Journal of KIIT. Vol. 24, No. 5, pp. 1-13, May 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.5.1

KoBERT - LLM 합의 기반 기타 라벨 재라벨링 기법

권수민*, 곽지호**, 박지영***, 정유철****

Consensus-based Relabeling of the “Other” Class using KoBERT and Large Language Models

Sumin Kwon*, Jiho Gwak**, Jiyoung Park***, and Yuchul Jung****

본 연구는 한국과학기술정보연구원(KISTI)의 위탁연구개발과제로 수행한 것입니다. 아울러 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-학·석사연계ICT핵심인재양성 (IITP-2026-RS-2022-00156394) 및 2025년도 교육부 및 경상북도의 재원으로 경북RISE센터의 지원을 받아 수행된 지역혁신중심대학지원체계(RISE)-(지역성장 혁신LAB)의 결과입니다 (2025-rise-15-105).

요 약

텍스트 분류에서 ‘기타’ 라벨은 서로 다른 주제와 모호한 의미를 한데 묶으면서 전체 성능의 바닥을 결정하는 요인으로 작용한다. 본 연구는 KoBERT 분류기와 대형 언어모델을 결합해 두 모델의 합의와 신뢰도를 이용하여 기타 라벨 문서를 자동으로 재라벨링하는 방법론을 제안한다. 기준 분류기의 예측과 언어모델의 의미 기반 예측을 함께 활용해 재라벨링 데이터를 구성하고, 이를 다시 학습에 반영한 결과, 기타 라벨에 속한 문서들이 보다 적절한 라벨로 분산되면서 전체 분류 성능과 기타 라벨 집단의 예측 안정성이 동시에 향상되는 것을 확인하였다. 향후 본 연구에서 발견된 특정 라벨의 성능 정체 현상을 해결하기 위해 계층적 분류 모델 및 프롬프트 최적화 기법을 도입하여 범용적인 데이터 정제 프레임워크로 확장하고자 한다.

Abstract

The “other” label in text classification aggregates heterogeneous and ambiguous samples, often limiting overall performance. We propose an automatic relabeling method that combines a KoBERT classifier with an LLM, using their agreement and confidence to reassign “other” documents to more suitable labels. Retraining on the relabeled data redistributes “other” samples and improves overall performance as well as the stability of “other” predictions. This suggests model-consensus relabeling as a practical way to mitigate the “other” label issue with minimal manual validation. Future work will extend this into a universal data refinement framework via hierarchical classification and prompt optimization.

Keywords

large language model, text classification, label noise reduction, KoBERT, automatic relabeling, label correction

* 국립금오공과대학교 컴퓨터공학부 학사과정
- ORCID: https://orcid.org/0009-0002-5909-0154
** 국립금오공과대학교 컴퓨터·AI융합공학과 석사과정
- ORCID: https://orcid.org/0009-0000-7524-0357
*** 한국과학기술정보연구원 선임연구원
- ORCID: https://orcid.org/0000-0002-8132-3834
**** 국립금오공과대학교 컴퓨터공학부 부교수(교신저자)
- ORCID: http://orcid.org/0000-0002-8871-1979

· Received: Dec. 22, 2025, Revised: Feb. 20, 2026, Accepted: Feb. 23, 2026
· Corresponding Author: Yuchul Jung
Dept. of Computer Engineering, Kumoh National Institute of Technology,

61 Daehak-ro (yangho-dong), Gumi, Gyeongbuk, [39177] Korea
Tel.: +82-54-478-7536, Email: jyc@kumoh.ac.kr

Ⅰ. 서 론

대규모 텍스트 데이터가 지속적으로 생산되고 축적되는 환경에서 이를 일정한 기준에 따라 체계적으로 구분하는 텍스트 분류는 필수적인 분석 도구이다[1]. 특히 연구 과제, 논문, 기술 보고서 같은 텍스트의 경우 문서에 기술된 연구 내용을 적절한 연구 분야 라벨로 매핑하는 과정이 핵심이다[2]. 이때 분류 클래스는 단순히 이름을 붙인 표지가 아니라 특정 문서가 어떤 연구 주제와 기술 영역에 속하는지를 규정하는 의미적 경계로 기능한다. 그리고 이를 통해 전체 연구 성과의 구조적 특성, 분야별 분포, 정책, 투자상의 주요 관심 영역을 효율적으로 파악할 수 있다. 과학기술 분야와 같이 다양한 주제의 연구 성과가 방대하게 축적되는 도메인에서는 이러한 분류 체계의 설계와 라벨 정의 방식이 분석 결과의 신뢰도와 이후 활용 가능 범위를 직접적으로 좌우한다[3].

그러나 실제 현장에서 축적되는 텍스트는 이상적인 연구 분야 분류 체계만으로 설명되기 어렵다. 도메인 전문 용어와 일반적인 표현이 문서 내에서 혼재하고, R&D 정책이나 사업 구조 변화에 따라 라벨 정의가 부분적으로 변경, 확장되며 서로 다른 연구 분야 분류 체계가 병행, 중첩되는 경우도 존재한다. 이 과정에서 일부 라벨들은 여러 분야의 경계에 걸쳐 포괄적으로 사용되거나 기타(etc)와 같은 형태로 처리된다. 이러한 라벨은 모델이 연구 분야 간 결정 경계를 명확히 학습하는 것을 방해하고, 라벨 간 성능 편차를 심화시키며 전체 성능의 하한선을 낮추는 주된 요인이 된다.

지금까지의 많은 연구와 실무 적용 사례들은 노이즈 라벨을 줄이기 위해 라벨 필터링을 수행하거나 노이즈에 강인한 손실 함수와 학습 기법을 설계하는 등 주로 모델 중심의 개선에 집중해 왔다[4]. 이러한 접근은 일정 수준의 성능 향상을 가져오지만, 비정형 텍스트가 본질적으로 갖는 모호함과 중첩된 의미 구조를 연구 분야 분류 체계와 데이터 수준에서 재구성하고 정제하는 데에는 한계가 있다. 특히 연구 분야 코드 체계에서 “__99(기타)”와 같이 의미가 확산된 라벨 집단에 대해서 이를 별도의 대상 집단으로 정의하고 반복적으로 재라벨링하는 데이터 중심 연구에 대한 논의가 상대적으로 부족하다.

이러한 문제의식을 바탕으로 본 논문은 두 가지 이상의 모델 예측을 결합하여 고신뢰 표본에 한해 라벨을 치환하는 반복적 재라벨링 기법을 제안한다.

Ⅱ. 관련 연구

2.1 노이즈 라벨 정제와 라벨 노이즈 학습

라벨 노이즈는 모델의 예측 정확도를 떨어뜨리고 필요한 훈련 샘플 수와 모델 복잡도를 증가시킨다[4]. 특히 딥러닝 모델이 노이즈 라벨에 쉽게 과적합된다는 문제가 있다는 점이 지적된다[5]. 이러한 문제를 해결하기 위해 라벨 노이즈를 다루는 연구들은 주로 학습 과정에서 노이즈 샘플을 탐지하고 선택적으로 교정하는 것에 초점을 둔다[6]. 대표적인 접근법 중 하나로, 훈련 중 손실값 분포를 분석해 손실이 높은 샘플을 노이즈 후보로 구분하고 트랜지션 매트릭스를 이용해 잘못된 라벨을 다른 클래스로 보정하는 방식이 제안되었다. 이 방식은 단순 데이터셋뿐 아니라 복잡한 이미지 데이터셋에서도 기존 방법 대비 의미 있는 수준의 정확도 향상을 가져온다[7].

노이즈 조건에서도 안정적인 학습을 유지하는 것을 목표로 하는 메타러닝 결합 동적 라벨 교정 기법도 제안되고 있다[8]. 이 기법은 학습 과정에서 각 샘플 데이터의 손실, 불확실성, 클래스별 난이도 정보를 시계열로 축적한다. 그리고 이를 입력으로 사용해 메타 라벨 교정기를 학습하여 실제 라벨 분포를 추정한다. 다양한 노이즈 유형과 수준을 가진 벤치마크에서 메타러닝 기반 방법은 기존 노이즈 대응 방법보다 더 높은 정확도와 강건성을 보여준다. 특히 고노이즈 환경에서도 분류 성능과 라벨 교정 정확도를 비교적 안정적으로 유지하는 것으로 보고된다[9][10].

한편, 이미 학습된 분류기의 예측 확률과 임베딩을 활용해 라벨 오류, 아웃라이어, 중복 데이터, 분포 변화를 탐지하는 모델 기반 진단 방식도 제안되어 왔다[11]. 이러한 도구들은 예측 분포와 벡터 표

현을 통계적으로 분석하여 라벨 오류 가능성이 높은 샘플을 효과적으로 선별하는 데 초점을 맞춘다. 다만 이와 같은 모델 중심 기법들은 주로 노이즈 라벨을 탐지하고 제거하거나 수정하는 데 집중한다[12].

2.2 데이터 중심 재라벨링과 HITL

라벨 노이즈 문제를 데이터 측면에서 해결하고자 하는 접근은 공통적으로 모델 예측을 단서로 사람이 반복적으로 개입하는 재라벨링 과정을 거친다. 대표적으로 사람이 라벨링한 초기 데이터셋을 모델에게 학습시킨 뒤, 모델 예측과 기존 라벨이 다른 샘플을 노이즈 후보로 간주하여 사람이 다시 검토하고 수정한 데이터셋으로 모델을 여러 차례 재학습시키는 방식이 있다[13].

이러한 반복 재라벨링 파이프라인을 의료 대화 데이터에 적용하여 문장 수준 기능 라벨을 학습하는 방식도 제안되었다. 해당 연구는 모델 예측과 기존 라벨이 불일치하는 샘플을 우선적으로 사람이 검토하도록 설계함으로써 문장 분류 정확도를 69.5%에서 82.5%까지 향상시켰다[14]. 특히 라벨 경계가 모호하거나 기타 라벨과 경계 라벨이 많이 포함된 발화에서 반복 재라벨링 구조의 개선 효과가 두드러지는 것으로 나타났다.

대규모 미라벨 텍스트 코퍼스를 효율적으로 분류하기 위해 few-shot 텍스트 분류 프레임워크를 제안한 연구도 있다[15]. 또한 사전학습된 워드 임베딩을 이용해 문서를 벡터로 표현한 뒤, 소수의 대표 문서만 사람이 각 범주의 프로토타입으로 지정하고 나머지 문서를 임베딩 기반 유사도에 따라 자동 분류하는 Human-in-the-Loop 방식도 제안되었다[16]. 이 방식은 기존 라벨이 거의 없거나 기존 분류 체계가 무의미해진 상황에서도 비교적 적은 수의 수동 라벨만으로 실용적인 수준의 분류 성능을 확보할 수 있음을 보여준다.

라벨이 없는 데이터나 기타로 묶인 데이터를 보완하기 위해 의사 라벨을 활용하는 연구도 보고되고 있다[17]. 먼저 라벨이 있는 일부 데이터로 분류 모델을 학습한 뒤, unlabeled 데이터에 대해 예측을 수행하고 이 가운데 신뢰도가 임계값 이상인 샘플에만 새로운 라벨을 부여하는 방식이다. 이렇게 생성된 의사 라벨을 기존 라벨 데이터와 함께 다시 학습에 사용하면 추가적인 수작업 라벨 없이도 분류 성능을 향상시킬 수 있다[18][19]. 최근에는 대형 언어 모델에 설명문과 예시를 함께 제시해 라벨이 없는 데이터에 대해 직접 라벨을 생성하게 하거나, 분류 모델의 예측과 LLM의 판단을 결합해 의사 라벨을 구성하는 시도도 이루어지고 있다[20].

이와 같이 데이터 중심 재라벨링 및 HITL 연구들은 공통적으로 모델 예측을 활용해 의심 샘플을 선별하고 사람 또는 별도의 절차를 통해 라벨을 재검토하고 수정한 뒤 재학습을 통해 성능을 점진적으로 개선하는 흐름을 갖는다. 그러나 대부분의 연구는 데이터의 기타 라벨을 별도의 대상 집단으로 설정하고 이를 구조적으로 재배치하는 프레임워크를 제시하지는 않는다. 또한 사람의 반복 개입이 필수적이기 때문에 대규모 데이터셋에 반복 적용하기에는 비용과 시간이 많이 소요된다는 한계가 있다.

따라서 본 논문에서는 기존 연구의 한계를 보완하기 위해 과학기술표준분류 데이터셋에서 특정 기타 라벨 집단을 대상 집단으로 설정하고 분류 모델의 예측과 LLM 기반 의미 정보를 결합한 고신뢰 재라벨링 규칙을 설계한다. 다음 장에서는 이를 기반으로 한 반복적 재라벨링 기법의 전체 구조와 데이터 흐름, 그리고 구체적인 구현 절차와 실험 설계를 자세히 설명한다.

III. 데이터셋 분석

본 논문에서 제안하는 재라벨링 기법을 적용하기에 앞서 과학기술표준분류 데이터셋과 기준 분류기의 동작 특성을 정량적으로 진단하는 데 목적이 있다. 데이터 구성과 전처리 과정을 통해 라벨 불균형과 문서 특성을 살펴보고 라벨 단위 성능 분석을 통해 전체 평균 지표에 가려진 하위 라벨군의 분포를 드러낸다. 이어서 키워드 기반 텍스트 특성과 “기타”로 수렴되는 코드 구조를 함께 분석함으로써, 성능 저하가 특정 라벨군과 데이터 특성에 집중된다는 점을 보이고 이후 장에서 다룰 반복적 재라벨링 전략의 필요성을 뒷받침한다.

3.1 데이터셋 구성 및 전처리

본 연구에서 사용한 데이터는 KISTI 국가 R&D 과제 정보를 기반으로 하되 선행 연구에서 전처리를 거쳐 구축한 버전을 제공받아 활용하였다. 각 과제에는 과학기술표준분류 대, 중, 소분류 코드 및 명칭이 부여되어 있다. 이 가운데 과학기술표준분류 코드1-중을 대표 라벨로 사용하여 중분류 수준의 연구 분야를 예측하는 다중 분류 문제로 설정하였다. 입력 텍스트는 미리 구성해 둔 통합 요약 필드(merged_text)를 그대로 사용하였으며 이 필드는 요약문_연구목표, 요약문_연구내용, 요약문_기대효과를 하나의 문장열로 병합한 것이다. 또한 요약문_한글키워드와 요약문_영문키워드 컬럼에는 요약문으로부터 추출, 정제된 핵심어가 포함되어 있어 이후 텍스트 특성 분석과 LLM 기반 의미 판단 단계에서 보조 정보로 활용된다.

선행 연구 단계에서 수행된 주요 전처리는 다음과 같이 정리할 수 있다. 과학기술표준분류 코드는 공백, 특수문자 등을 정규화하여 동일한 중분류가 일관된 기호 체계로 표현되도록 정리하였다. 그리고 요약문 필드의 줄바꿈, 중복, 공백, 특수 기호를 정규화한 뒤 세부 항목을 결합해 merged_text를 생성하였다. 또한 완전히 동일한 텍스트를 가지는 중복 과제는 제거하여 중복 샘플이 모델 학습에 미치는 영향을 최소화하였다.

본 연구에서는 이러한 선행 전처리 결과를 그대로 인계받아 사용하되 대표 중분류 코드가 존재하지 않거나 merged_text가 비어 있는 과제만을 추가로 제거하는 최소한의 필터링만을 수행하였다.

전체 데이터에는 중분류 기준 267개의 라벨이 존재하지만 문서 수가 극히 적어 통계적으로 의미 있는 성능 평가가 어려운 희소 라벨은 분석 대상에서 제외하였다. 라벨별 문서 수 분포를 기준으로 하위 구간에 위치한 희소 라벨을 제거한 뒤 최종적으로 217개 중분류 라벨을 대상으로 분류 모델의 학습, 평가 단계에 사용하였다. 이 과정 이후 남은 과제 수는 총 624,408건이다.

3.2 라벨별 성능 분석

표 1을 살펴보면 KoBERT로 학습된 분류기로 출력한 라벨의 F1-score 평균과 중앙값은 각각 약 0.45 수준이며 상위 25% 라벨은 0.61 이상, 하위 25% 라벨은 0.32 이하에 분포해 라벨 간 편차가 크게 나타났다. 특히 표 2와 같이 코드가 '99'로 끝나는 기타 라벨 13개 집단의 평균 F1-score는 약 0.31로, 나머지 일반 라벨 집단의 평균 약 0.46보다 약 0.15 낮게 나타났다.

표 1. F1-score 상/하위권 코드 및 점수
Table 1. Top- and bottom-performing codes by F1-score (scores included)

Top labels by F1-score	
S&T standard classification code 1 (Mid-level)	F1-score
LC11(Food Safety management)	0.8361
LC10(Dental science)	0.8298
EL09(Hydraulic system technology)	0.8138
LA01(Molecular and cell biology)	0.7738

Bottom labels by F1-score	
S&T standard classification code 1 (Mid-level)	F1-score
EA99(Other machinery)	0.1152
EB99(Other materials)	0.1092
EH14(Cleaner production/equipment)	0.0267
EC99(Other chemical engineering)	0.0222

표 2. 기타라벨 집단과 일반라벨 집단의 F1-score 비교
Table 2. F1-score comparison between the "Other" label group and the regular label group

Label type	F1-score
Label 99 ("Other")	0.3092
Regular labels	0.4856

이와 같은 분석 결과는 전체 평균 성능만을 기준으로 할 때에는 드러나지 않는 성능 편차가 라벨 간에 존재하며 특히 일부 라벨군이 성능 하한을 규정하는 병목 역할을 하고 있음을 시사한다.

3.3 텍스트 특성 분석

앞 절의 성능 분석에서는 일부 라벨 특히 "기타" 라벨에서 성능 저하가 집중적으로 나타난다는 점을 확인하였다. 이러한 성능 차이가 어디에서 기인하는

지 밝히기 위해 중분류 코드가 '99'로 끝나는 라벨들만을 대상으로 텍스트 구성과 키워드 분포를 중심으로 텍스트 특성을 추가로 살펴보았다.

라벨별 텍스트 특성은 각 라벨 i에 대해 문서 비율 doc_ratio(i)를 계산하여 정량화하였다. doc_ratio(i)는 식 (1)과 같이 "라벨 i에 속한 문서 중 해당 라벨 키워드가 한 번 이상 출현한 문서 수"를 "라벨 i의 전체 문서 수"로 나눈 값이다. 여기서 D_i는 라벨 i로 분류(할당)된 문서들의 집합, K_i는 라벨 i와 연관된 키워드 집합을 의미한다. 또한 식 (1)의 분자는 D_i에 속한 문서 d 중에서 K_i의 키워드 k가 하나 이상 포함되는 문서의 개수를 나타내며, 분모는 라벨 i에 속한 전체 문서 수를 의미한다. 따라서 doc_ratio(i)는 라벨 i 문서에서 라벨 특이적 키워드가 실제로 관측되는 정도를 나타내는 지표로 해석할 수 있다.

여기서 "해당 라벨 키워드"는 요약문_한글키워드 및 요약문_영문키워드에서 추출한 키워드 중 라벨 i와 연관된 키워드 집합 K_i를 의미한다.

$$doc\_ratio(i) = \frac{\left|\{d \in D_i \mid \exists\, k \in K_i, k \in d\}\right|}{\left|D_i\right|} \quad (1)$$

이를 통해 각 라벨이 소수의 특이적인 키워드로 어느 정도 잘 요약되는지 혹은 여러 라벨에서 두루 등장하는 범용적 표현에 의해 특징이 희석되는지를 비교하였다.

표 3. F1-score 및 데이터 개수 별 doc_ratio
Table 3. doc_ratio by F1-score and the number of samples

EG99 : High F1-score & large-data label	
Abstract keywords	**doc_ratio**
Nuclear power	0.0797
Workforce development	0.0505
International cooperation	0.0442
Nuclear nonproliferation	0.0394

OB99 : Low F1-score & small-data label	
Abstract keywords	**doc_ratio**
Artificial intelligence	0.0702
Biosignals	0.0594
EEG (Brainwaves)	0.0540
Virtual reality	0.0432

ED99 : Low F1-score & large-data label	
Abstract keywords	**doc_ratio**
Artificial intelligence	0.0340
Internet of things	0.0265
Deep learning	0.0213
Sensors	0.0165

NA99 : High F1-score & small-data label	
Abstract keywords	**doc_ratio**
Physics	0.1116
Mathematics (Field)	0.1116
Mathematics (Subject)	0.0862
Talent development	0.0761

분석 결과 성능이 상대적으로 높은 "기타" 라벨들은 소수의 특이적인 키워드가 비교적 높은 doc_ratio를 보이는 경향이 있었다. 표 3을 보면 EG99, NA99와 같이 상위권에 속한 라벨들은 원자력, 국제협력, 핵비 확산, 물리, 수학 등 각 라벨이 지칭하는 의미 영역에 특화된 고유 키워드들이 상위 키워드를 구성한다. 이들 키워드는 해당 라벨 문서의 상당 비율에서 반복적으로 등장하며, 결과적으로 "기타" 코드에 속해 있음에도 텍스트 수준에서는 하나의 주제를 중심으로 응집된 의미 구조를 형성한다. 라벨명과 요약문 내 핵심어가 비교적 일관되게 대응되고 라벨별로 고유한 키워드 조합이 존재하기 때문에 모델이 각 "기타" 라벨의 결정 경계를 학습하는 데 유리한 정보를 제공하며 높은 F1-score로 이어지는 양상을 보인다.

반면 성능이 낮은 하위 "기타" 라벨들은 여러 연구 분야에 걸쳐 공통적으로 사용되는 범용적 표현의 비중이 높았다. ED99, OB99의 경우 상위 키워드가 인공지능, 사물인터넷, 딥러닝, 센서, 가상현실 등 다양한 R&D 과제에서 폭넓게 사용되는 기술 일반어로 채워져 있다, 그리고 개별 키워드의 doc_ratio도 상위 라벨들에 비해 전반적으로 낮다. 동일한 키워드가 여러 "기타" 라벨에 동시다발적으로 출현하면서도, 어느 한 라벨에만 집중되지 않기 때문에 라벨 간 의미 경계를 선명하게 구분해 주기보다는 서로를 섞어 버리는 역할을 한다. 이처럼 일반적인 표현이 넓게 분포되고 라벨 내부의 의미 일관성이 충분히 확보되지 않은 경우, 모델은 특정 "기타" 라벨에만 고유한 패턴을 학습하기 어렵고

인접 라벨로의 오분류가 증가하여 F1-score가 하위권에 머무는 경향을 보인다.

종합하면, “기타” 라벨 내부에서도 일부 라벨은 소수의 고유 키워드에 의해 비교적 잘 응집된 의미 구조를 가지는 반면, 다른 라벨은 일반어 비중이 높고 여러 주제가 혼재된 텍스트 구성을 보인다. 이는 “기타” 집단에서 관찰되는 성능 저하가 단순히 표본 수 부족 때문이 아니라, 일반어 중심 텍스트와 모호한 라벨 정의가 결합된 구조적 문제에서 비롯된 것임을 시사한다. 특히 의미 경계가 흐릿한 “기타” 라벨에 대해서는, 고유성이 높은 키워드로 특징화되는 문서를 중심으로 라벨 구조를 세분화하고 이를 반복적인 재학습과 결합하는 등의 데이터 중심 재라벨링 기법이 필요함을 시사한다.

다음 장에서는 기준 분류기와 대형 언어 모델의 예측을 결합하여 이러한 모호 라벨 집단을 재구조화하는 구체적인 방법론을 제안한다.

Ⅳ. 방법론

본 절에서는 과학기술표준분류 중분류 데이터셋에서 “기타” 라벨을 중심으로 성능 하한이 형성되는 문제를 완화하기 위해, KoBERT 기반 기준 분류기와 설명문 기반 대형 언어 모델을 결합한 반복적 재라벨링 기법을 제안한다. 제안 방법은 1) 라벨 체계 및 재라벨링 대상 정의, 2) KoBERT 기준 분류기 학습, 3) 기준 분류기를 이용한 재라벨링 후보 집합 구축, 4) GPT-5 기반 의미 검증에 따른 보수적 재라벨링, 5) 재라벨링 결과를 반영한 반복 재학습 단계로 구성된다.

4.1 라벨 체계 및 재라벨링 대상 정의

본 연구에서 사용하는 라벨 체계는 Ⅲ장에서 설명한 과학기술표준분류 중분류 체계이다. 원시 데이터에는 중분류 기준 267개의 라벨이 존재하지만, 극히 소수의 문서만을 포함하는 희소 라벨과 H 계열 라벨은 분석과 학습 안정성을 위해 ZZ 라벨로 통합하였다. 이후 남은 217개 중분류 라벨을 대상으로 KoBERT 기준 분류기를 학습하고 라벨별 성능 및 텍스트 특성을 분석하였다. 그 결과, 코드가 “99”로 끝나는 “기타” 라벨 집단에서 성능이 전반적으로 낮고 일반 라벨과의 성능 격차가 크게 나타났다. 또한 텍스트 구성 측면에서도 특정 연구 주제를 지칭하는 고유 키워드보다 여러 분야에서 공통적으로 등장하는 일반어 비중이 높아 의미 경계가 모호한 문서들이 혼재한다는 점을 확인하였다.

따라서 이후 방법론에서는 기존 라벨 코드가 ‘99’로 끝나는 문서들을 재라벨링의 직접적인 대상으로 설정하고, 나머지 일반 라벨 문서들은 KoBERT 기준 분류기의 초기 학습 및 재학습을 위한 배경 데이터로 활용한다.

4.2 KoBERT 기준 분류기 학습

본 연구에서 기준 분류기는 한국어 사전학습 언어모델인 KoBERT를 기반으로 구현하였다. KoBERT는 대규모 한국어 말뭉치를 대상으로 학습된 BERT 계열 모델로 한국어 복합명사와 과학기술 용어를 안정적으로 처리할 수 있는 토크나이저와 어휘 집합을 사용한다. 과학기술표준분류 텍스트는 한글 복합명사, 약어, 전문 용어가 혼재된 형태를 띠고 다국어 BERT나 영어 중심 모델보다 한국어 전용 모델을 사용하는 편이 라벨별 의미 차이를 포착하는 데 유리하다. 또한 KoBERT는 국내 한국어 텍스트 분류 태스크에서 기준선 모델로 널리 활용되고 있으며, 파라미터 규모와 계산 비용이 비교적 작아 반복적인 재라벨링과 재학습을 수행하는 본 연구에 적합하다고 판단했다.

입력은 앞서 정의한 통합 요약 텍스트(merged_text)를 사용하며 각 라벨에 대한 확률 분포를 계산하도록 설계하였고, 데이터는 과제 단위로 무작위 분할하여 약 8:2 비율로 학습-검증 세트를 구성 하였다.

4.3 재라벨링 후보 집합 구축

재라벨링 대상은 기존 중분류 코드가 “99”로 끝나는 기타 라벨 문서들이다. 본 연구에서는 이들 문서를 모두 재라벨링 후보 집합으로 두고, 각 문서에

대해 기준 분류기와 GPT-5의 예측을 동시에 얻은 뒤, 두 예측의 관계에 따라 서로 다른 유형의 재라벨링 데이터를 구성한다.

먼저, 후보 집합에 속한 각 기타 문서에 대해 KoBERT 기준 분류기가 출력하는 라벨별 확률 분포를 계산한다. 이때 “기타” 코드는 후보 라벨에서 제외하고, 일반 라벨만을 대상으로 가장 높은 확률을 갖는 라벨과 그 확률을 기준 분류기의 Top-1 예측으로 사용한다.

동시에 동일한 문서에 대해 설명문 기반 GPT-5 분류를 수행한다. 입력 문서와 함께 과학기술표준분류 해설서를 제공하여 GPT-5가 가장 적합하다고 판단하는 라벨을 선택하도록 한다.

이 단계까지는 모든 기타 문서가 후보 집합에 포함된 상태에서 각 문서마다 (기준 분류기 예측 라벨, GPT-5 예측 라벨) 쌍이 부여된다고 볼 수 있다.

이후 이 쌍의 일치 여부와 신뢰도를 기준으로 두 가지 유형의 재라벨링 데이터(A/B)를 구성한다.

4.4 LLM 기반 의미 검증 및 재라벨링

재재라벨링 대상은 기존에 중분류 코드가 “99”로 끝나는 기타 라벨 문서들이며, 각 문서에 대해 KoBERT 기준 분류기의 Top-1과 GPT-5의 예측을 모두 얻는다. 그 후에 그림 1과 같이 예측을 결합하여 최종 재라벨링 데이터를 구성한다.

우선 각 과제에 대한 문서와 과학기술표준분류 해설서 그리고 KoBERT가 제안한 후보 일반 라벨에 대해서 GPT-5에 입력한다. 프롬프트는 “해설서를 참고하여 주어진 문서가 어떤 라벨 설명과 가장 잘 부합하는지 선택하고 선택한 라벨 코드 후보 5개까지 포함하여 출력하라”는 형태로 구성하여, GPT-5가 문서에 대해 후보 라벨까지 포함하여 제시하도록 했다. 이렇게 얻은 GPT-5의 선택 라벨을 KoBERT의 Top-1 라벨과 짝지어 문서별 쌍을 구성한다.

이후 두 예측의 관계에 따라 재라벨링 데이터는 두 가지 유형으로 나눈다. 집합 A는 KoBERT의 Top-1 라벨과 GPT-5가 선택한 라벨이 일치하는 문서들로 구성되는 모델 간 완전 합의 기반 재라벨링 데이터이다. 두 모델이 서로 다른 구조와 학습 데이터에 기반해 동의한 경우이므로 별도의 확률 임계값 없이도 상대적으로 신뢰도가 높은 후보로 간주하고, 기존 기타 라벨을 이 공통 라벨로 치환한다. 이렇게 얻은 문서들은 이후 실험에서 “완전 일치 기반 재라벨링 데이터(A 데이터)”로 사용한다.

집합 B는 GPT-5 신뢰도 기반 재라벨링 데이터로, GPT-5가 특정 라벨에 대해 충분히 높은 확률을 부여한 문서들로 구성된다. 여기에는 KoBERT의 Top-1 라벨과 GPT-5 라벨이 일치하면서 GPT-5가 해당 라벨에 대해 임계값 이상의 확률을 부여한 문서와 두 모델의 Top-1 라벨이 서로 다르더라도 GPT-5가 제안한 라벨의 확률이 임계값 이상인 문서가 모두 포함된다. 임계값은 GPT-5의 예측 확률 구간별로 일부 문서를 샘플링한 뒤 각 문서의 요약 텍스트와 해설서 상의 라벨 설명, GPT-5가 부여한 확률을 함께 눈으로 검토하면서 “이 수준이면 GPT-5 예측만으로도 자동 재라벨링에 활용할 수 있다”고 판단되는 최소 값으로 설정하였다. 실제 재라벨링에서는 GPT-5가 선택한 라벨의 확률이 이 임계값 이상인 경우에 한해 기존 기타 라벨을 GPT-5 라벨로 치환한다. 이렇게 얻은 샘플을 “GPT-5 신뢰도 기반 재라벨링 데이터(B 데이터)”로 정의하여 별도로 관리한다.

본 연구에서 ‘합의 실패’는 KoBERT와 GPT-5의 최종 예측 라벨이 불일치하면서 GPT-5 최대 확률이 임계값 τ 미만이거나 GPT-5 출력이 사전 정의된 라벨 집합 밖인 경우로 정의한다. 또한 GPT-5 응답 파싱 실패, 빈 응답, 다중 라벨 반환 등 예외가 발생한 경우에도 동일하게 합의 실패로 처리한다. 합의 실패 샘플은 자동 재라벨링을 수행하지 않고 기존 라벨을 유지하며 후속 분석을 위해 별도의 검토 대상 로그로 분리 저장한다. 한편 자동 재라벨링 결과의 신뢰성은 ‘보장’하기보다 오류 가능성이 높은 샘플을 보수적으로 배제하는 방식으로 위험을 완화하도록 설계하였다. 즉 모델 간 예측이 불일치하더라도 GPT-5 최대 확률이 τ 이상인 경우에 한해 선택적으로 재라벨링을 수행하며, 재라벨링 과정(원문, 두 모델 예측, 확률값, 최종 라벨)을 함께 기록하여 추적 및 사후 점검이 가능하도록 하였다.

재라벨링에 사용되지 않은 나머지 기타 문서들은 원래 라벨을 유지한 채 배경 데이터로 남겨 둔다. 이후 실험에서는 각 A, B 데이터셋에 대하여 KoBERT 기준 분류기의 재학습 성능과 기타 라벨 집단의 F1-score 개선 정도를 비교, 분석한다. 이를 통해 KoBERT - GPT-5 이중 예측 구조가 기타 라벨의 모호한 의미 경계를 재구조화하고 전체 분류 성능의 하한을 완화하는 데 어느 정도 기여하는지 정량적으로 평가한다.

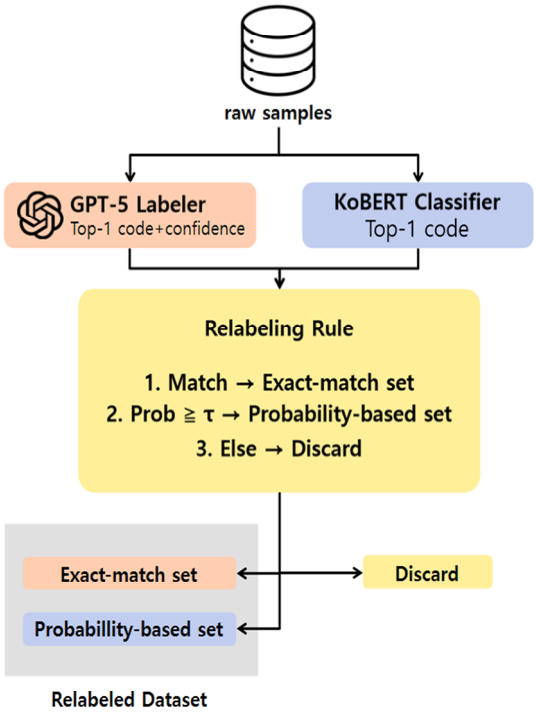

그림 1. 기타 데이터라벨링 절차
Fig. 1. Labeling workflow for the "Other" category

4.5 재학습

재라벨링한 데이터셋을 바탕으로 KoBERT 분류기를 다시 학습하고 동일한 설정으로 성능을 평가한다. 평가 지표는 전체 Macro F1 및 Macro Recall 뿐 아니라, 특히 모호 라벨 집단과 하위 성능 라벨군의 F1-score 변화를 중심으로 분석한다.

V. 실험 및 분석

5.1 평가 기준

실험에서 데이터 불균형과 기타 라벨의 특성을 함께 고려할 수 있도록 여러 지표를 사용하였다. 전반적인 분류 성능을 비교하기 위해 Top-1 / Top-3 / Top-5 Accuracy와 Macro F1-score를 사용하였다. Top-1 Accuracy는 모델이 가장 높은 확률로 예측한 라벨이 실제 정답과 일치하는 비율이며, Top-3와 Top-5 Accuracy는 각각 상위 3개, 5개의 예측 후보 중에 정답 라벨이 포함되는 비율을 의미한다. 이는 재라벨링으로 인해 정답 라벨이 상위 후보군 내에서 얼마나 더 잘 포착되는지를 확인하기 위한 지표이다.

기타 라벨 집단에 대해서는 라벨별 F1-score 변화를 추가로 분석하여 GPT-5 확률 정보를 활용한 재라벨링이 OB99(기타 인지/감성과학), EC99(기타 화공), EA99(기타 기계), EG99(기타 원자력) 등 대분류별 기타 라벨의 예측 성능을 어떻게 변화시키는지 상세하게 살펴보았다. 이를 통해 전체 Accuracy 및 Macro F1뿐만 아니라, 의미가 불분명하고 데이터가 희소한 기타 라벨의 성능 하한선이 실제로 개선되는지를 정량적으로 평가하였다.

5.2 실험 결과

아래 표 4는 기준 KoBERT 분류기와 제안한 재라벨링 데이터를 이용해 재학습한 모델들의 성능을 비교한 결과를 나타낸다. 기준 모델은 원본 라벨만을 사용해 학습한 경우로 검증 기준 Accuracy 0.7839 Macro F1 0.7527 Top-3와 Top-5 Accuracy는 각각 0.8834와 0.9161 수준을 보였다.

표 5는 GPT-5 예측에 기반한 재라벨링 과정에서 중분류 코드가 실제로 변경된 과제의 사례를 보여준다. 기존에 포괄적 코드로 묶여 있던 과제들이 재라벨링을 통해 식품 과학, 에너지 기계 시스템 등 보다 구체적인 중분류로 이동했음을 확인할 수 있다.

표 4. KoBERT 및 재라벨링 데이터 성능 비교
Table 4. Performance comparison between KoBERT and the relabeled dataset

Model setting	τ	Acc (Top1)	Acc (Top3)	Acc(Top5)	Macro-F1
KoBERT	-	0.7839	0.8834	0.9161	0.7527
KoBERT + A	-	0.7755	0.8778	0.9112	0.7380
KoBERT+ B(0.8)	0.8	0.7907	0.8903	0.9203	0.7630
KoBERT + B(0.85)	0.85	0.7849	0.8892	0.9190	0.7580
KoBERT + B(0.9)	0.9	0.7847	0.8858	0.9160	0.7547

표 5. 재라벨링된 데이터 예시
Table 5. Examples of relabeled data

Project title (Summary)	Mid-level code (Before relabeling)	Mid-level code (After relabeling)
Development of a novel cap and container with an automatic makgeolli gas-release function	OB99 (Other cognitive/affective science)	LB17 (Food science)
Development of a high-performance intelligent industrial pellet boiler (3,000,000 kcal/hr class)	EA99 (Other mechanical engineering)	EA07 (Energy/environmental mechanical systems)

KoBERT와 GPT-5의 예측 라벨이 완전히 일치하는 문서만을 대상으로 라벨을 수정한 A 데이터(수정 샘플 485건)를 재학습한 경우 전반적인 성능은 기준 모델과 비슷한 수준이지만 모든 지표에서 약간씩 낮아졌다. 재라벨링된 샘플 수가 전체 대비 매우 적고 일부 라벨 치환이 기존 분포와 완전히 맞지 않아 모델의 결정 경계를 개선하기보다는 오히려 미세하게 흔드는 효과가 더 크게 나타난 것으로 해석할 수 있다. 보수적으로 선별한 소량의 합의 샘플만으로는 전체 분류 성능을 끌어올리기에는 한계가 있었다고 판단했다.

반면 GPT-5가 특정 라벨에 대해 높은 확률을 부여한 문서만 선별하여 라벨을 수정하는 방식(B 데이터)에 대해서는, 확률 임계값 τ를 0.8, 0.85, 0.9로 설정한 세 가지 설정을 비교하였다. 여기서 τ는 GPT-5가 특정 라벨에 부여한 예측 확률의 하한을 의미한다. 전체 62만여 개 샘플 중에서 이 조건을 만족하는 문서는 각각 약 1.1만, 0.9만, 0.7만 건 수준이었으며, 해당 샘플의 라벨이 GPT-5 예측 결과로 교체되었다. 세 설정 Top-3/Top-5 Accuracy가 전반적으로 향상되는 경향을 보였으며, 특히 τ가 0.8인 설정(이하 B 데이터)이 가장 큰 성능 향상을 보였다. τ = 0.8 설정에서는 상대적으로 많은 고신뢰 샘플을 활용하면서도 노이즈를 일정 수준 이하로 억제하여 Macro F1과 Top-k 정확도 모두에서 기준 모델과 다른 임계값 설정보다 높은 값을 기록하였다.

τ = 0.9 설정은 학습 손실과 성능이 기준 모델과 유사한 수준으로 안정적으로 수렴해 성능을 해치지 않는 보수적 재라벨링에 가깝고, τ = 0.85 설정은 두 설정 사이에서 완만한 개선을 보이는 형태였다. 즉 전체 데이터 중 일부만 재라벨링했음에도 GPT-5 확률 정보를 이용해 τ = 0.8에서 충분히 많은 고신뢰 샘플을 확보해 학습에 반영했을 때 기준 KoBERT를 가장 크게 상회하는 성능을 얻을 수 있었다.

KoBERT와 GPT-5 완전 일치 샘플만을 사용한 매우 보수적인 재라벨링(A 데이터)이 뚜렷한 이득을 제공하지 못한 것과 달리, GPT-5 예측 확률을 활용해 여러 τ 값(0.8, 0.85, 0.9)을 탐색한 B 데이터 실험에서는 τ = 0.9 설정에서 성능을 희생하지 않는 안정성, τ = 0.8 설정에서 가장 큰 성능 향상이 확인되었다. 이는 GPT-5 기반 의미 검증을 통해 τ = 0.8에서 선택된 고신뢰 재라벨링 샘플을 중심으로 KoBERT를 재학습하는 방법이 기존 라벨 품질을 훼손하지 않으면서도 전체 분류 성능의 하한을 완화하고 상한을 끌어올리는 데 실질적으로 기여할 수 있음을 시사한다.

5.3 기타 라벨별 성능 변화

B 데이터의 임계값 0.8 설정 재라벨링이 기타 라벨 집단에 미친 영향을 확인하기 위해, 코드가 “99”로 끝나는 기타 라벨 19개에 대해 재라벨링 전후 F1-score를 비교하였다. 그 결과는 표 6과 같다. 표 6에서 확인할 수 있듯이, 기타 라벨 19개의 평균 F1-score는 원본 데이터 기준 0.26에서 재라벨링 후 0.29로 증가하였으며, 전체 평균 기준 약 0.03포인트의 성능 향상이 나타났다. 이는 GPT-5 신뢰도 기반 재라벨링이 기타 라벨 집단 전체의 분류 성능을 전반적으로 개선하는 방향으로 작용했음을 보여준다.

표 6. 기타 라벨 성능 변화
Table 6. Performance changes for the “Other” label

Label	Original dataset F1-score	Relabeled dataset F1-score	F1-score change rate
EA99	0.1152	0.147	+0.0318
EB99	0.1019	0.1588	+0.0569
EC99	0.0222	0.1687	+0.1465
ED99	0.1414	0.1559	+0.0145
EE99	0.2434	0.2511	+0.0077
EF99	0.196	0.2198	+0.0238
EG99	0.5948	0.5641	-0.0307
EH99	0.2585	0.2979	+0.0394
EI99	0.3036	0.3909	+0.0873
LA99	0.2349	0.2481	+0.0132
LB99	0.2538	0.2311	-0.0227
LC99	0.3168	0.332	+0.0152
NA99	0.5532	0.7442	+0.191
NB99	0.25	0.2143	-0.0357
NC99	0.1256	0.1157	-0.0099
ND99	0.3478	0.3333	-0.0145
OA99	0.3256	0.3636	+0.038
OB99	0	0.0769	+0.0769
OC99	0.5338	0.575	+0.0412
Overall average	0.26	0.29	+0.03

라벨별로 보면, 전체 19개 기타 라벨 중 14개 라벨에서 F1-score가 상승하였다. 특히 NA99는 0.5532에서 0.7442로 증가하여 가장 큰 성능 향상을 보였고, EC99 역시 0.0222에서 0.1687로 상승하여 기존에 낮은 성능을 보이던 라벨에서 뚜렷한 개선이 나타났다. 또한 EI99, OB99, EB99, OC99 등도 재라벨링 이후 F1-score가 비교적 크게 증가하였다. 이는 GPT-5가 문서 의미와 라벨 설명 간의 대응 관계를 활용하여, 일부 기타 라벨에 포함되어 있던 문서들을 보다 일관된 의미 범주로 재구성하는 데 기여했기 때문으로 해석할 수 있다.

반면 EG99, LB99, NB99, NC99, ND99의 5개 라벨에서는 재라벨링 이후 F1-score가 소폭 감소하였다. 이들 라벨은 모두 “기타(99)”에 해당하는 중분류 코드로, 라벨 내부에 다양한 연구 주제와 범용적인 기술 용어가 혼재되어 있는 경우가 많다. 따라서 일부 문서에 대해 GPT-5가 높은 신뢰도를 부여하더라도, 해당 라벨 자체의 의미 경계가 넓고 모호하기 때문에 재라벨링이 항상 성능 개선으로 이어지지는 않은 것으로 볼 수 있다.

특히 이러한 라벨들은 재라벨링 과정에서 텍스트 입력 자체는 거의 동일하게 유지된 반면, 라벨의 의미 범위가 넓은 “기타” 범주로 재정의되면서 라벨 내부 분산이 증가할 가능성이 있다. 이 경우 모델은 해당 기타 라벨을 하나의 일관된 의미 범주로 학습하기 어렵고, 인접한 일반 중분류 라벨과의 경계 역시 불명확해질 수 있다. 그 결과 일부 라벨에서는 정밀도와 재현율이 동시에 저하되며 F1-score 감소로 나타난 것으로 해석된다.

GPT-5 확률 정보를 활용한 재라벨링은 의미 중심이 비교적 뚜렷한 기타 라벨에서는 F1-score를 유의미하게 향상시키는 반면, 본질적으로 의미 범위가 넓고 내부 이질성이 큰 기타 라벨에서는 개선 효과가 제한적이었다. 그럼에도 표 6에서 전체 기타 라벨 평균 F1-score가 상승하고, 기존에 성능이 매우 낮았던 일부 라벨이 최소한의 예측 가능성을 회복했다는 것을 바탕으로 제안한 재라벨링 절차는 기타 라벨 집단의 성능 하한선을 끌어올리는 데 일정 부분 기여한 것으로 볼 수 있다.

VI. 결 론

본 연구는 과학기술표준분류 체계에서 반복적으로 발생하는 ‘기타’ 라벨의 모호성과 노이즈 문제를 완화하기 위해 KoBERT 기반 분류기와 대형 언어 모델(LLM)을 결합한 자동 재라벨링 프레임워크

를 제안하였다. 핵심 아이디어는 LLM을 단순 예측기가 아니라 의미 검증기로 활용하고, 예측 확률(신뢰도) 임계값(τ)을 통해 고신뢰 문서만 선별하여 재학습에 반영함으로써 대규모 수작업 검수 없이도 데이터 품질과 분류 안정성을 동시에 끌어올리는 데 있다.

실험 결과, 제안 방식은 Macro F1 및 Top-k 정확도를 기존 KoBERT 단독 학습 대비 일관되게 개선하였으며, 특히 기존에 F1-score가 0에 가깝던 극단적으로 취약한 라벨들이 재라벨링 이후 일정 수준의 예측 가능성을 회복하는 등 전체 분류기의 하한선 성능을 끌어올리는 효과를 확인하였다. 또한 정성 분석을 통해 포괄적인 '기타' 코드로 묶여 있던 과제들이 재라벨링을 거쳐 식품과학, 농화학 등 더 구체적인 중분류로 이동하며 라벨의 의미적 일관성이 강화됨을 확인하였다. 즉, 본 연구는 '기타' 라벨을 방치하지 않고 자동 정제·재편입 가능한 대상으로 전환함으로써, 분류 체계 운영의 실무적 부담을 줄이면서도 모델 성능과 데이터 해석 가능성을 향상시키는 실질적인 전략을 제시했다는 의의가 있다.

다만 본 접근은 몇 가지 한계를 갖는다. 전기/전자(ED99), 원자력(EG99) 등 내부 주제가 지나치게 이질적으로 혼재된 일부 라벨에서는 재라벨링 이후에도 성능 개선이 미미하거나 소폭 하락하는 현상이 관찰되었고, 사용한 임계값(τ)과 프롬프트 구성이 소규모 사전 탐색 기반의 경험적 설정이라는 점에서 도메인 변화 시 최적성을 보장하기 어렵다. 또한 특정 분류기(KoBERT)와 단일 LLM(GPT-5) 조합만을 고려했으므로, 다양한 모델 조합 및 환경에서의 재현성 검증이 추가로 필요하다. 더불어 합의 및 확률 임계값 기반 선별로 오류 가능성을 줄이도록 설계했더라도, 두 모델이 같은 방향으로 오분류하는 경우를 원천적으로 차단할 수는 없어 결과 신뢰성은 데이터 특성과 설정에 영향을 받을 수 있다.

향후 연구에서는 "기타" 라벨 내부를 의미적 군집으로 분해해 인접 코드로 재편입하거나, 1단계(기타 판별)와 2단계(세부 예측)로 나누는 계층적 분류 구조를 도입할 필요가 있다. 또한 불확실도 기반 가중치 조정, 커리큘럼 학습 등 불안정 라벨 특화 학습 전략을 결합해 성능 정체 라벨의 개선 효율을 높이고, 다양한 과학기술 정책 영역으로 실험을 확장하는 것이 요구된다. 나아가 최적의 임계값(τ)과 프롬프트를 자동 탐색하는 알고리즘을 도입하고, 표본 기반 점검, 추가 모델 검증, 보수적 적용 정책 등으로 검증 체계를 보완함으로써 사람의 개입을 최소화한 자동화 파이프라인으로 발전시키는 것이 목표가 될 수 있다.

References

[1] F. Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol. 34, No. 1, pp. 1-47, Mar. 2002. https://doi.org/10.1145/505282.505283.

[2] C. C. Aggarwal and C. Zhai, "A Survey of Text Classification Algorithms", Mining Text Data, Springer, pp. 163-222, Jan. 2012. https://doi.org/10.1007/978-1-4614-3223-4_6.

[3] Y. Noh, J. Yang, J. Kang, Y. Kim, and J. Lee, "A Study on the Improvement of the Revision Process Based on the Analysis of the Current Status and Relevant Issues of the Academic Standard Classification System", The Journal of Transdisciplinary Studies, Vol. 7, No. 1, pp. 1-19, Apr. 2023. https://doi.org/10.22685/JTS.2023.7.1.001.

[4] B. Frénay and M. Verleysen, "Classification in the Presence of Label Noise: A Survey", IEEE Transactions on Neural Networks and Learning Systems, Vol. 25, No. 5, pp. 845-869, May 2014. https://doi.org/10.1109/TNNLS.2013.2292894.

[5] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, "A Closer Look at Memorization in Deep Networks", Proc. International Conference on Machine Learning (ICML), Sydney, Australia, pp. 233-242, Aug. 2017. arXiv:1706.05394. https://doi.org/10.48550/arXiv.1706.05394.

[6] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning From Noisy Labels With Deep

Neural Networks: A Survey", IEEE Transactions on Neural Networks and Learning Systems, Vol. 34, No. 11, pp. 8135-8153, Nov. 2023. https://doi.org/10.1109/TNNLS.2022.3152527.

[7] Y. Zhang, G. Niu, and M. Sugiyama, "Learning Noise Transition Matrix from Only Noisy Labels via Total Variation Regularization", Proc. International Conference on Machine Learning (ICML), Online, pp. 12501-12512, Jul. 2021. https://doi.org/10.48550/arXiv.2102.02414.

[8] G. Zheng, A. H. Awadallah, and S. Dumais, "Meta Label Correction for Noisy Label Learning", Proc. AAAI Conference on Artificial Intelligence (AAAI), Online, Vol. 35, No. 12, pp. 11053-11061, Feb. 2021. https://doi.org/10.1609/aaai.v35i12.17319.

[9] Y. Wu, J. Shu, Q. Xie, Q. Zhao, and D. Meng, "Learning to Purify Noisy Labels via Meta Soft Label Corrector", Proc. AAAI Conference on Artificial Intelligence (AAAI), Online, Vol. 35, No. 12, pp. 10388-10396, Feb. 2021. https://doi.org/10.1609/aaai.v35i12.17244.

[10] Y. Wang, H. Pang, Y. Qin, S. Wei, and Y. Zhao, "Adaptive Estimation of Instance-Dependent Noise Transition Matrix for Learning with Instance-Dependent Label Noise", Neural Networks, Vol. 188, Art no. 107464, Aug. 2025. https://doi.org/10.1016/j.neunet.2025.107464.

[11] J.-H. Kim, S. Yun, and H. O. Song, "Neural Relation Graph: A Unified Framework for Identifying Label Noise and Outlier Data", Advances in Neural Information Processing Systems (NeurIPS), Vol. 36, pp. 43754-43779, Dec. 2023. https://doi.org/10.52202/075280-1898.

[12] T. Kim, J. Ko, S. Cho, J. Choi, and S.-Y. Yun, "FINE Samples for Learning with Noisy Labels", Advances in Neural Information Processing Systems (NeurIPS), Online, Vol. 34, pp. 24137-24149, Dec. 2021. https://doi.org/10.48550/arXiv.2102.11628.

[13] T. Guo, "The Re-Label Method For Data-Centric Machine Learning", arXiv preprint arXiv:2302.04391, Feb. 2023. https://doi.org/10.48550/arXiv.2302.04391.

[14] M. Wang, I. Valmianski, X. Amatriain, and A. Kannan, "Learning Functional Sections in Medical Conversations: Iterative Pseudo-Labeling and Human-in-the-Loop Approach", Proc. 8th Machine Learning for Healthcare Conference (MLHC), Proceedings of Machine Learning Research, Vol. 219, pp. 772-787, Aug. 2023. arXiv:2210.02658. https://doi.org/10.48550/arXiv.2210.02658.

[15] C. Pan, J. Huang, J. Gong, and X. Yuan, "Few-Shot Transfer Learning for Text Classification With Lightweight Word Embedding Based Models", IEEE Access, Vol. 7, pp. 53296-53304, Apr. 2019. https://doi.org/10.1109/ACCESS.2019.2911850.

[16] K. Bailey and S. Chopra, "Few-Shot Text Classification with Pre-Trained Word Embeddings and a Human in the Loop", arXiv preprint arXiv:1804.02063, Apr. 2018. https://doi.org/10.48550/arXiv.1804.02063.

[17] D.-H. Lee, "Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks", Proc. ICML 2013 Workshop on Challenges in Representation Learning (WREPL), Atlanta, GA, USA, pp. 1-6, Jul. 2013.

[18] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Ordonez, "Curriculum Labeling: Revisiting Pseudo-Labeling for Semi-Supervised Learning", Proc. AAAI Conference on Artificial Intelligence (AAAI), Online, Vol. 35, No. 8, pp. 6912-6920, Feb. 2021. https://doi.org/10.1609/aaai.v35i8.16852.

[19] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels", Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 4238-4247, Jun. 2022. https://doi.org/10.1109/CVPR

52688.2022.00421.

[20] M. Wan, T. Safavi, S. K. Jauhar, Y. Kim, S. Counts, J. Neville, S. Suri, C. Shah, R. W. White, L. Yang, R. Andersen, G. Buscher, D. Joshi, and N. Rangan, "TnT-LLM: Text Mining at Scale with Large Language Models", Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 5836-5847, Aug. 2024. https://doi.org/10.1145/3637528.3671647.

저자소개

권 수 민 (Sumin Kwon)

2023년 3월 ~ 현재 : 국립금오공과대학교 컴퓨터공학부 인공지능공학전공 학사과정

관심분야 : 자연어처리, 데이터 분석, 대규모 언어 모델

곽 지 호 (Jiho Gwak)

2025년 2월 : 국립금오공과대학교 컴퓨터공학부(공학사)

2025년 3월 ~ 현재 : 국립금오공과대학교 컴퓨터·AI융합공학과 석사과정

관심분야 : 딥러닝, 자연어처리, 텍스트 분류

박 지 영 (Jiyoung Park)

1998년 2월 : 서강대학교 영문학과(문학사)

2009년 8월 : 서울대학교 인지과학 석박사통합과정(이학박사)

2014년 4월 ~ 현재 : 한국과학기술정보연구원 (KISTI) 선임연구원

관심분야 : 뇌과학, 인지정보처리, 정보분석, 소셜 네트워크 분석

정 유 철 (Yuchul Jung)

2011년 2월 : 한국과학기술원 (KAIST) 전산학과(공학박사)

2009년 1월 ~ 2013년 7월 : 한국전자통신연구원 (ETRI) 연구원/선임연구원

2013년 8월 ~ 2017년 8월 : 한국과학기술정보연구원 (KISTI) 선임연구원

2017년 8월 ~ 2022년 9월 : 국립금오공과대학교 컴퓨터공학과 조교수

2022년 10월 ~ 현재 : 국립금오공과대학교 컴퓨터공학부 부교수

관심분야 : 거대언어모델, 자연어처리, 지식그래프, 한국어 음성 인식/합성, AI응용