국가데이터처
국가데이터연구원

Ministry of Data and Statistics(MODS)

Coding classification with AI



Data Science Research Team, DSRI

IM GYEONG MIN, Deputy Director

국가데이터처
국가데이터연구원

CONTENTS

I. Current Practice in Statistical Classifications

II. AI-based Classification System Development

III. Achievements and Future Steps



국가데이터처
국가데이터연구원

I

Current Practice in Statistical Classifications

Importance of Statistical Classifications

국가데이터처
국가데이터연구원

Statistical classifications enabling the detailed mapping of information by "region", "industry sector", and more

시·도별	사업체 수				종사자 수			
	2020년	2021년	증감	증감률	2020년	2021년	증감	증감률
Nationwide	6,032,022 (100.0)	6,075,912 (100.0)	43,890 (0.0)	0.7	24,813,449 (100.0)	24,992,505 (100.0)	179,056 (0.0)	0.7
Capital areas	2,972,805 (49.3)	2,976,201 (49.0)	3,396 (−0.3)	0.1	12,964,045 (52.2)	13,037,811 (52.2)	73,766 (−0.1)	0.6
Non-capital areas	3,059,217 (50.7)	3,099,711 (51.0)	40,494 (0.3)	1.3	11,849,404 (47.8)	11,954,694 (47.8)	105,290 (0.1)	0.9
Seoul	1,211,053 (20.1)	1,187,013 (19.5)	−24,040 (−0.5)	−2.0	5,868,926 (23.7)	5,816,141 (23.3)	−52,785 (−0.4)	−0.9
Busan	402,003 (6.7)	401,151 (6.6)	−852 (−0.1)	−0.2	1,537,281 (6.2)	1,543,764 (6.2)	6,483 (0.0)	0.4
Daegu	283,033 (4.7)	279,727 (4.6)	−3,306 (−0.1)	−1.2	1,010,557 (4.1)	1,009,758 (4.0)	−799 (0.0)	−0.1
Incheon	306,108 (5.1)	308,768 (5.1)	2,660 (0.0)	0.9	1,208,269 (4.9)	1,228,483 (4.9)	20,214 (0.0)	1.7
	170,085 (2.8)	170,958 (2.8)	873 (0.0)	0.5	667,435 (2.7)	675,942 (2.7)	8,507 (0.0)	1.3
	164,406 (2.7)	164,038 (2.7)	−368 (0.0)	−0.2	691,264 (2.8)	688,267 (2.8)	−2,997 (0.0)	−0.4
	117,247 (1.9)	115,389 (1.9)	−1,858 (0.0)	−1.6	543,424 (2.2)	549,102 (2.2)	5,678 (0.0)	1.0
	28,490 (0.5)	30,481 (0.5)	1,991 (0.0)	7.0	139,731 (0.6)	152,325 (0.6)	12,594 (0.0)	9.0
	1,455,644 (24.1)	1,480,420 (24.4)	24,776 (0.2)	1.7	5,896,850 (23.7)	5,993,187 (24.0)	106,337 (0.3)	1.8
	193,074 (3.2)	200,252 (3.3)	7,178 (0.1)	3.7	704,983 (2.8)	713,613 (2.9)	8,630 (0.0)	1.2
	191,265 (3.2)	194,446 (3.2)	3,181 (0.0)	1.7	808,018 (3.3)	817,157 (3.3)	9,139 (0.0)	1.1
	253,192 (4.2)	260,953 (4.3)	7,761 (0.1)	3.1	1,064,810 (4.3)	1,078,195 (4.3)	13,385 (0.0)	1.3
	225,964 (3.7)	231,084 (3.8)	5,120 (0.1)	2.3	794,929 (3.2)	790,945 (3.2)	−3,984 (0.0)	−0.5
	228,219 (3.8)	234,151 (3.9)	5,932 (0.1)	2.6	847,692 (3.4)	856,594 (3.4)	8,902 (0.0)	1.1
	321,061 (5.3)	328,539 (5.4)	7,478 (0.1)	2.3	1,225,829 (4.9)	1,240,651 (5.0)	14,822 (0.0)	1.2
	387,177 (6.4)	392,500 (6.5)	5,323 (0.0)	1.4	1,494,560 (6.0)	1,517,833 (6.1)	23,273 (0.0)	1.6
	94,001 (1.6)	96,042 (1.6)	2,041 (0.0)	2.2	318,891 (1.3)	320,548 (1.3)	1,657 (0.0)	0.5
...								

산업별	2020년	구성비	2021년	구성비	증감	증감률
A Agriculture, forestry and fishing	6,032,022	100.0	6,075,912	100.0	43,890	0.7
B. Mining & quarrying	12,707	0.2	11,409	0.2	−1,298	−10.2
C. Manufacturing	2,205	0.0	2,097	0.0	−108	−4.9
D Electricity, gas, steam	579,645	9.6	579,363	9.5	−282	0.0
E Water supply, sewage, waste management	74,092	1.2	88,491	1.5	14,399	19.4
F Construction	13,097	0.2	13,389	0.2	292	2.2
G Wholesale and retail trade	471,217	7.8	484,909	8.0	13,692	2.9
H Transportation and storage	1,567,298	26.0	1,536,031	25.3	−31,267	−2.0
I Accommodation and food service activities	593,274	9.8	616,884	10.2	23,610	4.0
J Information and communication	865,333	14.3	862,544	14.2	−2,789	−0.3
K Financial and insurance activities	113,304	1.9	120,425	2.0	7,121	6.3
L Real estate activities	64,108	1.1	63,054	1.0	−1,054	−1.6
M Professional, scientific and technical activities	278,339	4.6	284,479	4.7	6,140	2.2
N Business facilities management and business support services	221,460	3.7	222,951	3.7	1,491	0.7
O Public administration	151,264	2.5	142,151	2.3	−9,113	−6.0
P Education (schools and universities, etc.)	12,480	0.2	12,578	0.2	98	0.8
Q Human health and social work activities	234,741	3.9	248,928	4.1	14,187	6.0
R Arts, sports, and recreation related services	157,988	2.6	161,490	2.7	3,502	2.2
S Membership organizations, repair and personal services	143,161	2.4	143,454	2.4	293	0.2
	476,309	7.9	481,285	7.9	4,976	1.0

Statistical Classifications in Surveys
(In the case of Industrial Classifications)

국가데이터처
국가데이터연구원

The classification codes filled in by surveyors based on text information provided by respondents

- Industrial classifications categorization criteria:
  Business activity generating the most value-added, whether it be producing goods or providing services
- Industrial classifications hierarchical structure: Sections (21), Divisions (77), Groups (234), Classes (501), Subclasses (1,205)

< Questions for industrial classifications in a survey of establishments>

⑧ Types of business (main and incidental business lines)	• List your business lines in the order of sales amounts when incidental business lines are operated.				
	Where? (Place where business activities are conducted)	How? (Business means and procedures)	What is produced or provided? (Types of goods produced or services provided)	Percentage in sales(%)	★ Industrial classification code
Main business					
Sub-business 1					
Sub-business 2					
Sub-business 3					
★ Description of the individual classification of main business:					
★ Items to which special attention should be paid:					

통계청 Statistics Korea

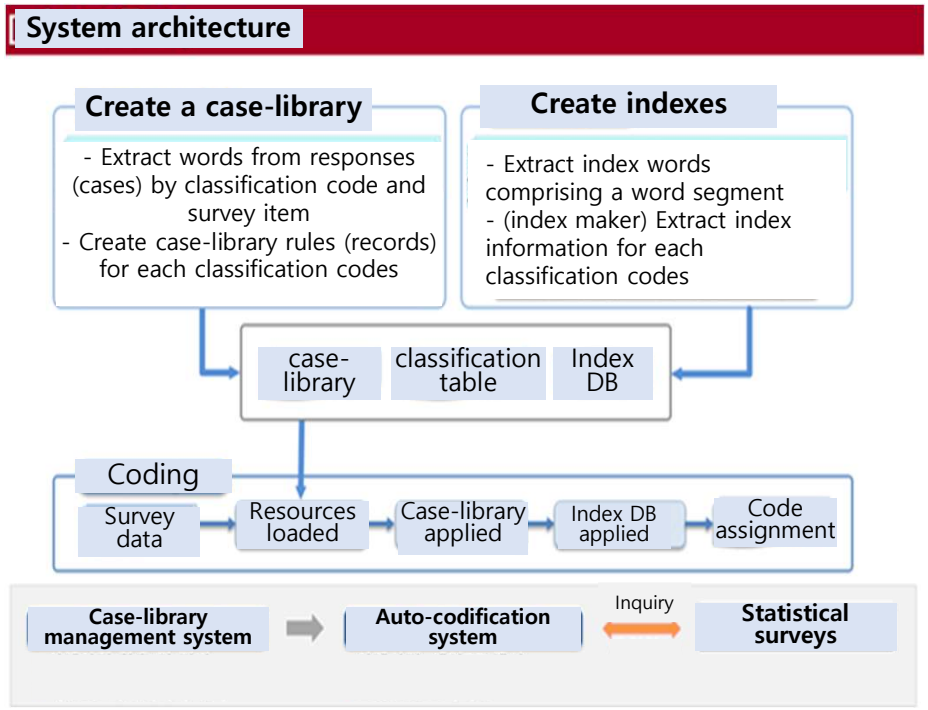
– 1 –

2020 Census

Data Editing for Statistical Classifications
(In the case of Industrial Classifications)

국가데이터처 국가데이터연구원

Limitation in computerized editing when using the legacy automatic coding systems



- **Case-library**: Editing rules derived from the professional judgement of skilled managers are applied automatically => high accuracy (95% or more) but low coverage (less than 60%)

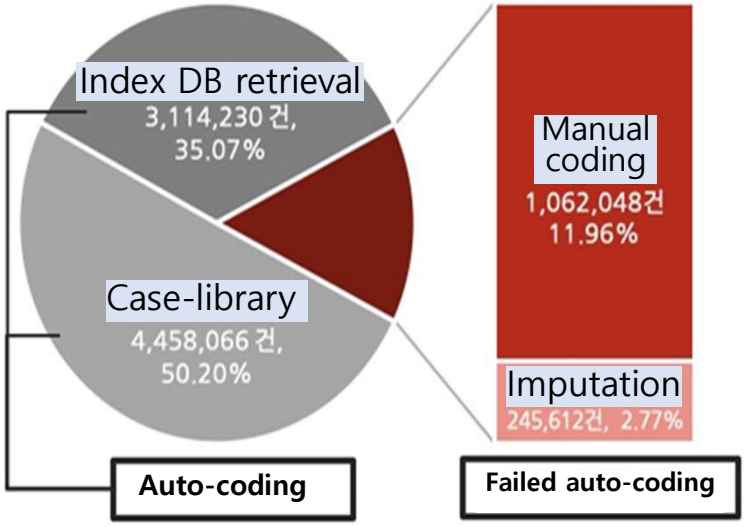
- **Index DB**: **Classification codes are determined based on scores calculated from frequency information for specific index words** => high coverage (99% or more) but low accuracy (less than 70%)

I Data Processing for Classification Items in Surveys

국가데이터처 국가데이터연구원

Substantial requirement for workforce and time to review statistical classification codes resulting from text-based classification information

- Limitations, such as the difficulty of fully understanding the vast scope of statistical classifications, hamper the accuracy of assigning and reviewing classification codes during field survey operations.
- Despite the use of a rule-based statistical coding system, finalizing classification codes often requires substantial human and time resources.



01 Importance of statistical classifications

- Code assignment relies heavily on text-based input.
- Statistical classification codes enable the detailed mapping of regional and industry-specific conditions.

02 Limitations of text analysis in existing rule-based systems

- A case-library approach is time-consuming and costly, with limited coverage.
- An index DB approach is unable to save labor in editing due to its low accuracy.

03 Substantial resources for processing statistical classification-related data

- Finalizing classification codes ends up requiring substantial labor and time.
- Processing large-volume data, particularly in cases such as census, is especially burdensome.



국가데이터처
국가데이터연구원

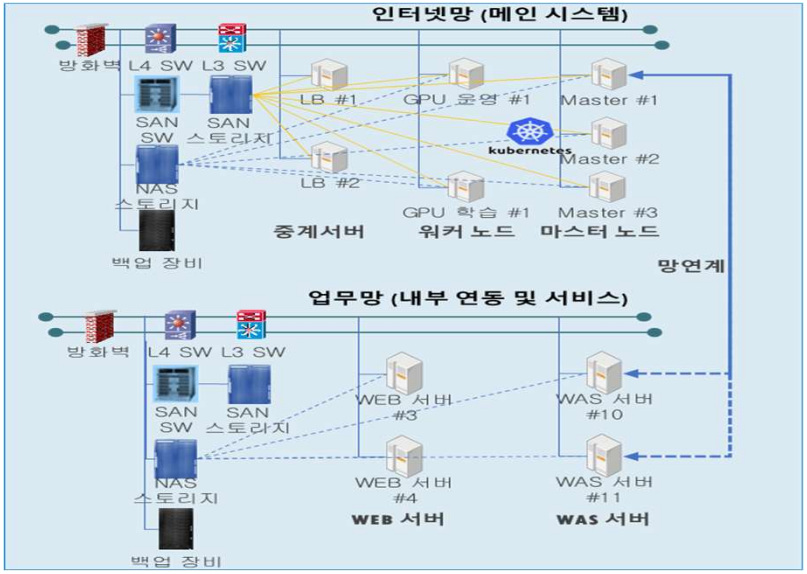
II

AI-based Classification System Development

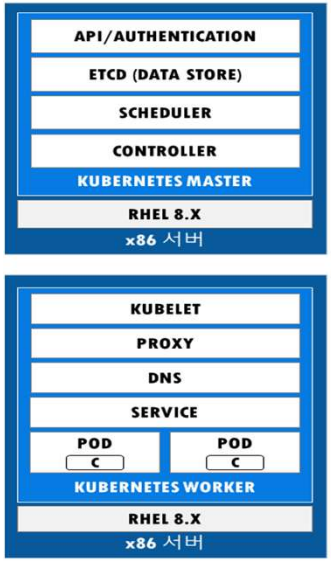
Development of AI-based statistical classification automation system (2022.7.~2023.1.)

- A development budget of 15.5 billion won was funded by the "Digital Public Service Innovation Project" of the Ministry of Science and ICT in 2022.
- The scope of development included 5 classification systems used in the 5 major surveys: the industrial/occupational classifications for the Regional Employment Survey, Population Census, Census on Establishments; the classifications for construction operations/clients for the Construction Survey; and the product classifications for the Online Shopping Survey

< Hardware design & structure >



< Software design & structure >



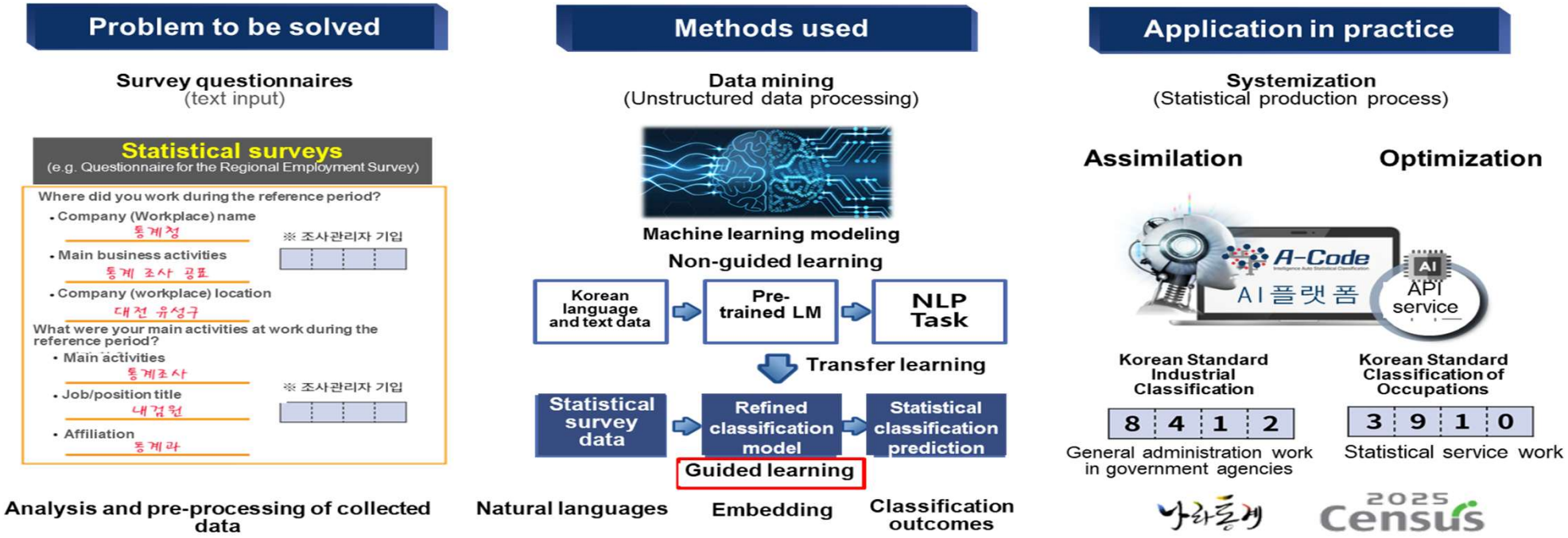
구분	구성	내용
NODE (Master)	Kubernetes Master	사용자 로그인, 애플리케이션 관리, 도메인 생성 등 모든 Kubernetes 명령 관장
	Kube-apiserver	Kubectl(관리 명령어 콘솔) RESTful API 처리
	ETCD	Kubernetes 클러스터 전체 정보에 대한 KEY-VALUE 저장소
	Kube-Scheduler	신규 POD 를 감지하고, 노드별 리소스 사용량에 따른 POD 배정
	Kube-controller-manager	POD(컨테이너) Lifecycle 관리
NODE (Worker)	Kubernetes Worker	컨테이너화된 서비스 파드(POD) 를 배치하고 실행
	Kubelet	컨테이너 정의에 따라 서비스가 건강하게 동작하는 것을 관리
	Kube-proxy	노드의 네트워크 규칙을 관리 내부 네트워크 세션이나 클러스터 바깥에서의 통신 관리
	DNS	파드(POD) 및 서비스를 위한 DNS 레코드 서버
	Service	실행중인 애플리케이션을 네트워크 서비스로 노출시키는 추상화 객체 (예. Loadbalancing)
	POD	Kubernetes 플랫폼에서 구분하는 컨테이너 최소 단위

II System Overview and Main Features

국가데이터처 국가데이터연구원

Overview of AI-based automatic statistical classification system

- Incorporating ① natural language processing (a pre-trained LM approach) and ② machine learning (a guided learning approach, the system processes relevant text input to generate predictive statistical classification codes and/or probabilities as an outcome.

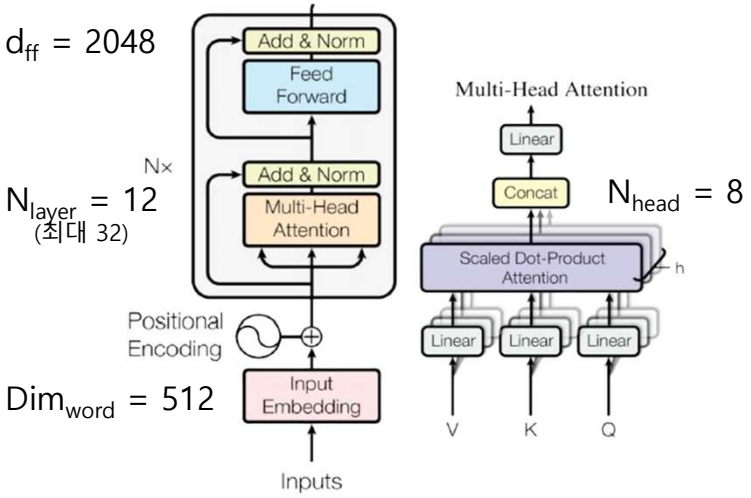


AI-based automatic statistical classification methodology (1): Language model choice

- Using RoBERTa* of Facebook as its language model which stemmed from Google's BERT model

  * Proven to work well to process a short sentence

- **BERT [Jacob Devlin, et al. 2018]**
  - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
  - Learns relational information between related words(tokens) using an Attention mechanism under a Bidirectional Transformer architecture



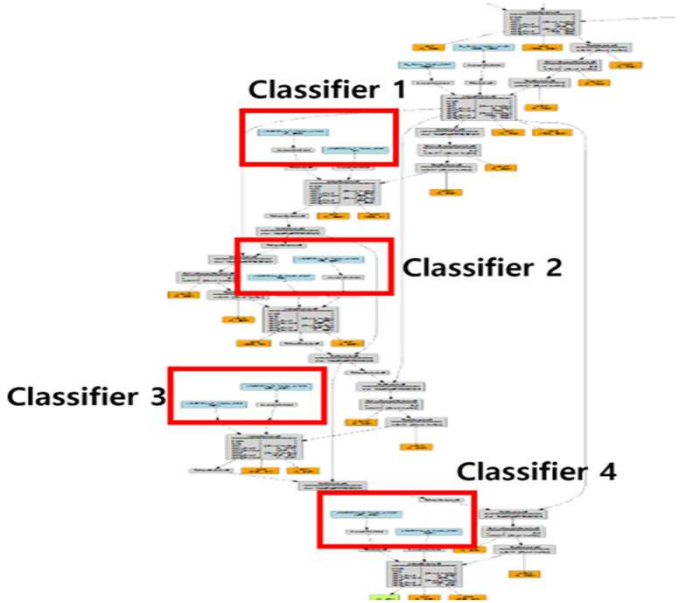
- **RoBERTa [Yinhan Liu, et al, 2019]**
  - Achieved an optimal method through experiments assessing whether the model's hyper-parameters affect performance for general tasks
  - Multi-task learning using various training datasets for multiple NLP task
  - Variations derived from the BERT model depending on the training corpus and hyper-parameter settings
  - Variation models (used in our system)
    - Multilingual parallel corpora: xlm-RoBERTa [Google. 2020]
    - Korean corpora: KLUE-RoBERTa [Park et al, 2021]
  - Classifications according to hyper-parameter setting

Model	Embedding size	Hidden size	# Layers	# Heads
KLUE-BERT-base	768	768	12	12
KLUE-RoBERTa-small	768	768	6	12
KLUE-RoBERTa-base	**768**	**768**	**12**	**12**
KLUE-RoBERTa-large	1024	1024	24	16

AI-based automatic statistical classification methodology (2): Choice of a guided-training method

- Using a Hierarchical Text Classification(Hi-TC) model due to the nature of hierarchical structure of statistical classifications i.e. section-division-group-class-subclass

Hierarchical Text Classification (Hi-TC) model



Example) Model structure for an industry/occupational classification system (in the 'subclass' level)

- **Basic structure**
  - A guided-training model designed specifically to process statistical data, enabling finetuning of text classification tasks
  - Utilizes a hierarchical structing of text classification pipelines consisting of sequential classifiers

- **Model characteristics**
  - Allows the use of various pre-trained models within the BERT family
  - Learns and predicts classification codes for lower-level categories by considering both natural language inputs and the classification outcomes in higher-level categories (e.g. Domain>Subdomain>Class>Subclass>Specific Item)

- **Strengths**
  - Fit-for-purpose design to handle real-problem classification, such as standard statistical classifications, involving a system of hierarchical categories or classes
  - Model's application levels cab be adjusted according to classification systems used in different surveys

System Overview and Main Features

Construction of classification models (5 classification systems, 5 surveys) and their performance evaluation results

- Assessing the accuracy of AI-predicted classifications on the basis of a total of 167 million text records from 5 surveys referencing 5 statistical classification systems. The text data were used for guided learning and evaluation following a series of pre-processing operation, including the removal of duplicates. It was aimed to achieve a 85% accuracy rate for the system development.

Classification name	Unit-in-charge	Target survey (cycle)	Approx. data counts	Basic data counts	Data pre-processing	Data refinement	Data subset	Accuracy / F-1	Operation wave (last updated)
10th Korean Standard Industrial Classification (KSIC)	Labor Statistics Div.	Regional employment survey (biannual)	0.43 mil. (6 month intervals)	3.47 mil. (8 survey rounds btw. 2019-22)	Missing: 1.09 mil. Duplication: 0.32 mil.	1.22 mil.	Learning: 1.10 mil. Validation: 0.12 mil. Evaluation: 0.43 mil.	Validation: 92.74% / 0.9271 Evaluation: 94.17% / 0.9414	4 (Dec.14)
	Economic Census Div.	Establishments Census (annual)	8.00 mil. (1-year intervals)	25.94 mil. (btw. 2018-21)	M: 4.84 mil. D: 5.53 mil.	8.34 mil.	L: 7.51 mil. V: 0.83 mil. E: 7.23 mil.	V: 85.04% / 0.8508 E: 86.48% / 0.8623	9 (Jan.11)
	Population Census Div.	Population Census (quinquennial)	10.00 mil. (5-year intervals)	9.81 mil. (2020 census round)	Processing: 5.71 mil. M: 0.40 mil. D: 0.65 mil.	4.10 mil.	L: 3.69 mil. V: 0.41 mil.	V: 86.01% / 0.8610 E: 85.60% / 0.8551	6 (Nov.6)
7th Korean Standard Classification of Occupations (KSCO)	Labor Statistics Div.	Regional employment survey (biannual)	0.43 mil. (6 month intervals	3.47 mil. (8 survey rounds btw. 2019-22)	M: 1.09 mil. D: 0.32 mil.	1.22 mil.	L: 1.10 mil. V: 0.12 mil. E: 0.43 mil.	V: 91.44% / 0.9141 E: 92.67% / 0.9259	5 (Dec.14)
	Population Census Div.	Population Census (quinquennial)	10.00 mil. (5-year intervals)	9.81 mil. (2020 census round)	P: 5.71 mil. M: 0.40 mil. D: 0.65 mil.	4.10 mil.	L: 3.69 mil. V: 0.41 mil.	V: 79.52% / 0.7915 E: 79.25% / 0.7896	6 (Nov.6)

II Performance Evaluation Results

국가데이터처 국가데이터연구원

Performance compared between existing system and AI-based statistical classification model: Degree of accuracy

Non-classification
- Case-library: (KSIC) 40.06%, (KSCO) 66.88%
- Index DB: (KSIC) 0.29%, (KSCO) 0.46%
- AI model: 0%

Precision (excl. non-classification)
- Case-library: (KSIC) 94.90%, (KSCO) 93.18%
- AI model: (KSIC) 84.20%, (KSCO) 79.64%
- Index DB: (KSIC) 69.44%, (KSCO) 63.42%

Accuracy (incl. non-classification) (=recall)
- AI model: (KSIC) 84.20%, (KSCO) 79.64%
- Index DB: (KSIC) 69.24%, (KSCO) 63.13%
- Case-library: (KSIC) 56.89%, (KSCO) 30.86%

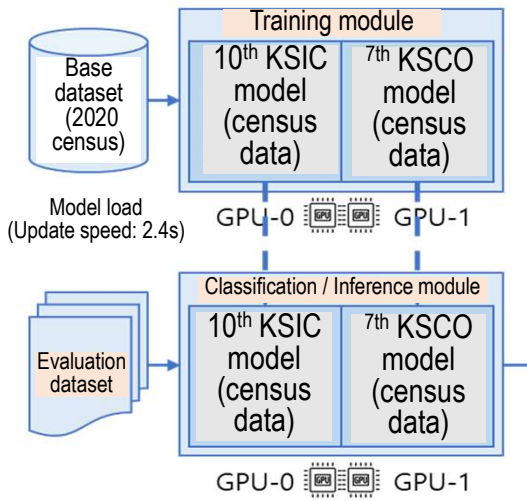
Classification type	Applied method	Data counts	Outcome (counts)		Non-classification	Ratio of correct (blue bar) /incorrect (pink bar) /non- classifications (grey bar)
			Correct classes	Incorrect classes		
Korean Standard Industrial Classification (KSIC)	AI model	409,901	**84.19%** (345,102)	15.81% (64,799)	0% (0)	
	Index DB		69.24% (283,806)	30.47% (124,908)	0.29% (1,187)	
	Case-library		56.88% (233,153)	3.06% (12,536)	40.06% (164,212)	
Korean Standard Classification of Occupation (KSCO)	AI model		**79.64%** (326,446)	20.36% (83,455)	0% (0)	
	Index DB		63.13% (258,779)	36.41% (149,247)	0.46% (1,875)	
	Case-library		30.86% (126,508)	2.26% (9,261)	66.88% (274,132)	

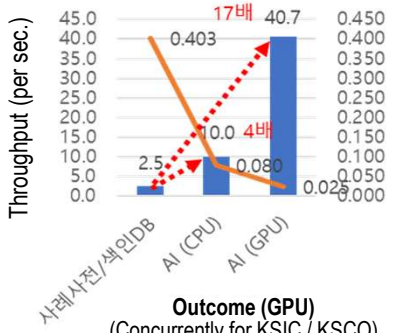
* Analysis of the 2020 Population Census data for their industry and occupation classification outcomes

Performance Evaluation Results

Evaluation for the processing speed of classification model

- **Performance analysis of processing the 2020 Population Census (with 20% sampling) data by system type**
  - Compared to the incumbent RMI server, a single instance processing time improved 4 times for CPU and 17 times for GPU
  - Dev. environment) NVIDIA RTX 3090(2식, 24GB), AMD Ryzen 9 5950X 16-Core(32 Thread) Processor, RAM 126GB, NVME(2식, 500GB, 2TB), SSD (500GB), HDD (2TB)



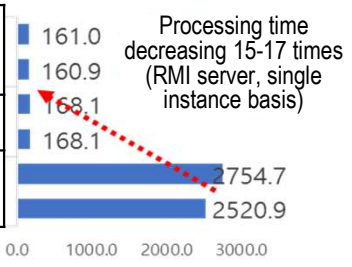


Outcome (GPU)
(Concurrently for KSIC / KSCO)
Single instance: 40-41 instance/s
0.0246 s/count
Multiple instances: 150 counts/s
0.0067 s/count

	Throughput (# of instance/s)	Latency (s / single instance)	Ratio
사례사전/색인DB	2.5	0.403	
AI (CPU)	10.0	0.080	5.04
AI (GPU)	**40.7**	**0.025**	**16.39**

Latency (s / count)

AI (2nd round)	Occupation verification Industry verification
AI (1st round)	Occupation Industry
Case-library / Index DB	Occupation Industry



80 min. to process 400k instances

- Inference time doubles to process KSIC and KSCO concurrently.
- Improvement from existing system (approx. 100ms/instance, concurrent basis)
  - 3-4 times when network speed included
  - 4-5 times when network speed excluded



국가데이터처
국가데이터연구원

III

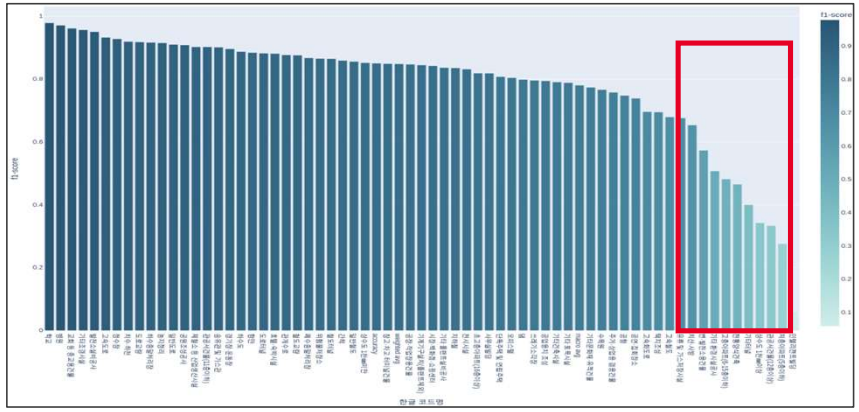
Achievement and Future Work

III Assessment of the AI-based Automated Classifications Pilot Operations

국가데이터처 국가데이터연구원

Outcome of the pilot operation of AI-based statistical classification automation

- Human-In-The-Loop approach: Involving an additional process in which human i.e. a reviewer engages in verifying and editing AI-predicted statistical classification output for ensuring data quality (See *Machine Learning for Official Statistics* (UNECE,2022))



	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

Selective editing
→ Determine editing priorities

1. Edit if certain classes showed poor prediction results
2. Edit if each specific case showed poor prediction results
3. Edit if a case had non-matching results compared to the outcomes of the existing rule-based automatic coding

Reinforcement and re-learning of editing outcomes
→ To improve prediction accuracy of AI classification

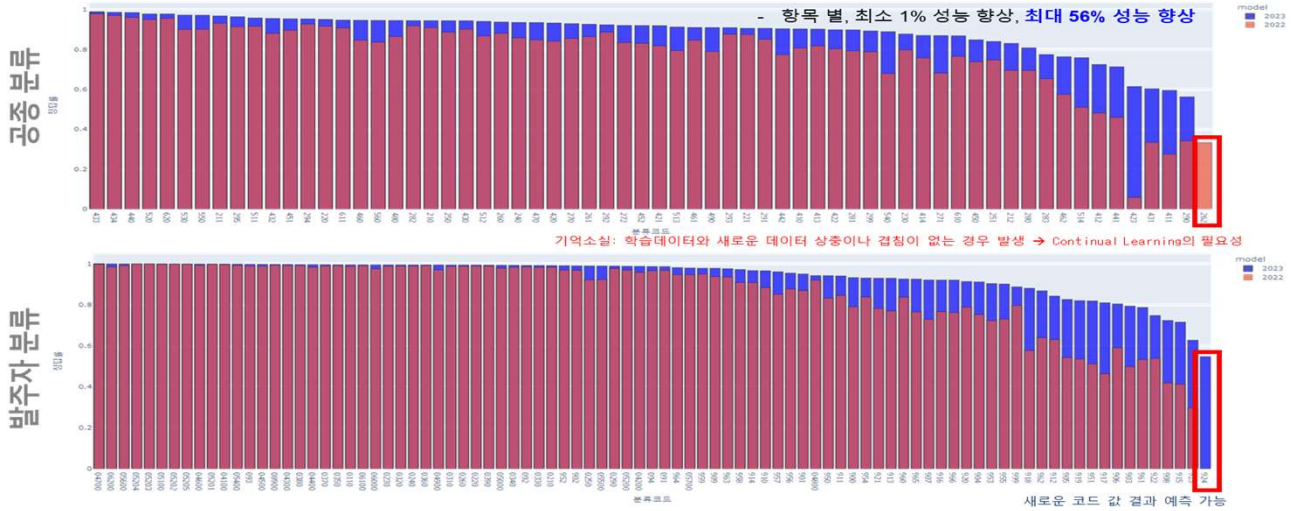
III Assessment of the AI-based Automated Classifications Pilot Operations

국가데이터처 국가데이터연구원

Improving AI model's inference by conditioning training data

- Updating the training datasets i.e. by incorporating corrections and refinements made by human reviewers and running a re-learning operation (without any refinement of LM or classification algorithm) has proven to improve the accuracy of the system, indicating potentials for reducing a workload for human reviewers over recurring cycles.

- At least 1% better performance across all variables
- Up to 56% better performance



공종 분류 (평가: 21년 내검 데이터 921,577건)

< 22년 구축 모델 >	
구분	수치
정분류율	85.05%
F1-macro	0.7796
학습데이터	18~20년 내검 데이터

선택적 에디팅 기존+신규 내검 데이터 학습

< 23년 고도화 모델 >	
구분	수치
정분류율	92.76% △7.7%
F1-macro	0.8757 △0.096
학습데이터	(기존) 18~20년 내검 데이터 + (추가) 21년 1~3차 내검 데이터

발주자 분류 (평가: 21년 내검 데이터 922,053건)

< 22년 구축 모델 >	
구분	수치
정분류율	88.96%
F1-macro	0.8530
학습데이터	18~20년 내검 데이터

선택적 에디팅 기존 + 신규 내검 데이터 학습

< 23년 고도화 모델 >	
구분	수치
정분류율	95.62% △6.7%
F1-macro	0.9410 △0.088
학습데이터	(기존) 18~20년 내검 데이터 + (추가) 21년 1~3차 내검 데이터

Future Work

Development of a generic platform for AI-based statistical classification automation

- Expanding the scope of the AI-based classification system beyond the current coverage of five surveys and five classification systems

< Statistics produced by KOSTAT based on industrial classifications >

구분	조사명	공표 주기	조사 대상	산업분류 조사항목
사업체	경제총조사	5년	390만	[세세분류단위] - 무엇을 가지고 (원재료, 영업장소 등) - 어떤 방법으로 (주요 영업, 생산 활동) - 생산·제공하였는가 (최종 재화, 용역)
	Census of Establishments	1년	490만	
	광업제조업조사	1년	7만5천	
	도소매서비스업조사	1년	10만	
	프랜차이즈업조사	1년	2만5천	
	소상공인실태조사	1년	4만4천	
가구	Population Census	5년	전체가구	[세분류(인총·지역별고용)/중(대)분류단위] - 사업체(직장) 이름 - 사업체가 하는 일 - 사업체 소재지
	Regional Employment Survey	반기	23만	
	경제활동인구조사	월	3만5천	
	가계금융복지조사	1년	2만	
	가계동향조사	분기	7천	
	이민자체류실태조사	1년	2만5천	
	사회조사	1년	2만7천	
	생활시간조사	5년	2만9천	
농림어업	농림어업총조사	5년	120만	[대분류 일부단위] - 주 종사 분야 선택 (농업.임업.어업.제조업.건설업. 도소매업.숙박음식업.기타)
	농림어업조사	1년	4만5천	
	농가경제조사	1년	3천	
	어가경제조사	1년	1천	

< COICOP* for the Household Finances and Living Conditions Survey >

* Classification of Individual Consumption by Purpose

Income & Expenditure Diary

28일 (목) 요일

번호	수입종류 및 지출의 품명과 용도	수입		지출		
		금액	현물	금액	외상	선물
1	카메라 할부 구입 (신용카드 6개월)		☐	600,000	☐	○
2	카메라 할부금 갚음		☐	100,000	☐	○
3	따로 사는 자녀 등록금 대납		☐	3,000,000	☐	✓
4	따로 사는 자녀에 줄 된장 구입		☐	200,000	☐	✓

	가계동향조사 (통합: 2019년~)		가계동향조사 (통합: 2019년~)
코드	항목명	코드	항목명
A - Income		M8 – Consumables	
B2 – Current income		M9 – Domestic services	
B4 – Earned income		N – Health	
B410 – Householder income		N1 - Medicine	
B420 – Householder wages & salaries		N2 – Medical consumables	
B430 – Householder bonuses		N3 – Medical supplies and devices	
B440 – Spouse income		N4 – Medical services	
B450 – Spouse wages and salaries		N5 – Dental services	

Topology of survey items and codes:
B420, B430, B450 …

Optimization of language models and classification algorithms for statistical uses

- Developing a general language model (LM) for statistical applications vs. domain-specific LM referencing relevant statistical classification systems

< Classifications of diseases and causes of death >

■ 의료법 시행규칙 [별지 제6호서식] <개정 2015.12.23.>

사망진단서(시체검안서)

※ [ ]에는 해당되는 곳에 "✔"표시를 합니다.

등록번호		연번호		원본 대조필인
① 성 명			② 성 별	[ ]남[ ]여
③ 주민등록번호	-	④ 실제생년월일	년 월 일	⑤ 직 업
⑥ 등록 기준지				
⑦ 주 소				
⑧ 발 병 일 시	년 월 일 시 분(24시간제에 따름)			
⑨ 사 망 일 시	년 월 일 시 분(24시간제에 따름)			
⑩ 사 망 장 소	주소			
	장소	[ ] 주택 [ ]의료기관 [ ] 사회복지시설(양로원, 고아원 등) [ ] 공공시설(학교, 운동장 등) [ ] 도로 [ ] 상업·서비스시설(상점, 호텔 등) [ ] 산업장 [ ] 농장(논밭, 축사, 양식장 등) [ ] 병원 이송 중 사망 [ ] 기타( )		

Causes of death	A. Direct cause
	B. Cause of A
	C. Cause of B
	D. Cause of C
	Other physical conditions
Type of death	Death from diseases ( ), Accidental death ( ) , Undetermined death ( )
Accident type	Traffic ( ), Addiction ( ), Fall ( ), Drowning ( ), Burning ( ), Other ( )

KCD-8 DB Masterfile

22.06.30. 현재

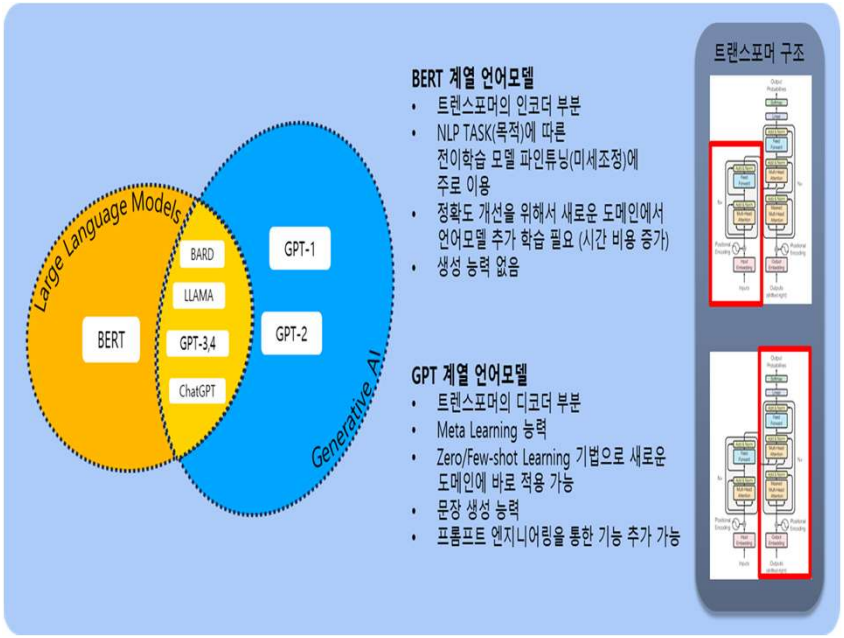
표제어	분류기준	질병분류코드	검별	주석	한글명칭	영문명칭	최하위코드	국내세분화코드	한의병명	국내추가진단명
1	대	A00-B99			I.특정 감염성 및 기생충성 질환(A00-B99)	Certain infectious and parasitic diseases(A00-B99)	0	0	0	0
1	중	A00-A09			장 감염 질환(A00-A09)	Intestinal Infectious Diseases(A00-A09)	0	0	0	0
1	소	A00			**콜레라**	**Cholera**	0	0	0	0
1	세	A00.0			비브리오 콜레라 01 콜레라형균에 의한 콜레라	Cholera due to Vibrio cholerae 01, biovar cholerae	1	0	0	0
		A00.0			고전적 콜레라	Classical cholera	1	0	0	0
1	세	A00.1			비브리오 콜레라 01 엘토르형균에 의한 콜레라	Cholera due to Vibrio cholerae 01, biovar eltor	1	0	0	0
		A00.1			엘토르 콜레라	Cholera eltor	1	0	0	0
1	세	A00.9			상세불명의 콜레라	Cholera, unspecified	1	0	0	0
1	소	A01			**장티푸스 및 파라티푸스**	**Typhoid and paratyphoid fevers**	0	0	0	0
1	세	A01.0			장티푸스	Typhoid fever	1	0	0	0
		A01.0			장티푸스균에 의한 감염	Infection due to Salmonella typhi	1	0	0	0
1	세	A01.1			파라티푸스 A	Paratyphoid fever A	1	0	0	0
1	세	A01.2			파라티푸스 B	Paratyphoid fever B	1	0	0	0
1	세	A01.3			파라티푸스 C	Paratyphoid fever C	1	0	0	0
1	세	A01.4			상세불명의 파라티푸스	Paratyphoid fever, unspecified	1	0	0	0
		A01.4			파라티푸스균에 의한 감염 NOS	Infection due to Salmonella paratyphi NOS	1	0	0	0
1	소	A02			**기타 살모넬라감염**	**Other salmonella infections**	0	0	0	0
		A02		포함	살모넬라 티피와 살모넬라 파라티피 이외의 살모넬라종에 의하여 생긴 감염 또는 식	Infection or foodborne intoxication due to any Salmonella species other than S. Typhi	0	0	0	0

KM-BERT: A BERT-based Korean language model trained on a Korean corpus of medicine and health

- Trained on 120 million words sourced from medical textbooks, news articles, and published papers
- Proven to outperform existing LMs when applied to medical datasets

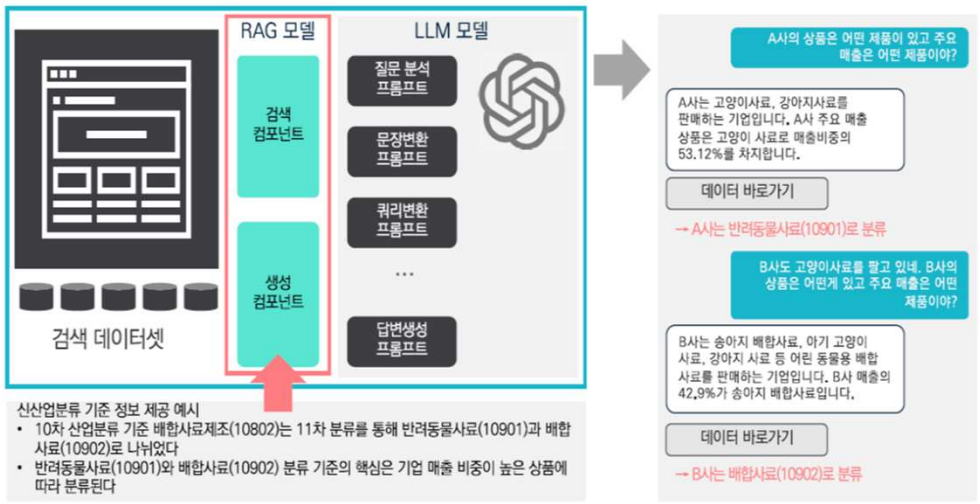
Adoption and application of generative large language models (generative LLMs)

- The current AI-based auto-coding system relies on a guided learning approach, using statistical data published in the past.
- This raises the need for adopting generative LLMs for problem-solving, especially in cases where sufficient 'correct' datasets are unavailable.



< Data refinement using an RAG* approach >
* Retrieval Augmented Generation

응답을 생성하기 전에 학습 데이터 소스 외부의 신뢰할 수 있는 지식 베이스를 참조하도록 하는 프로세스
아래는 RAG 기반 산업분류 LLM 모델 예시





국가데이터처
국가데이터연구원

Thank you.

