인공지능 기반 사회 통계 생산 방법론 고도화 방안: 가계동향조사와 생활시간조사 사례

오교중°, 최호진, 김일구, 한승우, 김건수
KAIST 전산학부, 아일리스프런티어 연구분석실, 통계개발원 통계방법연구실
{aomaru, hojinc}@kaist.ac.kr, {19kim, hsw90}@aift.kr, kgunsoo@korea.kr

Advancing Societal Statistics Processing Methodology through Artificial Intelligence: A Case Study on Household Trend Survey and Time Use Survey

Kyo-Joong Oh°, Ho-Jin Choi, Ilgu Kim, Seungwoo Han, Kunsoo Kim
KAIST School of Computing, AilysFrontier, Statistics Korea

요 약

본 연구는 한국 통계청이 수행하는 가계동향조사와 생활시간조사에서 자료처리 과정 및 방법을 혁신하려는 시도로, 기존의 통계 생산 방법론의 한계를 극복하고, 대규모 데이터의 효과적인 관리와 분석을 가능하게 하는 인공지능 기반의 통계 생산을 목표로 한다. 본 연구는 데이터 과학과 통계학의 교차점에서 진행되며, 인공지능 기술, 특히 자연어 처리와 딥러닝을 활용하여 비정형 텍스트 분류 방법의 성능을 검증하며, 인공지능 기반 통계분류 방법론의 확장성과 추가적인 조사 확대 적용의 가능성을 탐구한다. 이 연구의 결과는 통계 데이터의 품질 향상과 신뢰성 증가에 기여하며, 국민의 생활 패턴과 행동에 대한 더 깊고 정확한 이해를 제공한다.

주제어: 자연어처리, 기계학습, 통계생산 방법론, 국가통계

1. 서론

현대 사회에서 통계 데이터는 정책 결정, 사회과학 연구, 경제 분석 등 다양한 분야에서 중요한 역할을 수행한다. 이러한 통계데이터는 정부와 기업이 사회의 다양한 측면을 이해하고, 효과적인 결정을 내리는 데 필수적인 도구로 작용한다. 특히, 한국 통계청은 가계동향조사와 생활시간조사를 통해 국민의 소비 생활 패턴과 행동을 조사하고 분석하고 있다.

그러나, 이러한 조사들은 대규모의 데이터를 생성하며, 이를 효과적으로 다루기 위해서는 기존의 통계 방법론만으로는 한계가 있으며, 이 데이터를 효과적으로 분석하고 관리하는 것은 상당한 도전 과제이다. 인공지능 기술의 도입은 이러한 문제를 해결하고 통계 생산 방법론을 고도화할 수 있는 방법으로 간주되고 있다.

본 연구의 핵심 목적은 인공지능 기반의 통계 생산 방법론을 개발하고, 이를 가계동향조사와 생활시간조사에 적용했을 때 어떤 분석 결과를 낼수 있는지를 규명하는 것이다. 이 과정에서, 비정형 데이터의 처리와 분석에 있어 인공지능 기술의 활용 가능성에 중점을 둔다. 인공지능 기술의 도입은 이러한 데이터 분석을 더욱 효율적이고 정확하게 수행할 수 있게 하여, 통계 생산 방법론의 고도화를 추진할 수 있다.

본 연구는 가계동향조사와 생활시간조사의 데이터 처리와 분석 방법에 대한 새로운 통찰을 제공한다. 이를 통해, 해당 조사들의 결과가 더욱 신뢰할 수 있게 되며, 국민의 생활 패턴과 행동에 대한 더욱 깊고 정확한 이해를 가능하게 하며, 한국 사회의 다양한 측면을 조명하며, 정책 결정자들이 더욱 정보에 기반한 결정을 내릴 수 있도록 돕는다.

2. 연구 배경 및 관련 연구

2.1 통계청 데이터와 통계 생산 방법론

통계청은 국가의 주요 정책과 사회, 경제 현상을 반영하는 다양한 통계 자료를 작성하고 제공하는 기관으로, 통계 자료의 품질과 비교성 향상이 중요하다. 최근에는 빅데이터와 머신 러닝·딥 러닝 등 인공지능(AI) 기술을 활용하여 국가 통계 생산 과정에서의 정확성과 성능 개선의 필요성이 높아지고 있으며, 이를 위해 새로운 분석 방법론의 적용과 데이터 자체의 품질 향상이 추진되고 있다.

분류 체계는 통계학적 개념이나 현상을 체계적으로 분류하고 정의하는 방법으로, 인공지능 기반의 방법론은 텍스트 마이닝, 자연어 처리, 머신 러닝 등의 기술을 활용하여 통계학적 개념이나 현상을 자동으로 인식하고 분류하며, 분류 체계의 생성, 유지, 갱신, 매핑 등의 작업을 자동화하거나 지원할 수 있다. 통계청은 국제 분류를 기반으로 산업, 직업, 질병사인 등 다양한 표준 분류를 작성운영하고 있으며, 이 연구에서는 생활 시간 조사의 행동 분류, 사망 원인 통계의 사인 분류, 가계 동향 조사의 수입 지출 항목 코드를 확대 적용하고자 한다.

현재, 통계 자료 처리는 많은 시간, 예산, 인력이 투입되며, 주로 비정형(텍스트) 입력을 기반으로 분류 코드를 결정하는 방식으로 진행된다. 그러나, 시간이 지남에 따라 새로운 개념이나 현상이 발생하여 기존 분류 체계의 수정이 필요한 경우가 있으며, 이 과정은 인력과 시간이 많이 소요되고 통계 자료의 일관성과 연속성을 해치게 된다.

2.2 인공지능 기반 통계 데이터 처리

기존에 통계청에서는 규칙과 사전 기반의 산업직업 분류 자동 코딩 시스템과 사망 원인 자동 분류 시스템 등을 운영하고 있다. 이러한 시스템은 사전에 정해진 규칙과 입력 조건에 따라 분류 작업을 수행하며, 색인어나 사전 지식의 구축과 규칙 지식의 생성 및 관리가 필요했다. 이 방법론은 분류 항목이 적고 변화가 적은 경우에 적합하지만, 분류 항목의 다양성이 증가하면서 규칙과 조건 설정이 어려워지고, 유지 보수에 많은 시간과 비용이 발생하게 되었다. 또한, 데이터의 양이나 분류 항목이 변경될 경우 시스템의 유연성이 저하되어, 시스템을 재설정해야 하는 필요성이 있었다.

이러한 단점을 보완하기 위해 기계 학습 방법론을 적용하여 새로운 통계 분류 시스템을 구축하였다. 이 시스템은 대량의 데이터로부터 학습을 통해 자동으로 조사 자료 입력의 패턴을 학습할 수 있으며, 인력에 의한 수작업에 비해 더욱 정확하고 일관성 있는 분류 작업을 수행할 수 있다. 또한, 대량의 데이터를 빠르게 처리할 수 있어 분류 작업의 효율성을 높일 수 있으며, 분류 항목의 추가나 변경에도 빠르게 대처할 수 있는 유연성을 제공한다.

2.3 관련 연구

기존 통계청의 자동코딩 시스템은 [1, 2]의 연구를 기반으로, 사전 지식과 색인어 검색 기반의 방법론을 적용하여 구축하였다. 이 시스템은 규칙과 지식 기반 알고리즘을 사용하여, 사전에 정의된 규칙과 조건을 바탕으로 데이터를 분류한다. 이 방법은 특정 분류 항목이 상대적으로 적고 변화가 미미할 때 효과적이나, 분류해야 하는 데이터와 항목이 많아짐에 따라 그 한계가 드러나게 되었다. 또한, 시간이 흐르면서 새로운 형태의 조사 입력이 발생하며, 시스템의 응답율이 점차 감소하였다.

이러한 한계를 극복하기 위해, 기계학습 기반의 방법론[3]이 도입되었다. 이 방법론은 한국어 입력 데이터의 바이그램 정보를 활용하여 KNN (K-Nearest Neighbors) 알고리즘을 사용하여 군집화 학습을 수행한다. 그결과 산업 및 직업 코드를 N-best 형태로 자동으로 제공하는 시스템을 구축할 수 있었다.

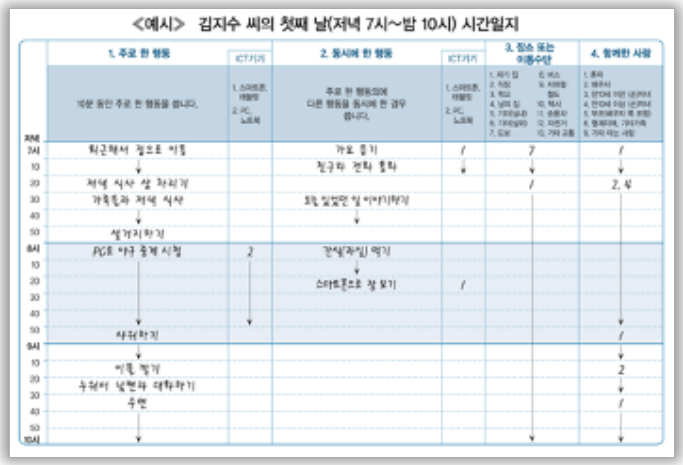
[4]에서는 문장 임베딩, 즉 문장의 의미를 수치적 벡터로 표현하는 기법을 활용하여 텍스트 분류 방법을 산업 분류에 적용하였다. [5]에서는 KoBERT, 한국어 텍스트를 위한 사전학습 언어 모델을 활용하여 텍스트 데이터의 학습 및 분류를 수행하였다. [6]에서는 다양한 언어의 입력 데이터를 처리할 수 있는 xlm-RoBERTa-Large 모델을 활용하여 관세청 수입 목록통관 데이터를 분류하는 연구를 수행하였다. [7]에서는 KoElectra와 DistilBERT 등 다양한 사전학습 언어모델을 전이 학습하여, 여러 분류 모델을 앙상블하는 방법을 적용하여 보다 정확한 분류 성능을 달성하였다.

수입과 지출 〈예시〉

28일 (목) 요일

번호	수입종류 및 지출의 품명과 용도	수 입		지 출		
		금 액	현물	금 액	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	☐	✓

그림 1. 가계동향조사 수입지출 조사표(예시)



<예시> 김지수 씨의 첫째 날(저녁 7시~밤 10시) 시간일지

그림 2. 생활시간조사 시간일지(예시)

3. 연구 방법론

3.1 데이터 소개 (가계동향조사와 생활시간조사)

본 연구에서는 "가계동향조사"와 "생활시간조사" 두 조사 자료를 활용하여 분류 모델의 학습과 분석을 진행한다.

가계동향조사는 국민의 가계 수입, 지출 및 주거 관련 사항을 상세히 파악하기 위한 통계 조사로, 매월 실시되며 연간(매월 9월) 공표 주기를 가진다. 1는 해당 조사의 조사표인데, 매 주기마다 약 770만건의 데이터가 수집되며, 가계 구성원의 수입/소득, 지출 항목에 대한 품명과 용도, 금액, 선물여부 등이 표시된다. 이 조사에서 수집된 데이터는, 한국표준 개별소비자출분류(COICOP-K, 2020 개정) 기반의 대분류 17개, 중분류 116개, 소분류 571개 범주로 이루어진 수입지출항목코드를 통해 체계적으로 분류된다. 이중 대분류 B, C는 수입과 소득에 해당하는 분류코드(170건)이며, 나머지는 지출 항목에 해당하는 분류코드(401건)으로 구성된다.

생활시간조사는 국민의 생활 행동 패턴과 시간 활용 현황을 깊게 이해하기 위해 5년마다 실시되는 통계 조사이다. 최근 조사는 2019년에 이루어졌으며, 계절성과 지역성을 반영하여 표본을 추출하여 연간 4회 조사가 진행된다. 조사 과정에서는 2와 같이 통계 조사원이 조사군 대상의 가구원의 행동을 10분 단위로 기록하며, 주행동, 동시행동, 그리고 행위장소, 함께한 사람 등의 정보가 포함된다. 이 데이터는 대분류 9개, 중분류 45개, 소분류 153개 범주로 구성된 행동분류 코드를 통해 체계적으로 분류된다.

3.2 데이터 처리 및 분석 방법

가계동향조사 자료에서 최종 공표된 2021년 데이터 7,788,234건을 학습과 검증 데이터를 9:1 비율 층화표집(Straitfied sampling)하여 구축하였으며, 올해(2023년) 공표 예정인 2022년 데이터 7,694,608건을 평가 데이터로 활용하였다.

생활시간조사 데이터에서는 입력 데이터의 중복이 많은 특징을 고려하여, 총 8,400,817건의 입력 데이터를 354,029건으로 정제하였으며, 학습 데이터 212,417건, 검증 데이터 70,806건, 그리고 평가 데이터 70,806건으로 각각 층화 표집하여 사용하였다.

분석 방법론 개발과 검증 과정에서는 정분류율(Accuracy)와 F-1 Score를 주요 평가 지표로 활용하였다. 정분류율은 분류 모델이 평가 데이터셋에 대하여 올바르게 분류한 항목의 비율을 나타내며, F-1 Score는 모델의 정밀도와 재현율의 조화 평균을 나타내는 지표로, 데이터 불균형이 있을 때 효과적인 평가 지표로 활용된다.

3.3 인공지능 모델 설명

본 연구에서는 통계 조사 데이터의 텍스트 분류 모델 학습을 위해 사용된 사전학습 언어모델은 KCBERT와 KLUE-RoBERTa를 활용하였다. 이러한 모델들은 텍스트 데이터의 복잡한 패턴을 파악하고, 분류 작업을 효과적으로 수행할 수 있는 높은 성능을 보유하고 있다. 또한, 허깅페이스(Huggingface)의 텍스트 분류(text classification) 파이프라인(pipeline)을 기반으로 한 SequentialClassfier를 전이학습 분류 모델로 사용하였다.

3.3.1 사전학습 언어모델

KCBERT[8]: KCBERT는 대규모의 온라인 뉴스 댓글과 대댓글 데이터를 활용하여 학습된 BERT 모델이다. 특히 신조어, 구어체, 특수문자, 이모지 등 다양한 형태의 텍스트에 유연하게 대응할 수 있도록 설계되었다. KCBERT는 Huggingface의 Transformers 라이브러리를 통해 쉽게 접근하고 사용할 수 있다. 단, 영문 처리에 있어서는 대소문자를 구분하는 Cased model을 사용하므로, 일부 제한적인 부분이 있을 수 있다.

KLUE-RoBERTa[9]: KLUE-RoBERTa는 다양한 자연어처리 태스크의 한국어 말뭉치를 기반으로 사전학습된 RoBERTa 모델로 한국어 텍스트에 대한 깊은 이해력을 보유하고 있다. BertTokenizer를 사용하여 텍스트를 처리하며, Huggingface Model Hub를 통해 쉽게 접근할 수 있다.

3.3.2 미세조정 분류모델

SequentialClassfier[10]: SequentialClassfier는 허깅페이스의 텍스트 분류 파이프라인을 기반으로 개발된 모델로, 주어진 텍스트 데이터에 대해 적절한 레이블 또는 클래스를 할당하는 작업을 수행한다. 이 모델은 감성 분석, 자연어 추론, 문법 정확성 평가 등 다양한 텍스트 분류 작업에 활용될 수 있다. 허깅페이스의 트랜스포머스(Transformers) 라이브러리를 통해 쉽게 사용할 수 있다.

4. 연구 결과 및 분석

4.1 가계동향조사 수입지출분류 결과 분석

21년도에 수집된 총 7,694,608건의 데이터를 학습 및 검증 목적으로 사용하였다. 이 데이터를 9:1의 비율로 층화표집하여, 6,925,152건은 학습 데이터로, 769,462건은 검증 데이터로 활용하였다.

학습 과정에서는 총 12회의 반복 학습이 이루어졌으며, 그 과정은 그림 3에서 확인할 수 있다. 이 중 9회 반복(6741 step)에서 모델이 최적의 성능을 보여주었으며, 이 시점에서의 학습셋 손실율은 0.0894, 평가셋 손실율은 0.1927로 측정되었다.

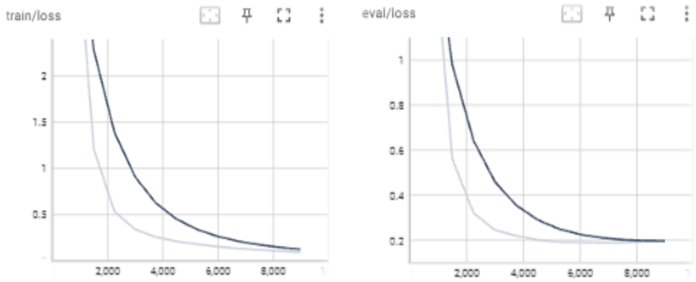


그림 3. 가계동향조사 학습 결과, train loss(좌), eval loss(우)

표 1. 가계동향조사 수입지출분류모델 정확도 결과

	학습 결과 (검증셋)	테스트 결과 (평가셋)
정분류율 (Accuracy, %)	95.66	98.00
F1-Score (Weighted, 0-1)	0.9565	0.9795
F1-Score (Macro, 0-1)	0.8829	0.8601

처리코드	항목명	f1-score	데이터 개수
B487	7번가구원급여소득	0	5
B495	5번가구원상여금	0	1
B536	6번가구원사업소득	0	9
B537	7번가구원사업소득	0	10
B545	5번가구원임대소득	0	11
B623	3번가구원배당소득	0	5
B624	4번가구원배당소득	0	1
B643	3번가구원개인연금소득	0	4
B645	5번가구원개인연금소득	0	3
B646	6번가구원개인연금소득	0	3
B652	2번가구원퇴직연금소득	0	21

그림 4. 수입지출분류 결과 예측 오류 항목

각 단계에서의 세부 평가 지표 수치는 표1에서 확인할 수 있다. 22년도에 7,788,234건의 평가 데이터를 기반으로 수행한 평가에서는 높은 정확도를 보여주고 있다. 정분류율은 98%, F-1(weighted) 점수는 0.9795, 그리고 ROC 점수는 0.9793으로, 이러한 결과들은 전반적으로 매우 높은 정확도를 나타낸다.

특히, 지출 부분을 나타내는 대분류 B와 C를 제외하면, F-1(macro) 점수가 0.9503으로 나타났다. 이러한 높은 점수는 실제 업무 시스템 구현이나 실무 적용을 고려할 때 충분히 시스템화가 긍정적으로 고려되는 수치이다.

전체 571개의 분류 항목 중에서 45개의 항목은 정분류율이 50% 이하이며, 38개의 항목은 50% 이상 80% 이하로 나타났다. 이러한 결과는 주로 학습 및 평가 데이터의 양이 1,000개 미만으로 부족하거나, 그림 4의 항목명으로 부터 유추할 수 있다시피, 입력에 실명 정보만 입력되었거나, 몇 번째 가구원인지에 대한 정보가 누락되어, 특정 가구원의 분류 정보를 추론할 수 없는 경우가 대부분이었다.

그 외에 특이 사례로는 21년에는 '가구주사회수혜금'으로 구분했던 입력이 22년에는 '가구주아동양육수당'으로 분류 기준이 변경되어 전체적으로 오류가 있는 항목이 있다는 점과, 코로나19의 영향으로 20 22년 해외 연수에 해당 하는 학습 데이터가 없어서 점차 늘어나는 어학연수나 국외 연수 비 등의 항목에서 오류가 생기는 것을 발견할 수 있었다.

4.2 생활시간조사 행동분류 결과 분석

본 연구에서는 최근 2019년 조사(5년 주기) 총 7,578,804건의 행동 기록 데이터에서 입력의 중복을 제거한 354,029건의 데이터를 활용하여 생활 시간 조사 결과를 분석하였다. 이 데이터는 층화표집(Stratified sampling) 방법을 통해 학습 데이터 212,417건, 검증 데이터 70,806건, 그리고 평가 데이터 70,806건으로 구분하였다. 모델 학습은 총 12회의 반복 학습을 거쳤으며, 이 중 10.6(그림 5)회 반복 시점에서 최적의 정확도을 보여주는 분류 모델이 학습되었다. 해당 시점에서의 학습셋과 평가셋의 손실율은 각각 0.4826과 0.7661로 측정되었다.

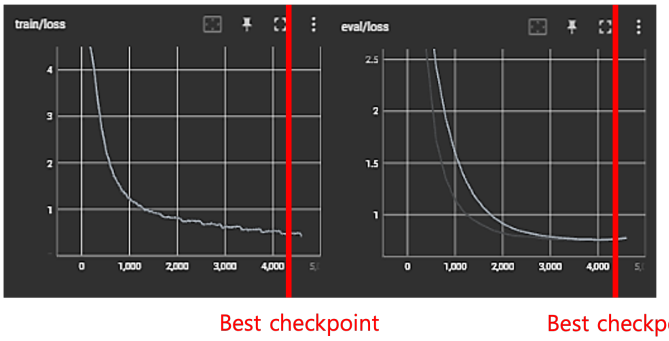


그림 5. 생활시간조사 학습 결과, train loss(좌), eval loss(우)

표 2. 생활시간조사 행동분류모델 정확도 결과

	학습 결과 (검증셋)	테스트 결과 (평가셋)
정분류율 (Accuracy, %)	77.80	77.80
F-1 점수 (Weighted, 0-1)	0.7751	0.7752
F-1 점수 (Macro, 0-1)	0.6547	0.6581

평가 단계에서는 70,806건의 평가 데이터를 기반으로 세부 평가 지표를 분석하였으며, 그 결과는 표2에서 확인할 수 있다. 통계청의 공표 및 집계 수준인 대분류 수준에서는 6을 참조하면, 자원봉사 및 무급연수 항목을 제외하면 대부분의 항목에서 0.9 이상의 높은 F-1 점수를 기록하였다.

더욱이, 정분류율이 50% 이하인 분류 항목이 32개에 해당한다. 이러한 항목들을 분석해본 결과, 대부분의 경우 학습과 평가 데이터의 수가 100개 이하로 매우 적었다. 분류 체계상 중/소 분류 항목에서 '기타'로 분류되는 경우가 많았으며, 또한 조사 대상의 산업/직업 분류 정보가 추가로 제공되면 더욱 명확한 분류 결과를 얻을 수 있을 것으로 예상되었다.

4.3 통합 결과 분석

본 연구에서 수행한 가계동향조사와 생활조사 행동분류 모델의 통합 결과 분석은 표 3에서 확인할 수 있다. 가계동향조사 모

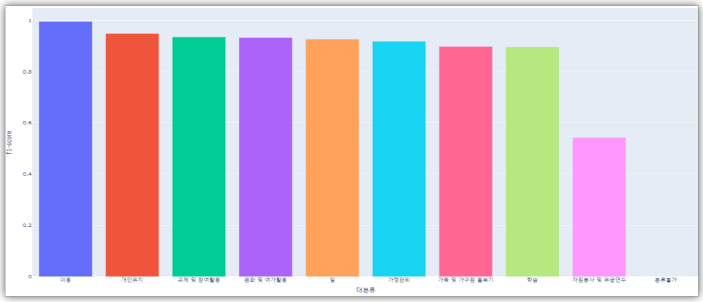


그림 6. 행동분류 예측결과(대분류)

표 3. 가계동향조사 및 생활시간조사 분류모델 정확도

	가계동향조사 수입지출분류 (평가셋)	생활시간조사 행동분류 (평가셋)
정분류율(%)	98.00%	77.80%
F-1 점수	0.9795	0.7752

델은 학습 결과와 테스트 결과에서 높은 정분류율과 F1-Score를 보여주었다. 특히, 테스트 결과에서 정분류율은 98.00%, 가중치가 적용된 F1-Score는 0.9795로 나타났다. 생활조사 행동분류 모델은 정분류율과 F1-Score (Weighted)는 각각 77.80%와 0.7752로 가계동향조사 모델에 비해 낮은 성능 지표를 보여주었다.

이러한 결과는 가계동향조사 데이터가 생활조사 데이터에 비해 높은 품질의 데이터가 수집되고, 품질과 양이 좋은 데이터로 구성되어 있기 때문이다. 이러한 차이점을 고려하여, 각 모델의 성능을 더욱 향상시킬 수 있는 방법을 연구하는 것이 필요하다.

5. 결론

본 연구는 인공지능을 활용하여 통계분류 작업의 효율성과 정확성을 높이는 방법을 탐구하였다. 전통적인 통계분류 작업은 인력 의존도가 높고 시간 소모가 크며, 숙련도 차이로 인한 정확도 저하와 일관성 결여 문제가 있었다. 따라서 현재 많은 분류와 관련된 통계조사 및 자료처리 업무에 상당한 인력, 예산, 시간이 투입되고 있다. 본 연구에서 제시된 자연어처리와 기계학습 방법은 이러한 문제점들을 해결할 수 있는 효과적인 방안으로 제시되었다.

본 연구에서는 가계동향조사와 생활시간조사의 분류 데이터에 집중하여 학습과 분석 결과를 제시하였으며, 분류 정확도가 떨어지는 분류 항목과 데이터 대한 선택적 에디팅(검증)이 가능한 항목에 대한 분석도 함께 제공하고 있다. 기존에 축적한 대량의 통계조사 데이터를 빠르고 정확하게 처리할 수 있으며, 학습을 통해 지속적으로 발전시킬 수 있다. 이러한 접근 방식은 일관성 있는 분류 작업을 지속적으로 수행하여, 통계분석 작업의 효율성을 높일 수 있으며, 고비용 문제를 해결할 수 있다.

향후 연구는 다양한 통계조사로의 확대와 지속적인 통계자료의 일관성과 연속성을 유지하는 것이 중요하다. 또한, 분류체계의 효율성과 정확성을 개선하고, 통계자료의 통합과 연계를 용이하게 하는 추가 연구 및 개발이 필요하다. 이러한 연구는 통계분석의 정확성과 효율성을 더욱 높이고, 국가 통계 시스템의 신뢰성을 강화할 것으로 기대된다.

감사의 글

이 논문은 2023년도 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원의 지원을 받아 수행된 연구임. (No. 1425171943, AI 시스템 및 서비스에 범용적 적용이 가능하고 비전문가도 쉽게 이해 가능한 eXplainable AI(설명가능한 인공지능) 솔루션 개발)

참고문헌

[1] 임희석, "예제기반의 학습을 이용한 한국어 표준 산업/직업 자동 코딩 시스템," *한국콘텐츠학회논문지*, Vol. 5, No. 4, pp. 169–179, 2005.

[2] S.-H. M. Y. Jung, J. Ryu and D.-C. Han, "A web based automated system for industry and occupation coding," *The 9th International Conference on Web Information Systems Engineering*, pp. 443–457, 2008.

[3] 임희석, "표준 통계 분류 코드 자동 생성," *한국산학기술학회 춘계학술발표논문집*, pp. 388–390, 2006.

[4] 오교중, 최호진, and 안현각, "기계학습 기반 단문에서의 문장 분류 방법을 이용한 한국표준산업분류," *제32회 한글 및 한국어 정보처리 학술발표 논문집*, 2020.

[5] 임정우, 문현석, 이찬희, 우찬균, and 임희석, "딥러닝 기법을 활용한 산업/직업 자동코딩 시스템," *한국융합학회*, Vol. 12, No. 4, pp. 23–30, 2021.

[6] 오교중, 최호진, 차원석, 김일구, and 우찬균, "언어모델 전이학습 기반 해외 직접 구매 상품군 분류," *제34회 한글 및 한국어 정보처리 학술발표 논문집*, 2022.

[7] 오교중, 최호진, 김진원, 차원석, and 김일구, "기업 비지니스 분석을 위한 한국표준산업코드 앙상블 분류," *제34회 한글 및 한국어 정보처리 학술발표 논문집*, 2022.

[8] SKTBrain, "Kcbert: Pretrained language model for korean comments," 2021. [Online]. Available: https://github.com/Beomi/KcBERT

[9] K. Team, "Klue-roberta: Pretrained language model for korean," 2021. [Online]. Available: https://github.com/KLUE-benchmark/KLUE

[10] H. Inc., "Huggingface's transformers: Text classification pipeline," 2021. [Online]. Available: https://huggingface.co/transformers/main_classes/pipelines.html#textclassificationpipeline