설
460: 공장·작업장용건물
461: 기계기구설치(플랜트제외)
462: 변·발전소용건물
470: 창고·차고·터미널건물
480: 위험물저장소
490: 기타건축시설

산업설비:
511: 하수종말처리장
512: 폐수종말처리장
513: 쓰레기소각장
514: 기타 환경시설공사
520: 발전소설비공사
530: 송유관 및 가스관
540: 유류 및 가스저장시설
550: 제철소 등 산업생산시설
560: 기타 플랜트설비공사

조경공사:
610: 수목원
611: 공원조성공사
620: 기타조경시설

중앙정부기관:
04100: 기획재정부
04200: 조 달 청
04300: 행정안전부
04400: 국 방 부
04500: 교 육 부
04600: 문화체육관광부
04700: 농림축산식품부
04800: 산업통상자원부
04900: 환 경 부
05000: 보건복지부
05100: 노 동 부
05200: 국토교통부
05201: 국토관리청(서울)
05202: 국토관리청(부산)
05203: 국토관리청(원주)
05204: 국토관리청(대전)
05205: 국토관리청(익산)
05400: 법 무 부
05500: 외 교 부
05600: 통 일 부
05700: 여성가족부
06000: 과학기술정보통신부
06100: 해양수산부
06200: 중소벤처기업부
08900: 기 타

지방자치단체:
0110: 서 울 시
0210: 부 산 시
0220: 대 구 시
0230: 인 천 시
0240: 광 주 시
0250: 대 전 시
0260: 울 산 시
0290: 세 종 시
0310: 경 기 도
0320: 강 원 도
0330: 충청북도
0340: 충청남도
0350: 전라북도
0360: 전라남도
0370: 경상북도
0380: 경상남도
0390: 제 주 도
0400: 기타/북한
0410: 중동
0420: 아시아
0430: 유럽
0440: 아프리카
0450: 오세아니아
0460: 아메리카

기타:
091: 공 공 단 체
092: 공기업
093: 주한외국기관
094: 해외

민간부분:
제조업(25개):
900: 식료품 제조업
901: 음료 제조업
902: 담배 제조업
903: 섬유제품 제조업(의복제외)
904: 의복,의복액세서리및모피제품
905: 가죽, 가방 및 신발 제조업
906: 목재 및 나무제품(가구제외)
907: 펄프, 종이 및 종이제품
908: 인쇄 및 기록매체 복제업
909: 코크스, 연탄 및 석유정제품
910: 화학물질·화학제품(의약품제외)
911: 의료용 물질 및 의약품 제조업
912: 고무 및 플라스틱제품
913: 비금속 광물제품 제조업
914: 1차 금속 제조업
915: 금속가공제품(기계 및 가구제외)
916: 전자부품, 컴퓨터 영상, 음향 및 통신장비 제조업
917: 의료,정밀,광학기기 및 시계
918: 전기장비 제조업
919: 기타 기계 및 장비 제조업
920: 자동차 및 트레일러 제조업
921: 기타 운송장비 제조업
922: 가구 제조업
923: 기타 제품 제조업
924: 산업용기계 및 장비수리업

비제조업(18개):
950: 농업, 임업 및 어업
951: 광업
952: 전기,가스,증기 및 공기조절공급업
953: 수도, 하수 및 폐기물처리, 원료재생업
954: 건설업
955: 도매 및 소매업
956: 운수 및 창고업
957: 숙박 및 음식점업
958: 정보통신업
959: 금융 및 보험업
960: 부동산업
961: 전문, 과학 및 기술서비스업
962: 사업시설관리, 사업지원 및 임대서비스업
963: 교육서비스업(국·공립 제외)
964: 보건업 및 사회복지서비스업
965: 예술, 스포츠 및 여가관련 서비스업
966: 협회 및 단체, 수리 및 기타 개인서비스업
999: 기타

통계청의 다른 통계분류 자료처리 과정과 동일하게 건설업조사 공종·발주자 분류에서도 ① 사례사전(단어들의 조합으로 이루어진 규칙체계)에 의한 분류코드 자동코딩과 ② 색인DB를 활용하여 색인어 빈도 정보로 추천 분류코드를 제공하는 분류시스템을 참고하여 ③ 최종적으로 내검원들이 분류코드를 확정하는 과정을 거치고 있다.

64종의 공종분류에 대해서는 128,336개의 사례사전과 48,969개의 고유색인어를 96종의 발주자분류에 대해서는 40,652개의 사례사전과 82,365개의 고유색인어를 활용하고 있으며, 사례사전에 의한 자동코딩 정확도는 80% 중반 수준이고 커버리지는 대략 60% 중반 수준을 보이고 있다.[24)]

<표 3-3> 건설업조사 공종·발주자분류 데이터 예시

	조사연도 (2017~2019)	발주자명	공사명	공종 자동코딩 코드	공종 공표 코드	발주자 자동코딩 코드	발주자 공표 코드
	BASE_YEAR	PPO_NM	CNST_NM	ATMC_STAT_KDCS_CD	STAT_KDCS_CD	ATMC_STAT_PPO_CD	STAT_PPO_CD
0	2019	롯데칠성	롯데칠성인천 1,2공장 PSM관련 전기공사	460	460	901	901
1	2019	인산시청	2019년 상록구 구역 교차로 개선 전기공사	210	210	0310	0310
2	2019	강원도교육청 강원도동해교육지원청	북평고 별관동 전원교체 전기공사	433	433	0320	0320
3	2019	경기도교육청 경기도화성오산교육지원청	기산초 화장실 개선 전기공사	433	433	0310	0310
4	2019	현대위어(주)	안산,평택공장 변전실 보수공사	460	460	920	920
...	...	...	...	...	...	...	...
1187850	2018	신젠타코리아(주)	2018년도 소방시설보수공사(작동)	...	550	910	910
1187851	2018	전주대학교	연구실 정밀안전점검에 실시에 따른 보완공사(소방)		433	963	963
1187852	2018	미원스페셜티케미칼(주)	50m3보수공사 중 소방공사		550	910	910
1187853	2018	경북대학교병원	경북대학교 소방공사		434	091	091
1187854	2018	신한농협하나로마트수양점	신한농협 하나로마트 수양지점 작동기능점검 보수공사		420	918	955

1187855 rows × 7 columns

기존 시스템 미적용 데이터

종류	분류개수 (종)	분류 코드 구성	사례사전/색인DB 구성	비고
공종분류	64	• 연속되지 않은 3자리 코드	• 규칙/조건: 128,336개 • 48,969개 고유색인어	• 산업/직업 분류와 비슷한 색인어 구성 • 발주자에 분류 코드가 결정되는 경우가 있음
발주자분류	96	• 중앙정부기관(25): 5자리 • 지방자치단체(24): 4자리 • 기타(4): 0으로 시작하는 3자리 • 민간부분(43): 9로 시작하는 3자리 • 제조업(25), 비제조업(18)	• 규칙/조건: 40,652개 • 82,365개 고유색인어 (4,565 색인어 중복)	• 지역 정보가 추가 자질로 적용 될 수 있음

사례사전의 낮은 커버리지(위의 <표 3-3>에서 '기존 시스템 미적용 데이터' 부분)와 분류 예측 정확도 저하로 인해 많은 내검원들이 장기간에 걸쳐 분류코드를 확인하는 작업을 수행하고 있는데, 경제성장에 따라 건설공사 데이터 양이 증가함에 따라 내검업무량 또한 지속적으로 증대되고 있다.

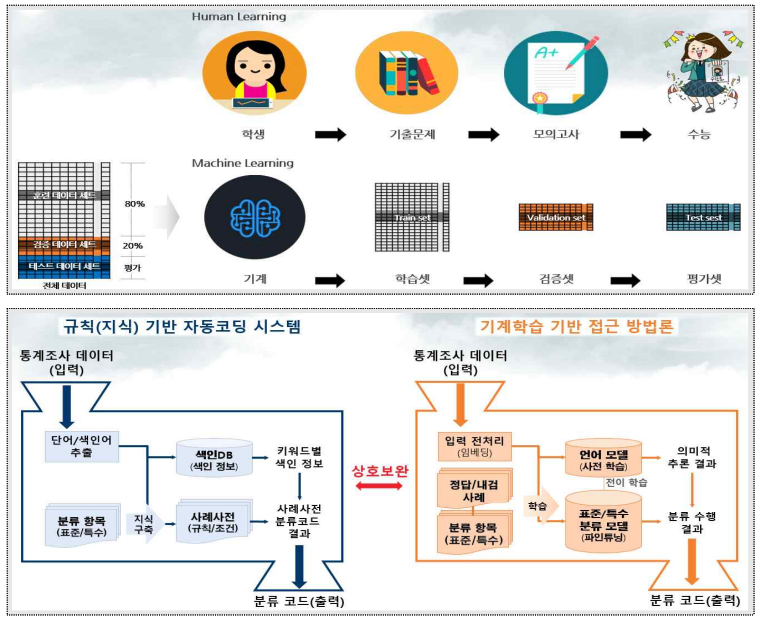
이를 해결하기 위해서는 사례사전과 색인DB의 커버리지와 정확도를 높이기 위한 규칙정비가 필수적인데 이 또한 업무담당자의 많은 노력을 요하는 상황이다.

24) 2017~2019년 건설업조사 일부 자료 1,187,855건에 대한 분석(2021) 결과

제2절 인공지능 통계분류 방법론

앞서 살펴본 바와 같이 기존 통계청이 운영하는 텍스트 분류 자료처리 과정은 규칙 기반 사례사전에 의한 자동코딩과 색인DB 정보를 활용하여 최종적으로 사람(내검원)이 최종 분류코드를 확정하는 방식으로 진행된다. 이 방법론은 분류 항목이 적고 입력 텍스트가 균질적인 경우에는 효율적일 수 있지만, 데이터 양이 많아지고 분류 항목의 다양성이 증가하는 경우 규칙과 조건 설정이 어려워지고 유지 관리에 많은 시간과 비용이 발생하는 한계를 지닌다.

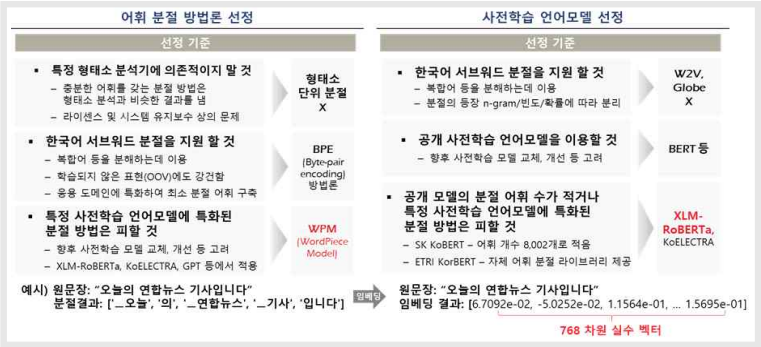
반면 인공지능 기반 통계분류는 대량의 데이터로부터 학습을 통해 자동으로 입력 텍스트의 패턴을 찾아내 사람에 의한 작업에 비해 더욱 정확하고 일관성 있는 분류 작업을 수행할 수 있다. 또한 대규모 데이터를 빠르게 처리할 수 있어 분류 작업의 효율성을 높일 수 있으며, 분류 항목의 추가나 변경 등에도 빠르게 대처할 수 있는 장점을 지닌다.



<그림 3-1> 기존 시스템과 AI 통계분류 시스템 방법론 비교

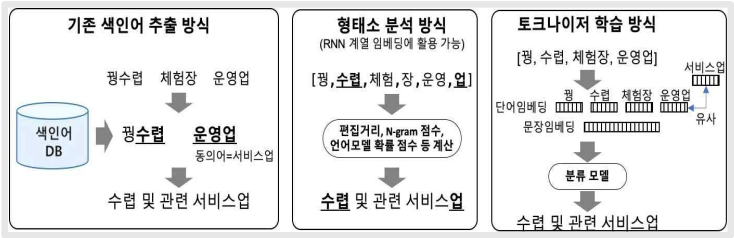
1. 자연어처리: 사전학습 언어모델 활용

AI 통계분류를 위해서는 우선 입력된 텍스트를 기계가 처리할 수 있는 숫자 형태의 정보로 변환하는 '자연어처리' 과정을 거쳐야 한다. 이를 위해 전통적으로는 형태소 분석 기준 어휘 분절 방식이 많이 사용되어 왔으나, 최근에는 구글 BERT나 OpenAI GPT 등 대규모 말뭉치를 Transformer와 같은 Attention mechanism[25]이 적용된 인코더-디코더 모델로 미리 사전에 학습하여 특정 자연어처리 태스크에 맞춰 전이학습을 수행할 수 있도록 개발된 '사전학습 언어모델'을 활용하는 것이 일반적이다.



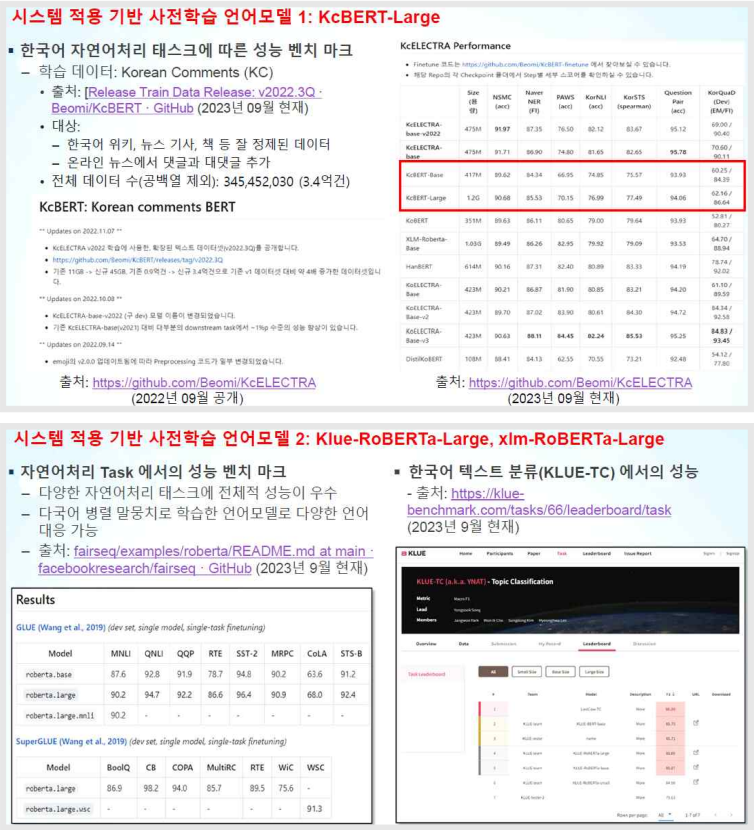
<그림 3-2> 사전학습 언어모델 선정기준

[참고] 입력 표현 분석 방식별 비교



25) Transformer 모델(Ashish Vaswani 등, 2017)은 특정표현에 대한 고유한 정보로 단어 간 상호등장(Co-occurrence) 관계 정보를 이용하여 문장 내 표현의 선행 단어(토큰)로부터 다음에 등장할 단어(토큰)의 확률을 학습/예측하는 자연어처리 알고리즘으로, BERT로 대표되는 분석에 특화된 인코더(Encoder)모델과 Chat GPT로 대표되는 생성에 특화된 디코더(Decoder)모델과 같은 LLM(Large Language Model)의 기초로 활용되고 있다.

통계청 인공지능 통계분류 시스템에서는 사전학습 언어모델로 KcBERT와 Klue-RoBERTa 등을 조사특성이나 분류체계 및 입력 데이터 특성에 맞춰 활용하였다. 이 모델들은 언어모델 벤치마크 평가 결과 우수한 성능을 보여주고 있다.26)



<그림 3-3> 시스템 적용 사전학습 언어모델 성능 벤치마크

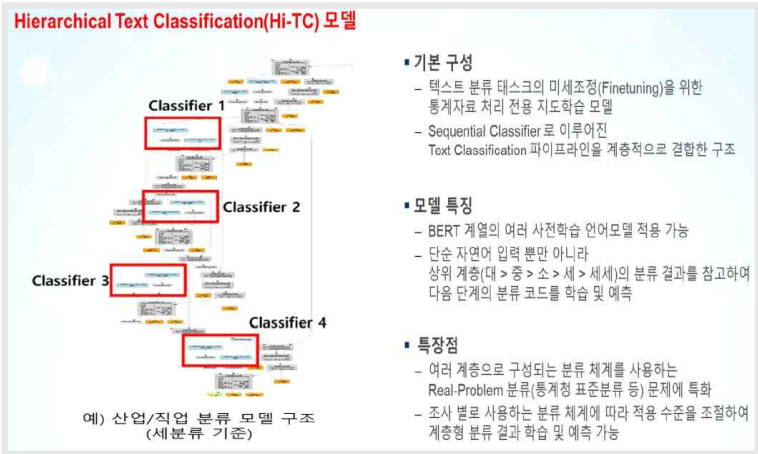
26) KcBERT는 대규모 온라인 뉴스 댓글 데이터를 활용하여 학습된 BERT모델로 특히 신조어, 구어체, 특수문자 등 다양한 텍스트에 유연하게 대응할 수 있도록 설계되었다. Klue-RoBERTa는 다양한 자연어처리 태스크의 한국어 말뭉치를 기반으로 사전학습된 RoBERTa모델로 한국어 텍스트에 대한 깊은 이해력을 보유하고 있다. - 오교중 외(2023)「인공지능 기반 사회통계 생산 방법론 고도화방안:가계동향조사와 생활시간조사 사례」제35회 한글 및 한국어 정보처리 학술대회 논문집 中

2. 지도학습 방법: 계층형 분류모델 적용

통계청 조사통계의 과거 공표자료를 AI 학습데이터로 사용함에 따라 통계분류를 위한 기계학습 방법으로는 문제와 정답을 함께 제공하는 사례기반 모델링으로 분류 예측에 높은 정확성을 기대할 수 있는 지도학습(supervised-learning) 방식을 적용하였다.

기존에는 규칙이나 사례 지식에 대한 추론을 적용하기 위해 의사결정 트리(decision tree) 혹은 데이터의 분포 정보를 활용한 선형 판별 분석(Linear Discriminant Analysis) 등의 방식을 분석에 많이 사용하였으나, 최근에는 컴퓨터 연산량의 증가에 따른 딥러닝 연구가 활성화되면서 순방향 신경망(feed-forward neural network)을 활용한 분류 모델 구축이 일반적으로 활용되고 있다.

통계청 인공지능 통계분류 시스템에는 자연어 추론 이외에도 감성 문석, 문법 정확성 평가 등 다양한 텍스트 분류 작업에 높은 성능을 보이는 허깅페이스의 텍스트 분류 파이프라인을 기반으로 개발된 SequentialClassfier 모델을 적용하였다.27) 그리고 보통 대-중-소-세-세세분류 등의 계층형 분류구조를 지니는 통계분류 태스크 특성을 반영한 미세조정 모델로 Hierarchical Text Classification 모델을 개발하였다.



<그림 3-4> 시스템 적용 지도학습 모델

27) 앞에서 설명한 KcBERT, Klue-RoBERTa 등의 사전학습 언어모델 및 SequentialClassfier 같은 분류모델은 허깅페이스(Huggingface) Model Hub의 트랜스포머스(Transformers) 라이브러리를 통해 쉽게 접근할 수 있다.

제3절 인공지능 통계분류 예측결과 및 실무활용

2017~2021년 건설업조사 전수 데이터 883만 건을 공종·발주자분류에 대한 인공지능 학습에 사용하였다. 이 가운데 중복 및 결측을 제거한 데이터를 약 9:1의 비율로 학습용 및 (분류예측 모델 성능) 검증용 데이터로 나누어 활용하였다. AI 예측 정확도 평가 기준으로는 정분류율(accuracy) 및 F-1 Score를 주요 평가지표로 활용[28]하였으며, 2021년 건설업조사 내검대상 데이터 92만 건을 평가데이터셋으로 활용하여 분류 예측 정확도 성능을 평가하였다.

<표 3-4> 건설업조사 공종·발주자분류 학습데이터

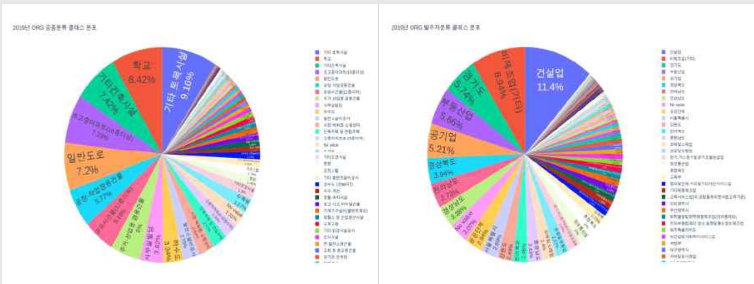
- 공사종류 데이터 - KDCS

	CNST_NM	ORG_STAT_KDCS_CD	STAT_KDCS_CD	ATMC_STAT_KDCS_CD
0	월계중 화장실 보수공사	433.0	433	433.0
1	대학원 건축물 구조부 보수 및 보강공사시	433.0	433	433.0
2	덕산초등학교 화장실보수 및 기타공사	490.0	433	433.0
3	경남교육연수원 화장실 보수공사	490.0	433	433.0
4	노형초 유치원, 화장실 수선 및 관리실 재배치공사	433.0	433	433.0
...	...	...	...	...
1187850	동막해수욕장 주변 보행환경 조성사업 중 상하수도 공사	292.0	210	NaN
1187851	온수리 일원 하수(배수)관거 정비사업	292.0	292	NaN
1187852	마그네사거리~신현리 보행환경 개선공사 중 상하수도 공사	292.0	292	NaN
1187853	국제화복합단지 완충녹지공사중 토공및부대공사	611.0	299	NaN
1187854	하자보수 울타리 보수보강 및 배수 트렌치 토목공사	412.0	620	NaN

- 발주자 데이터 - PPO

CNST_NM	PPO_NM	ORG_STAT_PPO_CD	STAT_PPO_CD	ATMC_STAT_PPO_CD
월계중 화장실 보수공사	광주광역시서부교육지원청	4500	240	240.0
대학원 건축물 구조부 보수 및 보강공사시	광주과학기술원	4500	91	91.0
덕산초등학교 화장실보수 및 기타공사	경상남도창원교육지원청	380	380	963.0
경남교육연수원 화장실 보수공사	경상남도교육청	380	380	380.0
노형초 유치원, 화장실 수선 및 관리실 재배치공사	제주시교육지원청	390	390	390.0
...	...	...	...	...
동막해수욕장 주변 보행환경 조성사업 중 상하수도 공사	인천광역시 강화군	230	230	230.0
온수리 일원 하수(배수)관거 정비사업	인천광역시 강화군 길상면	230	230	230.0
마그네사거리~신현리 보행환경 개선공사 중 상하수도 공사	인천광역시 강화군	230	230	230.0
국제화복합단지 완충녹지공사중 토공및부대공사	인천경제자유구역청	230	230	230.0
하자보수 울타리 보수보강 및 배수 트렌치 토목공사	장유청용예가입주자대표회의	999	960	960.0

- 공사종류 데이터 특징
  - KDCS – 64종류의 Label로 학습/예측을 진행
  - 조건 : ORG_STAT_KDCS와 STAT_KDCS값이 동일한 데이터
  - 중복 제외 17~21년 전체 데이터 : 6,328,076 건
- 발주자종류 데이터 특징
  - KDCS – 88종류의 Label로 학습/예측을 진행
  - 상당수 데이터가 공사명에 발주자 이름이 포함되어 있음
  - 조건 : ORG_STAT_PPO와 STAT_PPO값이 동일한 데이터
  - 중복 제외 17~21년 전체 데이터 : 5,887,909 건



<그림 3-5> 건설업조사 공종·발주자분류 데이터 분포 예시

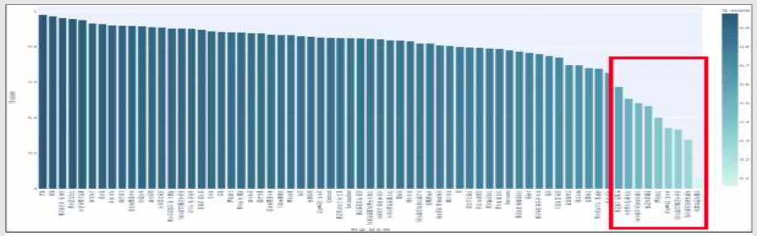
28) 정분류율(accuracy)은 분류 모델이 평가 데이터셋에 대하여 올바르게 분류한 항목의 비율을 나타내며, F-1 Score는 모델의 정밀도와 재현율의 조화평균을 나타내는 지표로 데이터 분포의 불균형이 심할 때 보다 효과적인 것으로 평가된다.

앞서 살펴본 사전학습 언어모델을 활용하여 학습용 데이터에 대해 지도학습 방식 학습을 거쳐 평가용 데이터에 대한 계층형 분류모델을 적용한 결과, 공종분류의 경우 약 85%, 발주자분류의 경우에는 89%라는 비교적 높은 정확도로 분류코드를 예측함을 확인할 수 있었다.

<표 3-5> 건설업조사 공종·발주자분류 성능평가 결과

분류 구분	관련 부서	적용조사 (주기)	데이터양 (추이/예측)	기초 데이터 이수 건수	데이터 전처리	데이터 정제	데이터 분리	학습/평가 품질 (Accuracy / F-1)	구축횟수 (최종학습일)
한국표준 산업 (10차)	고용 통계과	지역별 고용조사 (연 2회)	약 43만 (6개월 간격)	총 347만 (44~51회 총 8회차 19~22 4년 자료)	결측: 109만 중복: 32만	122만	학습: 110만 검증: 12만 평가: 43만	검증: 92.74% / 0.9271 평가: 94.17% / 0.9414	4회 (12/14)
	경제 총조사과	전국사업체 조사 (연 1회)	약 800만 (1년 간격)	총 2,594만 (18~21년)	결측: 484만 중복: 553만	834만	학습: 751만 검증: 83만 평가: 723만	검증: 85.04% / 0.8508 평가: 86.48% / 0.8623	9회 (01/11)
	인구 총조사과	인구총조사 (5년)	약 1,000만 (5년 간격, 전가구원 종합)	총 981만 (2020 인구총조사 기초데이터)	처리: 571만 중복: 40만 결측: 65 개	410만	학습: 369만 검증: 41만	검증: 86.01% / 0.8610 평가: 85.60% / 0.8551	6회 (11/06)
한국표준 직업 (7차)	고용 통계과	지역별 고용조사 (연 2회)	약 43만 (6개월 간격)	총 347만 (44~51회 총 8회차 19'~22' 4개년 자료)	결측: 109만 중복: 32만	122만	학습: 110만 검증: 12만 평가: 43만	검증: 91.44% / 0.9141 평가: 92.67% / 0.9259	5회 (12/14)
	인구 총조사과	인구총조사 (5년)	약 1,000만 (5년 간격, 전가구원 종합)	총 981만 (2020 인구총조사 기초데이터)	처리: 571만 중복: 40만 결측: 65	410만	학습: 369만 검증: 41만	검증: 79.52% / 0.7915 평가: 79.25% / 0.7896	6회 (11/06)
건설업 공종	산업 통계과	건설업 조사 (연 1회)	평균 180만 (1년 간격)	총 883만 (17~21년 전수)	중복: 250만	633만	학습: 192.1만 검증: 21.3만 평가: 92.1만 (21년, 내검 대상)	검증: 84.22% / 0.8411 평가: 85.03% / 0.8482	8회 (01/04)
건설업 발주자	산업 통계과	건설업 조사 (연 1회)	평균 180만 (1년 간격)	총 883만 (17~21년 전수)	중복: 294만 결측: 5만	589만	학습: 192.6만 검증: 21.4만 평가: 92.2만 (21년, 내검 대상 Y)	검증: 91.81% / 0.9177 평가: 89.32% / 0.8937	6회 (01/04)
물품통관 분류	서비스업 동향과	온라인쇼핑 동향 (월 1회)	평균 230만 (1월 간격)	총 9,730만 (21 ~ 22년 2년 자료)	중복: 8,315만 (21' 01.~22' 08.)	상품군: 93만 (22년 04월) HS코드: 1,182만 (21' 01.~22' 08.)	학습: 26만 검증: 2.9만 평가: 9.3만 (상품군, 22년 04월)	검증: 96.54% / 0.9653 평가: 96.86% / 0.9685	4회 (12/02)

2023년에는 AI 공종·발주자분류 예측결과를 건설업조사 자료처리 과정에 활용하기 시작하였는데, 아래의 그림과 같은 AI 분류코드 예측확률값 정보를 내용검토 우선순위 및 중요도 판단에 활용하는 방식으로 실무에 사용하였다.



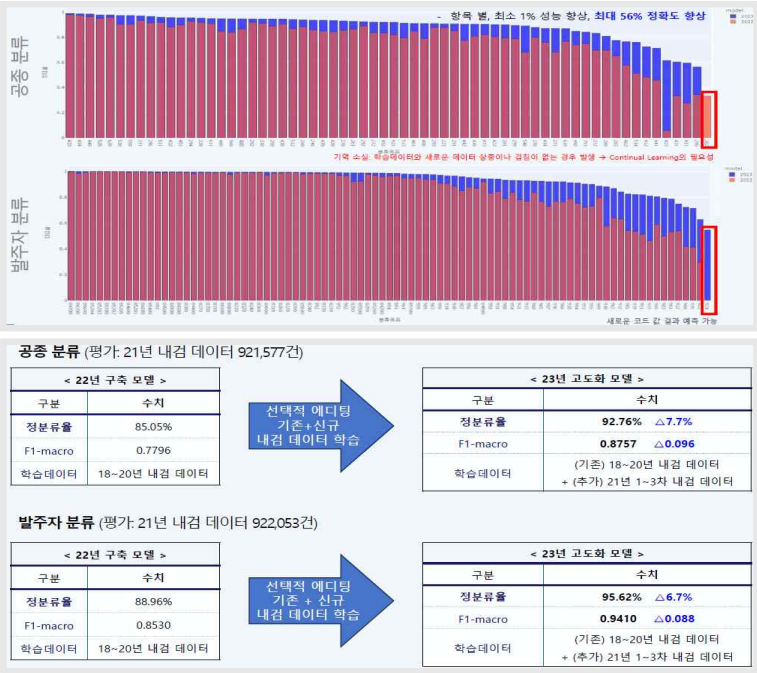
	text	prob
0	어린이집에서 보호자로부터위탁을받아 보호자로부터위탁을받아	0.898575
1	사업장에서 소매상 대상 소매상 대상	0.853393
2	영업장에서 고객의뢰를 받아 고객의뢰를 받아	0.602864
3	교회에서 기독교종교활동 기독교종교활동	0.334787
4	초등학교 초등교육서비스 초등교육서비스	0.572152

<그림 3-6> AI 통계분류 예측결과를 활용하여 내검효율성 제고

제4절 학습데이터 갱신 및 재학습을 통한 분류정확도 향상

2023년 잠정 공표자료를 학습데이터에 추가하여 재학습을 수행[29]하고, 기존 성능평가 데이터셋이었던 2021년 92만 건에 대한 분류예측 정확도를 재측정한 결과 공종분류 7.7%, 발주자분류의 경우 6.7%의 정확도 상승 결과를 확인하였다.

이는 향후 AI 분류예측 결과를 활용한 자료처리 결과를 반영하는 학습데이터 갱신 및 재학습만으로도 (사전학습 언어모델이나 분류알고리즘 등의 개선 없이도) AI 분류예측 정확도가 지속적으로 개선될 수 있으며, 그에 따라 사람(내검원)이 확인해야 하는 내용 검토 업무량을 계속적으로 줄일 수 있음을 보이는 매우 고무적인 결과라고 생각한다.



<그림 3-5> 재학습을 통한 분류정확도 향상 (붉은색: 기존, 파란색: 추가학습)

29) 지속적 학습(continual learning) 최적 방법론으로 과거 학습데이터에 신규 데이터셋을 추가 사용하는 iCarl (Sylvestre-Alvise Rebuff 등, 2017) 및 Rehearsal Methods를 적용하였다.

제 4 장

인공지능 통계분류 실무활용성 제고방안

앞 장에서 살펴본 바와 같이 학습데이터 갱신 및 재학습을 거친 건설업조사의 공종·발주자 분류코드에 대한 AI 분류예측은 90%를 상회하는 높은 정확도 수준을 보이고 있으며 향후에도 학습데이터의 체계적인 분석 및 관리를 통해 정확도 개선이 지속될 것으로 기대한다.

그럼에도 불구하고 100% 분류예측 정확도에 이르지 않는 한 통계청에서는 여전히 내검원에 의한 분류코드 확인작업을 수행할 것으로 보이는바[30], 이러한 상황에서 AI 통계분류를 활용한 통계분류 자료처리의 효율성을 높이기 위한 방안을 검토하고자 한다.

제1절 분류체계 개선

건설업조사의 공종·발주자분류는 조사표의 '공사명'과 '발주자명'에 대한 입력값을 토대로 그에 해당하는 분류코드를 기입하는 방식으로 조사되고 있다. 2023년 현재 AI 통계분류 예측 정확도는 정분류율(accuracy) 기준 공종분류 92.76% 발주자분류 95.62% 수준을 보이고 있는데, 예측 정확도가 떨어지는 분류항목에 대한 원인 분석으로 다음과 같은 요인을 생각해 볼 수 있다.

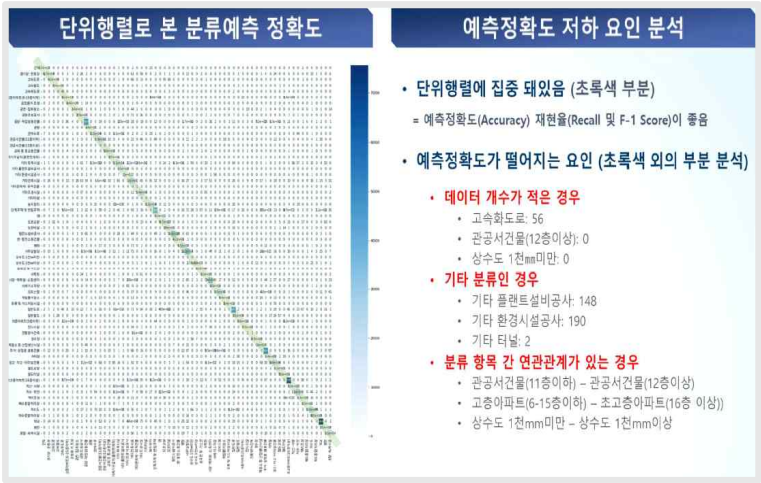
(1) 데이터 개수가 적은 경우: 고속화도로, 관공서건물(12층 이상), 상수도(1천mm 미만) 등의 분류코드는 해당 레코드 수가 극히 미미하여 충분한 학습이 이루어지기 어려운 데이터 특성을 보이고 있는데, 이는 주요한 분류예측 저하요인이 된다.[31]

30) 장기적으로는 표본조사의 신뢰수준과 같이 AI 분류예측 정확도의 일정 수준(예를 들어 95%)을 기준으로 오차를 인정하고 통계분류 자료처리를 완전 자동화하거나, 분류항목별 예측정확도 기준을 설정하고 그 이하에서만 내검을 수행하는 방안 등을 검토할 필요가 있다고 생각한다.

31) 예측정확도를 위해 분류라벨당 필요한 최소 학습데이터 개수는 사전학습 언어모델의 성능이나 생성형 언어모델 활용 여부 등에 따라 달라질 수 있는 것으로 일률적인 기준은 존재하지는 않는 것으로 보이나, (과거 공표자료라는 정답데이터셋이 명확하고 조사문항이라는 입력데이터가 비교적 정형화된 특성에 맞춰) RoBERTa 언어모델과 지도학습 방식 분류예측 방법론을 적용한 AI 통계분류 자동화 시스템에서는 분류라벨당 최소 400개 정도의 학습데이터를 확보하는 경우 안정적인 분

(2) 분류항목 간 연관관계가 있으나 구별정보가 조사문항에는 존재하지 않는 경우: 관공서건물(11층 이하)-관공서건물(12층 이상), 저층아파트(5층 이하)-고층아파트(6-15층 이하)-초고층아파트(16층 이상), 상수도(1천mm 미만)-상수도(1천mm 이상)의 경우 해당 공종분류코드는 분리되어 있으나 공사명에 층높이 등의 구별정보가 없는 경우가 많은 상황이다. 이러한 분류 영역에서 AI 분류예측은 빈도수의 영향을 크게 받아 오류 가능성이 높아질 수밖에 없다.[32)]

(3) 기타 분류인 경우: 기타 분류의 경우 속성상 해당 카테고리 내 텍스트의 패턴이 이질적일 수밖에 없고 그에 따라 AI 분류예측 효율이 낮아지게 된다.



<그림 4-1> AI 통계분류 예측 정확도 저하 요인 분석

향후 데이터 개수가 작은 분류영역이나 분류 항목 간 연관관계가 있는 경우에는 분류코드를 통폐합하고, 기타 분류는 가급적 축소하는 방향으로 분류체계 개편을 검토할 필요가 있다고 보인다.[33)]

류예측 성능이 도출되는 것으로 분석되었다.

32) 이 경우 내검원에 의한 확인작업도 조사문항 자료 이외 홈페이지 검색이나 전화문의 등 추가 정보를 확인하여 최종 분류코드를 정하는 방식으로 진행되게 된다.

33) 이미 통계청에서는 저층·고층·초고층 아파트를 '아파트'로, 상수도 1천mm 미만과 이상을 '상수도'로, 관공서 건물 11층 이하와 12층 이상을 '관공서건물' 등으로 통합하여 결과를 공표하고 있다. 그럼에도 불구하고 조사표 입력 시에는 '건설산업기본법'에 명시되어 있는 현행 64종 공종세분류를 기준으로 조사가 이루어지고 있는 상황이다. 향후 건설산업기본법상의 공종세분류 개편 시 데이터

제2절 조사단계에서 인공지능 통계분류 예측결과 활용

건설업조사는 건설업체에서 응답자 기입방식으로 각 협회 시스템을 통해 조사 대상 연도의 건설경영실태 및 건설공사실적 등의 자료를 입력하고 통계청은 건설 관련 협회 대상 인터넷조사를 통한 방식으로 자료가 입수되고 있다.

통계청 자료처리 과정의 효율을 높이기 위해서는 조사단계의 정확한 자료입력이 무엇보다 중요할 것이다. 건설업조사 공종·발주자 분류코드에 관련해서는 다음 두 가지 방안을 제안하고자 한다.

(1) 최초 입력된 분류코드와 최종 공표된 분류코드가 변동된 케이스(이는 AI 통계분류 예측 정확도가 낮은 분류영역의 경우와 관련성이 매우 높다고 보인다)에 대한 과거자료를 분석하고 그 결과를 대한건설협회 등 협회 통계업무 담당자에게 교육하는 과정이 필요하다. 이를 통해 조사단계에서 보다 정확한 응답이 가능할 것으로 본다.

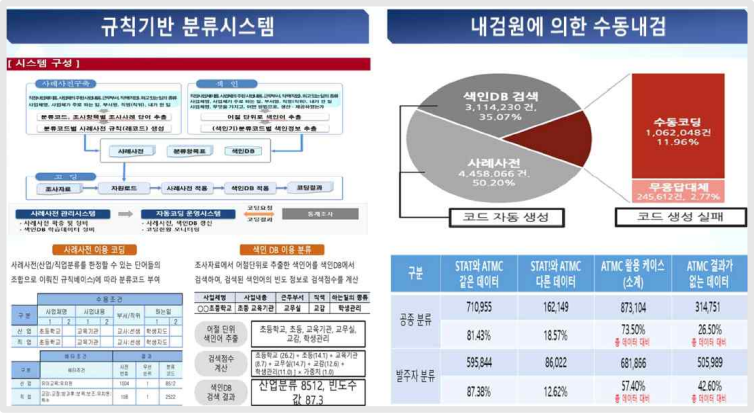
(2) 협회 통계업무 담당자들이 AI 통계분류 자동화 시스템을 활용하여 조사입력 과정에서 보다 정확한 분류코드를 입력할 수 있게 지원하는 방안을 고민할 필요가 있다. 통계청이 현재 서비스하고 있는 '통계분류포털'은 분류기준 정의에 대한 설명자료 기준으로 색인어 위주의 검색결과를 제공하고 있어 조사에 활용이 제한적인 데 비해, AI 통계분류 자동화 시스템은 과거 응답자료에 기반한 분류예측 결과를 보다 효과적으로 제공할 수 있을 것으로 생각한다.

<그림 4-2> 현장조사에 AI 통계분류 예측결과 활용 예시

개수가 작은 분류영역 등에 대한 통폐합이 필요하다고 생각한다.

제3절 기존 시스템을 인공지능 통계분류 시스템으로 대체

통계청의 기존 통계분류 자료처리 과정은 규칙(rule) 기반 사례사전 방식의 자동코딩과 색인DB 빈도수 정보에 기반한 추천분류코드를 활용하여 내검원이 최종 분류코드를 확정하는 방식으로 진행되어 왔다. 앞서 살펴본 바와 같이 이와 같은 방식은 규칙 정비 및 관리에 업무담당자의 많은 노력이 필요하며, 낮은 커버리지로 인해 사람이 확인해야 하는 업무량이 많다는 한계를 가지고 있다.



<그림 4-3> 기존 규칙기반 통계분류 시스템 활용 예시

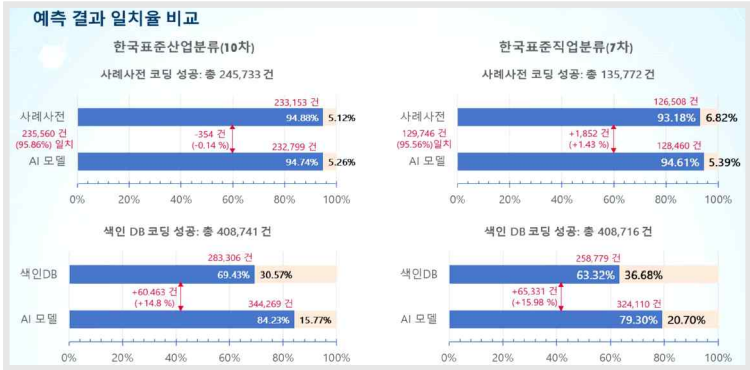
건설업조사의 경우 사례사전 자동코딩은 공종분류 81.43% 발주자분류 87.38%라는 비교적 높은 정확도를 보이지만, 이는 사례사전에 의한 자동코딩이 이루어지지 않은 경우를 제외하고 계산된 것으로써 (공종분류 26.5% 발주자분류의 42.6%에 대해서는 사례사전 자동코딩 결과 없음) 전체 데이터 기준으로 계산한 정확도는 공종분류 73.5% 발주자분류 57.4% 수준으로 떨어지게 된다.

사례사전 자동코딩의 경우 낮은 커버리지 및 정확도 문제뿐 아니라 정확도에 대한 정보를 내검원이 활용할 방법이 없다는 점 또한 문제가 된다. 윗 단락에서 표현한 정확도는 사례사전 자동코딩 결과가 최종 공표결과와 같은지 여부로 사후적으로 측정되는데, 막상 자료처리 과정 중에는 개별 데이터별 사례사전 자동코딩 결과를 어느 정도 신뢰할 수 있는지에 대한 정보가 없이 내검을 진행할 수밖에 없다.

내검원이 사실상 전체 데이터 모두를 확인해야 하는 상황이 발생하는 것이다.

이와 같은 상황에도 불구하고 업무담당자의 전문성에 기초한 규칙체계인 사례사전에

대한 신뢰가 지속되고 있는 것으로 보인다. 낮은 커버리지와 정확도 문제는 업무담당자가 더욱 열심히 규칙체계를 추가 및 재정비함으로써 극복할 수 있다고 보는 인식도 광범위하게 존재하고 있는 것으로도 보인다.



<그림 4-4> 사례사전 자동코딩과 AI 통계분류 정확도 비교[34]

위 그림은 통계청에서 가장 엄밀한 규칙체계로 관리하고 있다고 평가되는 산업직업분류에 대한 사례사전 자동코딩[35]과 AI 통계분류 결과를 비교 분석한 결과이다.

(커버리지 문제를 제외하기 위해 사례사전 자동코딩이 존재하는 데이터에 한정해) 사례사전 자동코딩 결과와 AI 통계분류 결과를 비교한 결과 놀랍게도 AI 통계분류 예측 결과가 사례사전 자동코딩과 유사하거나 오히려 더 높은 정확도를 보임을 확인할 수 있다. 이는 사람이 발견할 수 있는 규칙체계의 경우 AI 기술 또한 그 패턴을 반영하여 적용함을 의미하는 것으로 해석할 수 있다고 본다.

또한 앞의 <그림 3-6>에서 확인할 수 있는 바와 같이 (사례사전 자동코딩과는 달리) AI 통계분류 자동화 시스템은 예측 분류코드와 함께 정확도에 대한 확률정보를 개별 레코드 단위로 제공할 수 있는 바, 이 정보를 활용하여 사람이 집중해서 확인해야 하는 내용검토 대상을 선별하여 자료처리를 진행할 수 있다는 장점이 존재한다.

이러한 점에서 규칙기반 사례사전 시스템을 유지할 실익은 크지 않다고 생각한다.

---

34) 2020년 인구주택총조사의 산업직업분류 데이터 분석 결과임 - 통계청 '인공지능 기반 통계분류 자동화 시스템 구축사업' 완료보고회 발표자료 中

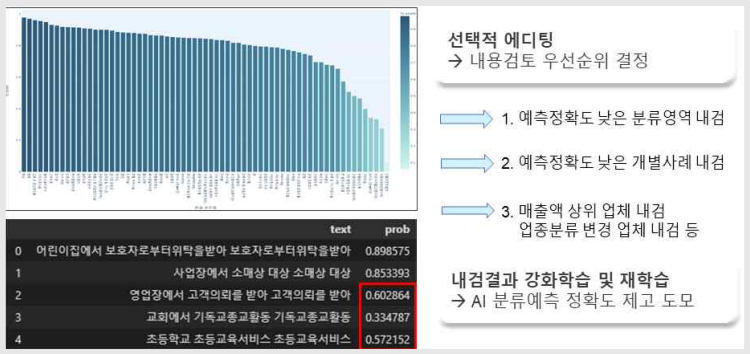
35) 산업직업분류의 경우 10년이 넘는 기간 동안 동일한 전담 업무담당자가 2022년 6월 기준 37,424건의 사례사전 규칙체계를 관리하여 왔으며, 사례사전 자동코딩 결과가 도출되는 경우에는 95%라는 매우 높은 수준의 정확도를 보이고 있다.

제4절 인공지능 통계분류 예측 결과를 활용한 내검 효율화

2023년 통계청은 건설업조사 자료처리에 AI 통계분류 예측결과를 시범 활용하였다. 기존 사례사전 자동코딩 결과와 함께 AI 분류예측 코드 및 확률정보를 내검원에게 제공하여 최종 분류코드 확정 시 참고하는 방식이었다. 이와 같은 제한적인 활용만으로도 ① 내검 업무 효율성 증대(내검원당 업무처리량 대략 5% 이상 증대) 및 ② 선택과 집중에 의한 내검 결과 재학습으로 AI 분류예측 정확도 약 7% 향상이라는 효과를 확인할 수 있었다.

향후 AI 통계분류 예측결과의 보다 효과적이고 적극적인 활용을 위해서는 내검체계 정비가 필요하다고 본다. 앞 절에서는 기존 사례사전 자동코딩을 AI 통계분류로 대체할 수 있음을 보였는데 여기서 한 발 더 나아가 AI 통계분류 예측 정확도 정보를 활용하여 내검원에 의한 확인 대상 데이터 범위를 한정하여 업무량을 대폭 경감할 것을 제안한다.

예를 들어 AI 분류예측 정확도가 95% 이상인 데이터는 AI 분류코드를 적용하고, 그 이하 수준인 경우에만 내검원에 의한 확인을 진행하는 방식을 적용할 경우 대폭적인 업무량 감축이 가능할 것으로 본다.[36]



<그림 4-5> AI 통계분류 예측결과를 활용한 선택적 내검(안)

위 그림에서와 같이 내검원에 의한 확인대상을 AI 분류예측 정확도가 낮은 영역에 집중하고, 집계에 큰 영향을 주는 사업체와 전년 대비 업종 변경 사업체 등을 보충적으로 내검하는 경우 통계 정확성을 저해하지 않으면서 효율성을 제고할 수 있을 것이다.

36) 현재 통계청 산업통계과에서는 건설업조사 자료처리를 위해 매년 약 10명 내외의 내검요원을 채용하여 10개월 이상 내용검토를 진행하고 있는데, 작업 중 많은 부분이 공종·발주자분류 확인과 관련된다(건수로는 매년 약 80만 건의 데이터에 대해 내검원에 의한 확인작업이 수행되고 있다).

제 5 장

결론 및 시사점

본 연구에서는 AI 통계분류 연구 추진경과와 사전학습 언어모델을 활용한 자연어처리 등 AI 통계분류 방법론을 살펴보았고, 건설업조사 공종·발주자분류에 대한 AI 통계분류 분석결과와 향후 AI 통계분류 실무활용성을 제고하기 위한 방안들을 검토하였다.

통계개발원에서는 미래 통계환경 변화에 대한 대응의 일환으로 데이터과학(Data Science) 기법의 국가통계 적용 분야로 'AI 기반 통계분류 자동화' 연구를 추진하였다. 2020~2021년 AI 통계분류 가능성 및 기존방식 대비 효과성 개선 여부 등에 대한 기초 연구를 거쳐 2022년에는 'AI 통계분류 자동화 시스템'을 개발하여 2023년부터는 일부 조사통계를 대상으로 실무에 시범 적용하고 있다.

'18~'21년 건설업조사 4개년 자료에 대해 Klue-RoBETa 언어모델과 지도학습 방식의 Hierarchical Text Classification 분류 알고리즘을 적용 결과 공종분류 85.03%, 발주자분류 89.32% 수준의 분류 예측 정확도를 확인하였다. 이를 토대로 2023년에는 AI 분류예측 코드 및 정확도에 대한 확률정보를 분류코드 내검과정에 활용하여 업무효율을 제고하였으며, 내검이 종료된 데이터를 추가하여 재학습한 결과 분류 예측 정확도가 공종분류 92.16%, 발주자분류 95.62%로 대폭 향상되는 결과를 도출하였다.

그리고 AI 통계분류 실무활용성을 제고하기 위해 다음과 같은 네 가지 방안을 제안하였다. 첫째, AI 통계분류 과정의 기술적 특성을 고려하여 분류코드당 데이터 개수가 적거나 상호연관성 있는 분류코드의 통폐합, 기타 분류 항목의 축소 등의 분류체계 개선이 필요하다. 둘째, 조사입력 단계에서부터 정확한 분류코드가 작성될 수 있도록 현장조사에 AI 통계분류 예측결과의 활용이 필요하다. 셋째, 기존 규칙(사례사전) 기반 분류시스템을 AI 통계분류 자동화 시스템으로 대체하는 것이 더 효율적이다. 넷째, 예측정확도가 95% 이상인 데이터에 대해서는 AI 예측 분류코드를 적용하는 등의 AI 통계분류 예측결과를 활용하는 체계적인 내검 개선안을 마련할 필요가 있다고 보았다.

이러한 분석 결과는 자연어처리 등 AI 기술이 적용 가능한 국가통계 전문영역을 개척하고 관련된 기술을 검증하여 실무에 지속적으로 적용할 수 있는 토대를 마련한 것으로써 향후 AI 기술 확산에도 크게 기여할 것으로 생각한다. 또한 통계청의 통계분류 자료처리의 정확성·일관성 및 시의성을 제고하는 한편 반복적 업무의 자동화·지능화를 통한 행정 효율성 향상에도 이바지할 것으로 기대한다.

참고문헌

강유경. (2001). **산업/직업 분류 자동코딩 시스템**, 한국조사연구학회 2001년 추계학술대회.

김정민 외. (2020). **AI 기술의 국가통계 활용 사례 및 국내 도입 촉진 방안**, 소프트웨어정책연구소(SPRi) ISSUE REPORT.

김혜란, 임경민. (2020). **데이터과학 기초연구**, 통계개발원.

박성현. (2018). **4차 산업혁명 시대의 데이터 사이언스의 역할과 미래비전**, 미래한국 기고문.

오교중 외. (2021). **인공지능 기반 통계조사 분류시스템 구축 방안 연구**, 통계개발원.

오교중, 우찬균 외. (2022). 언어모델 전이학습 기반 해외직접구매 상품군 분류, **제34회 한글 및 한국어 정보처리 학술발표 논문집**.

오교중 외. (2023). 인공지능 통계분류 자동화 확대 적용 연구, 통계개발원.

오교중, 김건수 외. (2023) 인공지능 기반 사회통계 생산방법론 고도화 방안: 가계동향조사와 생활시간조사 사례, **제35회 한글 및 한국어 정보처리 학술발표 논문집**.

우찬균, 임희석. (2020). 딥러닝 기반 한국표준산업분류 자동분류 모델 비교, **한국정보처리학회 2020 춘계학술발표대회 논문집**.

임경민, 김혜란. (2020). **데이터과학 기법의 국가통계 시험적용 분석**, 통계개발원.

임경민. (2021). **인공지능 기반 산업분류 자동화 처리 연구**, 통계개발원.

임경민. (2023). 인공지능 기반 통계분류 자동화, **통계개발원 국가통계방법론 심포지엄 자료집**, 통계개발원.

임정우, 우찬균 외. (2021). 딥러닝 기법을 활용한 산업/직업 자동코딩 시스템, **한국융합학회논문지 제12권 제4호**.

임희석. (2005). 예제기반의 학습을 이용한 한국어 표준 산업/직업 자동코딩 시스템, **한국콘텐츠학회논문집**.

전영일 외. (2020). **인공지능(AI)과 데이터 사이언스, 국가통계의 도약**, KOSTAT 통계플러스.

K.Team, “Klue-RoBERTa: Pretrained language model for Korean”, 2021
[Online]. Available: https://github.com/KLUE-benchmark/KLUE

H.Inc.,“Huggingface’s transformers: Textclassification pipeline”, 2021
[Online]. Available: https://huggingface.co/transformers/mainclasses/pipelines.html#textclassificationpipeline

통계청. 건설업조사「조사지침서」,「분류내검 사례집」.

통계청. 산업직업분류 자동코딩시스템 직무편람.

통계청. 인공지능 통계분류 자동화시스템 구축사업, 사업계획서 및 완료보고회 발표자료.

부 록

2022년 AI 통계분류 자동화 시스템 구축 결과

Ⅰ 추진배경 및 경과

1 추진배경

□ 4차 산업혁명 시대를 맞아 다양한 대용량 자료를 최신 IT 기술과 연계하여 신속·정확하게 정보를 창출하고 활용하는 데이터과학(Data Science)의 중요성이 강조되고 있음

○ 이미 해외에서는 ‘비정형 텍스트 분석을 통한 분류 자동화’ 분야를 중심으로 **인공지능(AI) 기술을 국가통계 분야에 적용**하고 있음

[참고: 비정형 텍스트 분석을 통한 분류 자동화 적용 해외사례*]

국가	활용 내용
미국	- (BLS) 업무상 상해 및 질병 설문조사 응답결과 자동 코드 분류
	- (Census) 미국 지역사회 설문조사의 산업 및 직업에 대한 코드 분류
독일	- 국내 법인에 대한 유럽 국민계정체계 기관 분류 할당
	- 온라인 구인 공고 분류
캐나다	- 월별 소매거래 및 분기별 소매상품 조사에 스캔 데이터 활용
핀란드	- 경찰청 교통사고 보고 문건에 기반해 사고 내역 자동 분류
벨기에	- 구인 공고별 알맞은 경제활동 통계 분류 코드 예측
룩셈부르크	- 민간 기업의 스캐너 데이터를 활용해 상품 분류 코드 추천
스위스	- 항공 이미지에 기반한 토지 피복 및 토지 사용 코드 분류 자동화

* 「데이터 과학 기법의 국가통계 시험적용 분석」 연구보고서 (2020, 통계개발원)

□ 통계청에서도 AI 등의 데이터과학 기법을 도입하여 통계생산 및 분석 프로세스를 개선하기 위한 다양한 노력을 경주하고 있음*

* 개발원 데이터과학 연구팀(‘20), 교육원 데이터사이언티스트 양성과정(‘21) 신설 등

○ 특히 **텍스트에 기반한 통계분류 자료처리에 AI 기술을 활용**하기 위한 다양한 연구를 집중적으로 진행하고 있음

> ❶ (고용과, '17년) '딥러닝 기술 활용 산업직업분류 정비' 내부 검토
> ❷ (인총과, '20년) '인공지능 기반 산업직업분류 자동코딩 시스템' 연구용역
> ❸ (개발원, '20년) 데이터과학 관련 '산업분류 코딩 자동화 시험분석' 연구
> ❹ (개발원, '21년) '인공지능 기반 통계분류 자동코딩시스템 구축' 연구용역
> ❺ (산통과, '22년) '건설업조사 분류코딩 머신러닝 기법 도입 및 활용' 연구용역
> ❻ (인총과, '22년) '인구주택총조사 업무 프로세스 개선방안' 연구용역

2 추진현황

□ 인공지능 기반 통계분류 자동화 시스템 구축 사업 추진

○ 해외사례 및 선행연구에 기반하여 범용 통계분류 플랫폼으로 '인공지능 기반 통계분류 자동화 시스템 구축' 사업을 진행 ('22.7월~'23.1월)

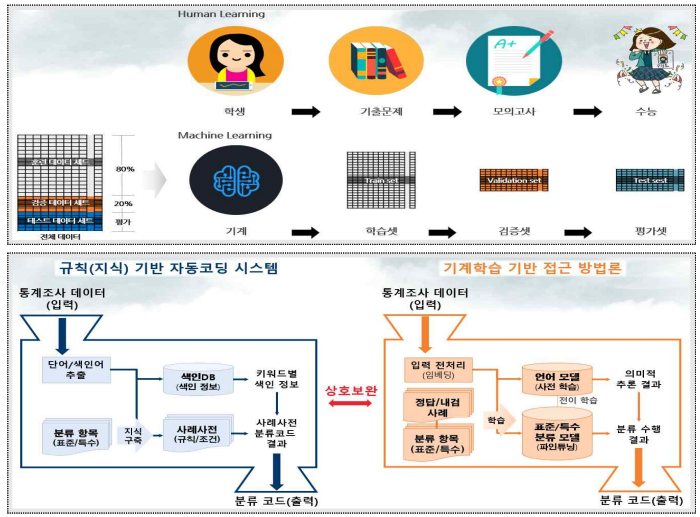
- '22년 **과기부「디지털 공공서비스 혁신 프로젝트*」지원 사업**으로 선정되어 개발예산 확보 (1,750백만 원, 한국지능정보사회진흥원 주관)
  * AI 등 첨단 ICT 기술을 공공분야에 선도적으로 적용하기 위해 중앙행정기관 및 지자체 공공기관 대상으로 사업공모를 통해 총 11개 과제를 선정 197억 원 지원
- 청내 인공지능 통계분류 유관부서*가 참여하는 TF 회의('22.2.24.)를 거쳐 사업추진 방향 확정 및 세부 사업계획서 작성 ('22.3월)
  * 조사시스템관리과, 통계방법연구실, 통계기준과, 경제총조사과, 고용통계과, 산업통계과, 서비스업동향과, 인구총조사과 등
- '제1회 **통계청 인공지능 활용 산업분류 자동화 경진대회**'('22.3.7.~.4.13.) 수상팀의 최적 통계분류를 위한 AI 알고리즘을 개발 사업에 반영

○ **인공지능 기반 통계분류 자동화 시스템 구축 사업 개요**

- 사업기간: 2022.07.24. ~ 2023.01.20. (6개월)
- 사업금액: 15.4억 (국고지원사업, NIA 주관)
- 사 업 자: ㈜아일리스프런티어 컨소시엄
- 사업범위: 기계학습 기반 통계분류 자동코딩 시스템 구축*
  * 5종 분류(산업/직업, 공종/발주자, 상품군별분류)
    5개 조사(인총, 전국사업체, 건설업, 지역별고용, 온라인쇼핑조사) 대상

Ⅱ 인공지능 통계분류 방법론

1 기존 시스템과의 차이



□ 기존 시스템은 규칙(지식) 기반 분류 자동코딩 지원

○ **(사례사전)** 업무담당자의 전문적 판단에 의한 내검규칙 적용*
  * 사전에 설정된 수용조건 및 배타조건 적용하여 분류코드 부여

○ **(색인DB)** 어절 단위로 추출된 색인어(keyword) 빈도 정보 활용*

⇨ (한계) - keyword 인식 문제: 오탈자, 띄어쓰기 등
  - 낮은 연계율: 사례사전 확충을 위한 지속적인 노력 필요
  - 시스템 부하: 내검규칙 확인을 위한 처리시간 지연 문제

□ **인공지능 기반 통계분류 자동코딩의 차별점**

○ **(자연어처리)** 자연어처리* 기술 적용하여 keyword 인식 문제 해결
  * 컴퓨터를 이용해 사람의 자연어를 분석하고 처리하는 기술 ex) BERT, GPT

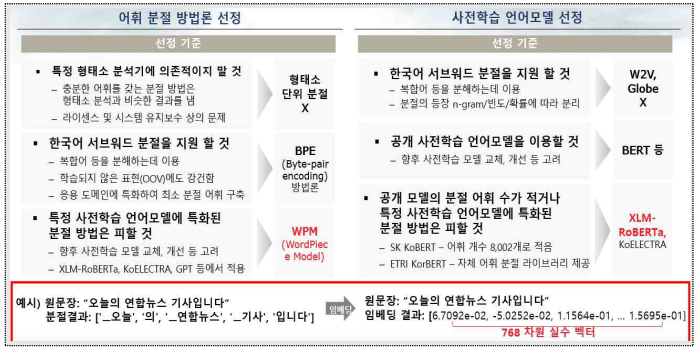
○ **(기계학습)** 데이터 학습을 통해 통계분류 자동처리 cf) 사례사전/색인DB 대체

○ **(GPU서버)** AI 분석에 특화된 GPU서버 활용으로 처리속도 향상

2 인공지능 통계분류 방법론

□ 통계 분야 특화 자연어처리

○ 입력된 텍스트를 컴퓨터가 처리할 수 있는 숫자 데이터로 변환



- **XLM-RoBERTa**: 페이스북에서 개발한 신경망 기반 언어모델*

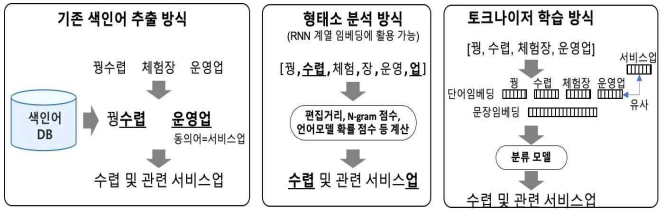
  * 구글 BERT 기반 다국어 어휘 학습이 강조된 모델 (2.5TB의 100개 언어로 학습)

- **WordPiece**: 구글이 2016년 발표한 토크나이징* 알고리즘**

  * 빈번히 결합되어 사용된 글자 패턴을 통해 복합어 및 개체명의 경계 자동학습

  ** 학습데이터에 대해 likelyhood(우도)를 최대화하는 방식으로 기호 쌍을 선택
  유연성, 효율적인 디코딩 등의 장점이 있으며, BERT 모델에 적용되어 성능을 입증

[참고: 입력 표현 분석 방식별 비교]



⇨ (장점) - 띄어쓰기와 무관하게 입력 결과 처리 분석 가능

- 통계분류 분야 말뭉치 추가 수집으로 학습데이터 자동 갱신
- 색인어 지식이 없는 분야(Out-of-vocabulary)에 강점
- 처리 속도 대폭 향상 ex) 1초에 약 15,000 문장의 토크나이징 가능

□ (학습)데이터 전처리

○ 데이터 분포 분석: 분류영역별 학습데이터 충분성 및 편향성 파악

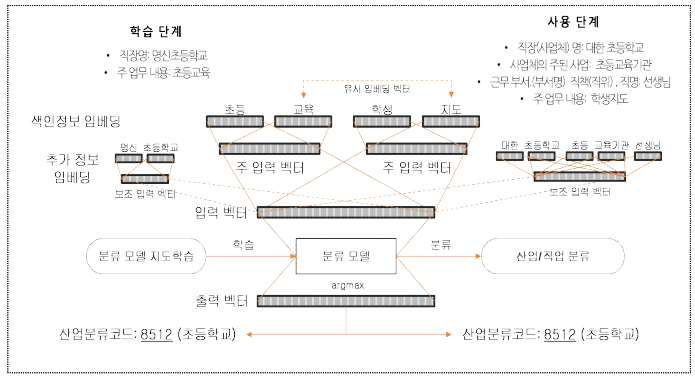
○ 중복데이터, 텍스트 입력 누락, 오입력 데이터 등 처리*

  * 전체 텍스트 입력 누락 및 존재하지 않는 분류코드 입력 여부 등 확인 및 처리

□ 인공지능 기반 통계분류 알고리즘 적용

○ 통계분류별 및 조사별 데이터 특성 고려한 최적 분류알고리즘 선택

  ⇨ 분류항목별 예측 분류코드 정확도 확률(신뢰도)을 제공



- **FlatBERT**: 기본형 모델로 계층 구조가 단순한 분류체계에 적합*

  * BERT 최적화 모델인 RoBERTa(Robustly Optimized BERT)를 베이스 모델로 활용

- **HiBERT***: 계층형 분류구조(대-중-소-세분류 등) 특성을 반영한 모델**

  * HiBERT(Hierarchy Bidirectional Encoder Representations from Transformer)

  ** 통계청 인공지능 활용대회('22.4월) 대상 수상팀 사용 알고리즘

Ⅲ 인공지능 통계분류 시험분석 결과

1 시험분석 결과 요약

□ 인공지능 통계분류 시험분석 결과 요약

○ **인공지능 통계분류 예측 정확도** (공표자료 기준)

(1) 인구총조사	- 산업분류 84.2%, 직업분류 79.6%
(2) 전국사업체조사	- 산업분류 84.1%
(3) 지역별고용조사	- 산업분류 88%, 직업분류 83%
(4) 건설업조사	- 공종분류 86%, 발주자분류 95%
(5) 온라인쇼핑조사	- HS코드 94%, 상품군별분류 94%

○ **기존 산업직업 자동코딩시스템과의 성능 비교** (산업분류 기준)

- 기존 시스템 자동코딩 결과보다 AI 분류예측 정확도가 더 높음

(가구조사) · AI 통계분류 예측: 100% 커버리지 84.2% 정확도

· 사례사전 자동코딩: 59.9% 커버리지 94.9% 정확도

*사례사전 자동코딩 커버리지 59.9%에 대한 AI 분류예측 정확도 94.7%

(사업체조사) · AI 통계분류 예측: 100% 커버리지 84.1% 정확도

· 색인DB 활용 예측: 99.5% 커버리지 73.8% 정확도

○ **범용 통계분류 자동화 시스템 가능성 확인**

- 통계분야에 특화된 자연어처리 및 분류 알고리즘 적용
- 산업·직업분류 이외 공종·발주자분류 및 상품군별분류에 있어서도 높은 수준의 AI 기반 분류 자동화 예측 정확성 확인*

  * 건설업조사 공종분류 86%, 발주자분류 95%, 온라인쇼핑조사 상품군별분류 94%
- 향후 질병·사인분류, 행동분류 등 기타 분류체계에 적용 확대 검토 필요

○ **인공지능 통계분류 처리 속도 개선**

- 기존 시스템 대비 최대 17배 처리속도 향상 확인*

  * (RMI 서버, 단건 요청 대비) 수행시간 15~17배 단축, (배치 시) 10분 이내 10만 건 처리
- 조사 및 내검기간 AI 통계분류 예측결과 제공 및 활용 강화

2 세부 조사별 분석 결과

(1) 인구총조사 산업·직업분류

□ 분석방향

○ AI 기반 통계분류 예측결과 정확도 확인

○ 현 산업직업 자동코딩시스템과 AI 통계분류 예측결과 비교 분석

○ AI 통계분류 시스템 분류예측 수행속도 측정

□ 분석데이터

○ 2020년 인구주택총조사 공표자료 전체를 대상으로 분석 수행

- 가구원 기준 총 9,810,963건 전체 데이터 중 산업·직업분류 항목 조사대상 가구원 4,099,010건*을 대상으로 분석

  * 국내 거주, 만 나이 15세 이상, 경제활동인구(일하지 않았음 제외)인 가구원

- 산업·직업분류 조사항목* 중 일부 누락이 있는 경우도 포함

  * 직장(사업체)명, 직장(사업체)가 하는 일, 근무부서, 직책(책무), 내가 하였던 일

- 성별, 나이, 교육정도, 종사상지위, 근로장소 정보 추가하여 분석
- 4,099,010건 중 90%는 학습·검증용, 10% **409,901건을 평가셋**으로 활용

□ 분석결과

○ **(AI 분류예측 정확도) 산업분류 84.2%, 직업분류 79.6%** 정확도* 확인

* 409,901건에 대한 AI 예측 분류코드의 최종 공표 분류코드와의 일치 여부로 측정

○ **(기존 산업직업 자동코딩시스템과의 비교)**

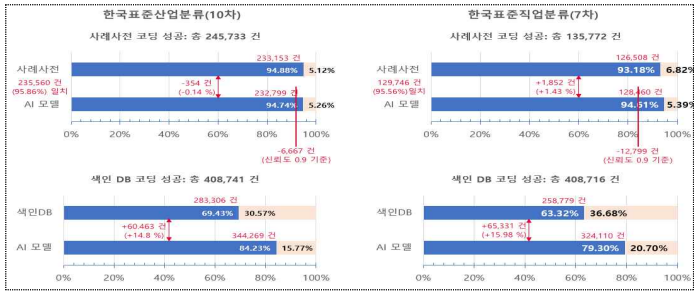
■ 정분류 ■ 오분류 ■ 미분류

분류	분류 방식	데이터 수 (건)	분류 (건) 정분류	분류 (건) 오분류	미분류 (건)	정/오/미 분류 비율 비교
산업분류	AI	409,901	**84.19%** (345,102)	15.81% (64,799)	0% (0)	
	색인DB		69.24% (283,806)	30.47% (124,908)	0.29% (1,187)	
	사례사전		56.88% (233,153)	3.06% (12,536)	40.06% (164,212)	
직업분류	AI		**79.64%** (326,446)	20.36% (83,455)	0% (0)	
	색인DB		63.13% (258,779)	36.41% (149,247)	0.46% (1,875)	
	사례사전		30.86% (126,508)	2.26% (9,261)	66.88% (274,132)	

- **산업직업 자동코딩시스템보다 AI 분류예측 정확도가 더 높음***

  * (시스템 분류예측 코드와 최종 분류코드가 같은 개수/409,901)로 비교

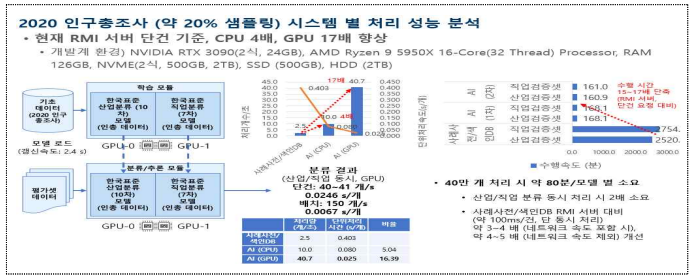
[참고: AI 분류 예측과 산업직업 자동코딩시스템 영역별* 비교]



* 산업직업 자동코딩시스템은 ①사례사전 기반 자동코딩과 ②색인DB 활용 코드추천 두 영역으로 구성 [상세내용은 별첨 산업직업 자동코딩시스템 현황 참고]

○ **(AI 분류예측 속도) 최신 GPU서버 활용으로 처리시간 대폭 단축***

* GPU 기준 처리속도 17배 향상 ⇨ 인총 조사기간 중 분류 예측결과 제공 가능



(2) 전국사업체조사 산업분류

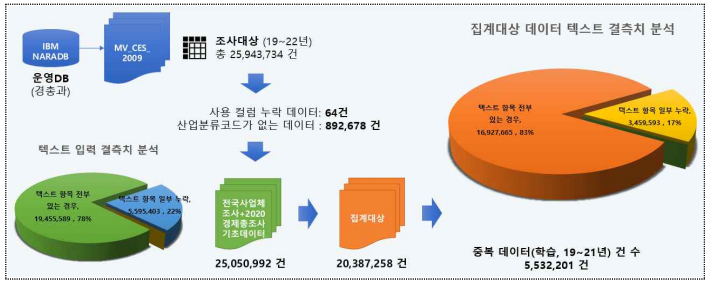
□ **분석방향**

○ AI 기반 통계분류 예측결과 정확도 확인

○ 현 산업직업 자동코딩시스템과 AI 통계분류 예측결과 비교 분석

□ **분석데이터**

○ '22년 전국사업체조사 공표(예정)자료 전체를 대상으로 분석 수행



- '19년~'21년 조사자료 17,822,619건을 학습·검증용 데이터로 사용*

  * 텍스트 미입력 및 중복데이터 제거 후 8,344,118건의 90% 학습용, 10%는 검증용

- **'22년 조사 집계대상 사업체 6,075,515건 전체를 평가셋**으로 활용

- 산업분류 조사항목* 중 일부 누락이 있는 경우도 포함

  *「사업체명, 무엇을 가지고, 어떤 방법으로, 생산·제공하였는가」4가지 항목

- 사업체명, 종사자 수, 사업자등록번호, 법인등록번호, 사업체 구분코드 정보 추가하여 분석

□ 분석결과

○ **(AI 분류예측 정확도) 산업분류 84.1%** 정확도* 확인

* 6,075,515건에 대한 AI 예측 분류코드의 공표 분류코드와의 일치 여부로 측정

활용데이터	언어모델 (Pretraining)	분류모델 (Finetuning)	Acc (%)	F1 (weight)	F1 (macro)	비고
전국사업체조사(19~22년) (25,943,670 건)	KLUE RoBERTa	HiBERT	검증: 87.56 평가: 81.05	검증: 0.873 평가: 0.808	검증: 0.693 평가: 0.655	22년 조사전체(7,228,373건) 기준
			검증: 88.71 평가: 84.15	검증: 0.8751 평가: 0.8393	검증: 0.7043 평가: 0.6809	22년 집계대상(6,075,515건) 기준

○ (기존 산업직업 자동코딩시스템과의 비교)

- 산업직업 자동코딩시스템보다 AI 분류예측 정확도가 더 높음*

1순위만 보여줄 경우

분류	분류방식	데이터 수 (건)	분류 예측율 (건, %)	미분류 (No value) (건, %)	정밀도 (Precision) (건, %)	오분류 (Error) (건, %)	정분류율 (Accuracy) (건, %)	미분류 및 오분류 합계 (건, %)	내검 예상치 (건)
산업분류	색인DB	6,075,515	6,053,016 99.63%	**22,499 0.37%**	4,467,322 73.80%	1,585,694 26.20%	**4,467,322 73.53%**	**1,608,193 26.47%**	-
	AI	6,075,515	6,075,515 100%	-	5,113,055 84.15%	962,460 15.84%	**5,113,055 84.15%**	**962,460 15.84%**	**1,819,640**

• AI 모델의 경우 신뢰도 90% 미만으로 예측한 데이터를 내검 대상으로 삼음
• 기존 방식의 경우, 미분류 +오분류 기준

3순위까지 보여줄 경우

분류	분류방식	데이터 수 (건)	분류 예측율 (건, %)	미분류 (No value) (건, %)	정밀도 (Precision) (건, %)	오분류 (Error) (건, %)	정분류율 (Accuracy) (건, %)	미분류 및 오분류 합계 (건, %)	내검 예상치 (건)
산업분류	색인DB	6,075,515	6,053,016 99.63%	**22,499 0.37%**	5,301,937 87.59%	751,079 12.41%	**5,301,937 87.27%**	**773,578 12.73%**	
	AI	6,075,515	6,075,515 100%	-	5,789,398 95.29%	286,117 4.71%	**5,789,398 95.29%**	**286,117 4.71%**	**484,406**

• 3순위 정답의 경우 신뢰도는 3순위 예측의 신뢰도들 합한 후 90% 미만 산출

- 1순위 추천 분류 정확도*: AI 84.1% vs 색인DB 73.5%

  * (시스템 1순위 분류예측 코드와 최종 분류코드가 같은 개수/6,075,515)로 비교

- 3순위 추천 분류 정확도*: AI 95.3% vs 색인DB 87.6%

  * (시스템 3순위 분류예측 코드 내 최종 분류코드가 있는 개수/6,075,515)로 비교

○ (자동/수동내검 기준 제공) 영역별/레코드별 예측정확도 정보 활용

정답률 상위 10개

	Precisio	Recall	F1-score	Supp	Korean_index
8721	0.99	0.99	0.99	4827	보육시설 운영업
9491	0.99	0.99	0.99	2626	종교 단체
8440	0.99	0.98	0.99	3893	사법 및 공공 질서 행정
9691	0.99	0.98	0.99	775	세탁업
8432	0.99	0.98	0.98	3475	국방 행정
7120	0.98	0.98	0.98	1500	회계 및 세무관련 서비스업
8610	0.98	0.98	0.98	1051	병원
9611	0.98	0.98	0.98	3747	이용 및 미용업
6110	0.98	0.98	0.98	814	공영 우편업
6513	0.97	0.98	0.97	415	사회보장보험업

정답률 하위 10개

	Precisio	Recall	F1-score	Supp	Korean_index
1630	0	0	0	14	코르크 및 조물 제품 제조업
5020	0	0	0	27	내륙 수상 및 항만 내 운송업
520	0	0	0	24	원유 및 천연가스 채굴업
610	0	0	0	7	철 광업
620	0	0	0	12	비철금속 광업
729	0	0	0	13	그 외 기타 비금속광물 광업
7622	1	0.17	0.29	6	음반 및 비디오물 임대업
3040	0.97	0.21	0.35	141	자동차 재제조 부품 제조업
3812	0.6	0.25	0.35	108	지정 폐기물 수집, 운반업
3191	0.91	0.23	0.37	43	전투용 차량 제조업

- AI 분류예측 정확도 확률값에 기초하여 내검원 업무량 관리 가능

  [예시] 조사별 예산 및 자료처리 기간 고려하여 자동내검/수동내검 비중 결정

  ① AI 분류 예측 정확도 95% 이상 산업분류 영역 ⇨ AI 분류코드 적용
  ② 예측 정확도 95% 이하 ⇨ AI 분류코드 3순위 추천 값 활용 수동내검 진행
  ③ 전주기 분류코드와 다른 경우 확인 내검 등 순차 진행 등

(3) 지역별고용조사 산업·직업분류

□ 분석방향

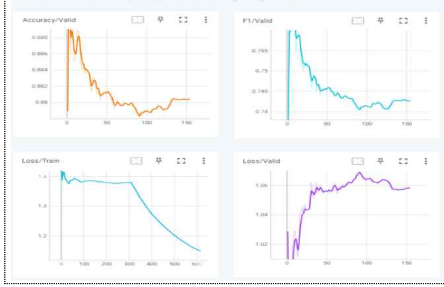
○ AI 기반 통계분류 예측결과 정확도 확인

○ 현 자동코딩시스템과 AI 통계분류 예측결과 비교 분석

□ 기초 분석결과

○ 산업분류 88%, 직업분류 83% 정확도의 학습모델 시험구축

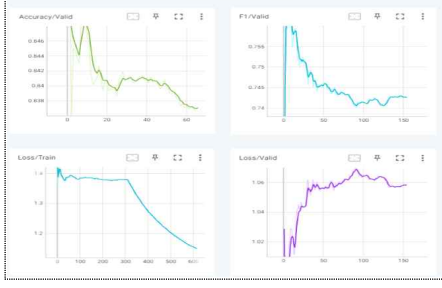
10차 산업 분류 코드(4자리) 학습 결과 A_IND_CD 기준



학습데이터 : 157,674 건
검증데이터 : 39,410 건
분류 종류: 489
Accuracy : 88%
F1-score : 74%
Loss/Train : 1.15
Loss/Valid : 1.058

산업 분류 작업(489항목)에서
정확도 88% 성능의 모델이 학습됨

7차 직업 분류 코드(4자리) 학습 결과 A_JOB_CD 기준



학습데이터 : 129,237 건 (49차)
검증데이터 : 14,349 건
분류 종류: 443
Accuracy : 83%
F1-score : 72%
Loss/Train : 1.15
Loss/Valid : 1.058

직업 분류 작업(443항목)에서
정확도 83% 성능의 모델이 학습됨

□ 추가 분석계획

○ 2022년 하반기 지역별고용조사 자료 분석 및 제공

- 과거 3개년(6회차) 공표자료를 학습데이터로 사용*

  * 산업직업 조사항목 통합모델 적용, 지역 성별 나이 등 추가 입력변수 활용

- 산업직업 자동코딩시스템과 AI 분류예측 결과 비교 분석*

  * 전체적인 분류예측 정확도 및 사례사전/색인DB 영역별 AI 분류 정확도 비교

(4) 건설업조사 공종·발주자분류

□ 분석방향

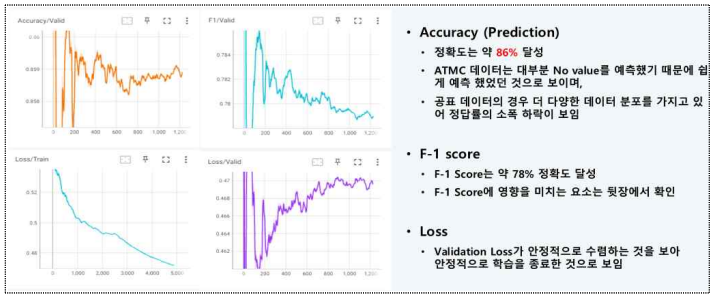
○ AI 기반 통계분류 예측결과 정확도 확인

○ 현 자동코딩시스템과 AI 통계분류 예측결과 비교 분석

□ 기초 분석결과

○ **공종분류 86%, 발주자분류 95% 정확도의 학습모델 시험구축***

* '17년~'21년 조사결과를 학습데이터로 사용(공종 6,328,076건, 발주자 5,887,909건)



□ 추가 분석계획

○ **'22년 건설업조사 공표(예정) 자료 분석**

- 과거 5개년 공표자료를 학습데이터로 사용*

  * '22년 최종 공표(예정)자료의 공종·발주자코드와 AI 분류예측 결과 비교 분석

- 공종·발주자분류 자동코딩시스템과 AI 분류예측 결과 비교 분석*

  * 전체적인 분류예측 정확도 및 사례사전과 AI 분류의 커버리지/정확도 비교 검토

(5) 온라인쇼핑조사 상품군별분류

□ 분석방향

○ AI 기반 통계분류 예측결과 정확도 확인

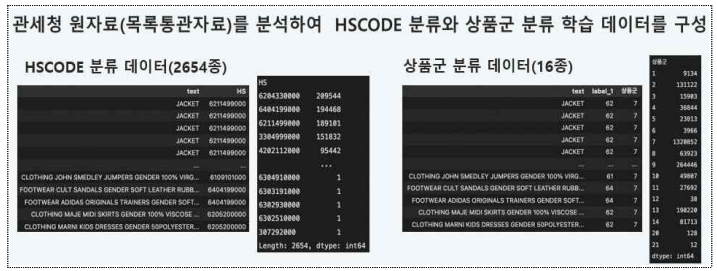
○ 현 자동코딩시스템과 AI 통계분류 예측결과 비교 분석

○ AI 통계분류 시스템 분류예측 수행속도 측정

□ 분석데이터

○ (월별) **관세청 목록통관자료* ⇨ HS코드(10자리) ⇨ 상품군별분류**

* 관세청 목록통관자료에는 HS코드 2자리만 존재 (월 200만~500만 건 자료)



○ **데이터 편향이 매우 큼**

- (HS코드) 상위 10개의 HS코드가 전체 데이터 비중의 55% 차지*

  * 6204330000(합성섬유직물제의류 8.7%), 6404199000(신발류 8.5%), 6211499000(기타직물제의류 6.8%), 3304999000(기타화장품류 4.3%), 4202112000(가죽제가방 4.0%)

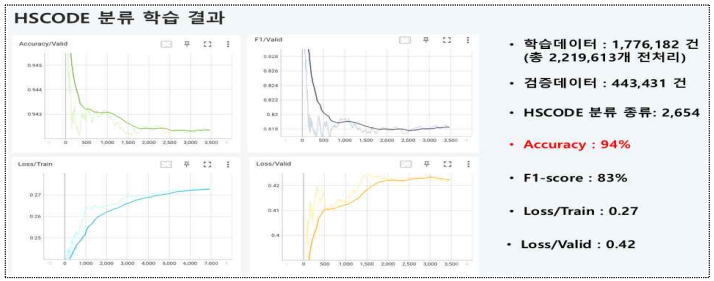
- (상품군별분류) 상위 5개 분류가 전체 목록통관자료의 89% 차지*

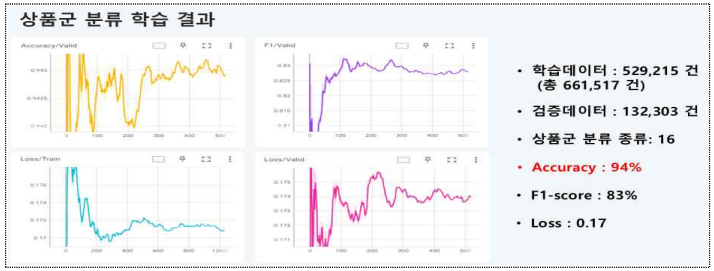
  * 7(의류및패션관련상품), 9(화장품), 13(생활및자동차용품), 2(가전전자통신기기), 14(가구)

○ **코드연계에 장시간 소요** ex) 시스템 부하로 200만 건 처리에 일주일 소요

□ 기초 분석결과

○ (현 자동코딩 시스템 대비) **94% 정확도의 학습모델 시험구축**





○ GPU서버 활용 시 **코드연계 시간 대폭 감소 가능***

* 현재 200만 건 코드연계 소요시간 7일 ⇨ GPU 활용 AI 분류처리시 4시간으로 단축

		HSCODE 분류		상품군 분류	
활용데이터	모델	acc	F1 (Macro)	acc	F1 (Macro)
AI기초데이터(text, HS 활용)	HiBERT	94%	83%		
AI기초데이터 (text, HS, 상품군 활용)	FlatRoberta			94%	83%
AI기초데이터(text, HS(변환), 상품군 활용)	FlatRoberta			학습 예정	학습 예정

***10만건의 데이터 처리 약 10분이 소요**

*HS코드를 상품군 text로 변경하여 학습 예정
**HSCODE분류 6자리 예측으로 학습 후 상품군 분류에 활용가능
*** 학습데이터 부족으로 정답을 못 맞춘 분류의 장표는 제외함

□ 추가 분석계획

○ '22년 온라인쇼핑조사 공표(예정) 자료 분석

○ AI 통계분류 시스템 분류예측 수행속도 측정

Ⅳ 향후 추진계획

2020년~2021년	2022년	2023년	2024년
[기초연구] AI 통계분류 자동화 시험 분석 (지고,건설업조사)	**[시스템 구축]** AI 통계분류 시스템 개발 (과기부 지원사업)	**[시스템 시험 적용]** AI 통계분류결과 실무 시험적용 (5개 분류, 5종 조사)	**[시스템 확산]** AI 통계분류결과 적용대상 확대 (나라시스템 적용)

□ 인공지능 통계분류 자동화 시스템 시험적용

○ AI 통계분류 결과를 일부 조사통계 업무에 시험 적용하여 통계분류 자료처리의 효율성 및 분류 정확도 제고 지원

- 적용분류(3종): 산업·직업분류, 공종·발주자분류, 상품군별분류
- 대상조사(5종): 인총시험조사[산업직업분류], 전국사업체조사[산업분류], 지역별고용조사[산업직업분류], 건설업조사[공종발주자분류], 온라인쇼핑조사[상품군별분류]

□ 인공지능 통계분류 자동화 시스템 고도화 및 확산

○ AI 통계분류 예측 정확도 향상을 위한 학습데이터 품질 제고 노력*

* 분류 예측 정확도 하위 분류영역에 대한 재검토 및 검토 결과 재학습 등

○ AI 통계분류 자동화 적용 대상 분류 및 조사 확대 검토

- AI 통계분류 적용 대상 조사 확대 가능성 분석 (산업·직업분류)

. * 현재 AI 시스템은 전사체, 인총, 지역별고용조사 3종의 조사 대상으로 개발

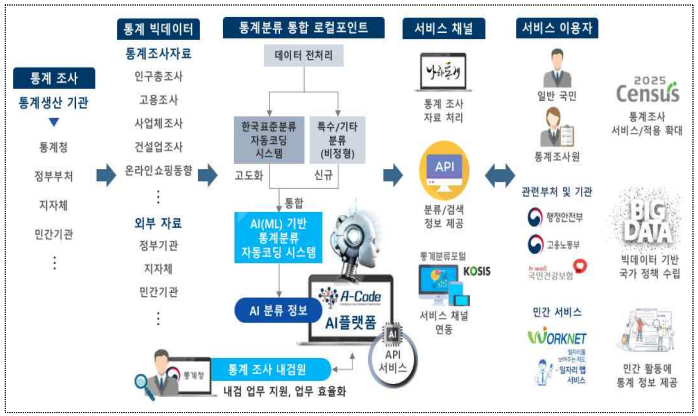
주기	담당과	조사명
사업체 대상 조사	경제총조사과	경제총조사, **전국사업체조사**
	산업통계과	광업제조업조사, 도소매서비스업조사, 프랜차이즈업조사
	경제통계기획과	소상공인실태조사
가구 대상 조사	인구총조사과	**인구주택총조사**
	고용통계과	경제활동인구조사, **지역별고용조사,** 이민자체류실태조사
	사회통계기획과	사회조사, 생활시간조사
	가계수지동향과	가계동향조사
	복지통계과	가계금융복지조사, 초중고사교육비조사

- AI 적용 대상 통계분류 확대 가능성 분석 (질병·사인분류, 행동분류 등)

□ AI 통계분류 시스템 안정화를 위한 예산·조직 확보 노력

V 기대효과

[참고: 범용 인공지능 통계분류 시스템 활용도(안)]



□ **정확한 통계분류 서비스 제공**

○ 통계청이 보유한 방대한 데이터를 AI 기술을 활용하여 학습

○ 높은 정확도의 일관된 통계 분류 예측 정보 제공

* 인공지능 예측 정확도는 학습데이터의 크기 및 품질에 의해 결정

□ **신속한 통계분류 서비스 제공**

○ AI에 특화된 최신 GPU서버 활용하여 자료처리 기간 단축*

○ 조사기간 및 내검 작업시 실시간/배치 통계분류 예측 결과 신속 제공

* 3세대 Ampere 아키텍처로 현재 GPU서버 중 가장 강력한 성능의 A100 2식 도입

□ **효율적인 통계분류 자료처리 지원**

○ AI 통계분류 예측값의 확률정보를 토대로 자동/수동내검 체계화*

ex) AI 통계분류 예측 정확도 확률값이 95% 이상인 경우 자동코딩 분류코드 적용
그 외는 AI 통계분류 예측 정확도 3순위 분류코드 대상 수동내검 진행 등

□ **범용 인공지능 통계분류 플랫폼으로 확장**

○ Open API 서비스로 AI 통계분류 결과 통계작성기관 등에 제공*

ex) 향후 나라통계시스템, 통계분류포털 등의 채널을 통한 서비스 확대 추진

Abstract

An analysis of AI statistical classification results and a study on the improvement of practical utilization

GyeongMin Im

In the era of the 4th Industrial Revolution, the importance of data science is being emphasized as a methodology for the quick and accurate production of information by linking big data with IT. Recently there has been increasing demand for the application of data science techniques such as artificial intelligence (AI) to national statistics.

In the area of national statistics, 'classification coding automation using statistical classification text information' where AI technology can be widely used is attracting attention. Many countries, including the United States, are making efforts to apply atypical text classification automation using natural language processing technology to the statistical production process.

As part of data science research to respond to changes in statistical environment in the future, Statistics Korea has also conducted research on the automated statistical classification coding, using natural language processing and supervised learning-based machine learning classification models since 2020. Based on the research, Statistics Korea developed the 'automated artificial intelligence statistical classification system' in 2022 and applied this system to some surveys on a pilot basis.

Abstract

This study examined the progress of data science research that the Statistical Research Institute conducted, and introduced the development results of the artificial intelligence-based statistical classification methodology and the automated artificial intelligence statistical classification system. In addition, focused on the classification cases of construction types and orderers in the Construction Survey, the application results of AI-based statistical classification were analyzed. And this study examined the methods to provide practical usability to more effectively utilize AI statistical classification prediction results.

These analysis results are expected to contribute to improving the accuracy, consistency, and timeliness of Statistics Korea's statistical classification data processing, and also contribute to improving administrative efficiency through automation and intelligence of repetitive tasks. Besides, by pioneering specialized areas of national statistics where natural language processing technology can be applied and verifying related technologies, this study is expected to contribute to the dissemination and spread of 'the automated classification through atypical text analysis' in the future.

Keywords: artificial intelligence and national statistics, natural language processing, classification automation

연구진

○ 임경민 (통계청 통계개발원 통계방법연구실 사무관)

* 연구진의 소속 및 직급은 연구과제 완료 시 기준임을 알려드립니다.

연구보고서 2023-16

AI 통계분류 결과분석 및 실무활용성 제고방안 연구

- 건설업 공종·발주자분류 사례 검토-

인 쇄	2024년 4월
발 행	2024년 4월
발 행 인	박상영
발 행 처	통계청 통계개발원 35220 대전광역시 서구 한밭대로 713 TEL.(042)366-7100 Fax.(042)366-7123
홈페이지	http://sri.kostat.go.kr
ISSN(Online)	2733-4120

통계청
통계개발원