KIPS Trans. Softw. and Data Eng.
Vol.12, No.4 pp.159~172
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2023.12.4.159

Building Sentence Meaning Identification Dataset Based on Social Problem-Solving R&D Reports

Hyeonho Shin[†] · Seonki Jeong[††] · Hong-Woo Chun[†††] · Lee-Nam Kwon[††††] · Jae-Min Lee[†††††] · Kanghee Park[††††††] · Sung-Pil Choi[†††††††]

ABSTRACT

In general, social problem-solving research aims to create important social value by offering meaningful answers to various social pending issues using scientific technologies. Not surprisingly, however, although numerous and extensive research attempts have been made to alleviate the social problems and issues in nation-wide, we still have many important social challenges and works to be done. In order to facilitate the entire process of the social problem-solving research and maximize its efficacy, it is vital to clearly identify and grasp the important and pressing problems to be focused upon. It is understandable for the problem discovery step to be drastically improved if current social issues can be automatically identified from existing R&D resources such as technical reports and articles. This paper introduces a comprehensive dataset which is essential to build a machine learning model for automatically detecting the social problems and solutions in various national research reports. Initially, we collected a total of 700 research reports regarding social problems and issues. Through intensive annotation process, we built totally 24,022 sentences each of which possesses its own category or label closely related to social problem-solving such as problems, purposes, solutions, effects and so on. Furthermore, we implemented four sentence classification models based on various neural language models and conducted a series of performance experiments using our dataset. As a result of the experiment, the model fine-tuned to the KLUE-BERT pre-trained language model showed the best performance with an accuracy of 75.853% and an F1 score of 63.503%.

Keywords : Social Problem-Solving Research, Natural Language Process, Data Building, Pre-trained Language Model

사회문제 해결 연구보고서 기반 문장 의미 식별 데이터셋 구축

신 현 호[†] · 정 선 기[††] · 전 홍 우[†††] · 권 이 남[††††] · 이 재 민[†††††] · 박 강 희[††††††] · 최 성 필[†††††††]

요 약

일반적으로 사회문제 해결 연구는 과학기술을 활용하여 다양한 사회적 현안들에 의미있는 해결 방안을 제시함으로써 중요한 사회적 가치를 창출하는 것을 연구 목표로 한다. 그러나 사회문제와 쟁점을 완화하기 위하여 많은 연구들이 국가적으로 수행되었음에도 불구하고 여전히 많은 사회문제가 남아 있는 상황이다. 사회문제 해결 연구의 전 과정을 원활하게 하고 그 효과를 극대화하기 위해서는 사회적으로 시급한 현안들에 대한 문제를 명확하게 파악하는 것이 중요하다. 사회문제 해결과 관련된 기존 R&D 보고서와 같은 자료에서 중요한 사안을 자동으로 식별할 수 있다면 사회문제 파악 단계가 크게 개선될 수 있다. 따라서 본 논문은 다양한 국가 연구보고서에서 사회문제와 해결방안을 자동으로 감지하기 위한 기계학습 모델을 구축하는 데에 필수적인 데이터셋을 제안하고자 한다. 우선 데이터를 구축하기 위해 사회문제와 쟁점을 다룬 연구보고서를 총 700건 수집하였다. 수집된 연구보고서에서 사회문제, 목적, 해결 방안 등 사회문제 해결과 관련된 내용이 담긴 문장을 추출 후 라벨링을 수행하였다. 또한 4개의 사전학습 언어모델을 기반으로 분류 모델을 구현하고 구축된 데이터셋을 통해 일련의 성능 실험을 수행하였다. 실험 결과 KLUE-BERT 사전학습 언어모델을 미세조정한 모델이 정확도 75.853%, F1 스코어 63.503%로 가장 높은 성능을 보였다.

키워드 : 사회문제 해결 연구, 자연어처리, 데이터구축, 사전학습 언어모델

※ 이 논문은 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2018R1D1A1B07048839).
† 준 회 원 : 경기대학교 문헌정보학과 석사
†† 비 회 원 : 경기대학교 문헌정보학과 석사과정
††† 비 회 원 : 한국과학기술정보연구원 미래기술분석센터장
†††† 비 회 원 : 한국과학기술정보연구원 기술지능연구팀 책임연구원
††††† 비 회 원 : 한국과학기술정보연구원 기술지능연구팀 팀장
†††††† 비 회 원 : 한국과학기술정보연구원 기술지능연구팀 선임연구원
††††††† 비 회 원 : 경기대학교 문헌정보학과 부교수
Manuscript Received : December 20, 2022
First Revision : January 25, 2023
Accepted : January 30, 2023
* Corresponding Author : Sung-Pil Choi(spchoi@kgu.ac.kr)

1. 서 론

과학기술의 역할이 과거와 다르게 경제성장뿐만 아니라 환경, 건강, 안전 등 국민 생활 및 사회문제 해결에 중요한 요소가 되며 그 역할에 대한 필요성이 부각되고 있다. 이에 따라, 2000년대 이후 선진국은 선진화된 과학기술 역량을 바탕으로 경제성장과 삶의 질 향상을 위한 과학기술을 지향하면서 사회문제 해결 연구 개념을 도입하였다. 국내에서도 과학기술정보통신부(구 미래창조과학부)가 2013년부터 사회문제

해결형 기술개발 사업을 시행하면서, 사회구실 통합기획, 최종사용자가 연구개발 활동에 참여하는 리빙랩, 사회적 활용을 위한 R&D 등과 같은 사회문제 해결 연구를 도입해왔다[1].

이에 공공연구기관에서 사회문제 해결 연구에 관심을 가지고 연구개발사업을 진행하고 있다. 하지만 『과학기술 기반 사회문제 해결 종합실천계획』에 의거 2014년~2018년에 착수되어 추진된 사회문제 해결 실천과제에 대한 이행점검결과에 의하면 과학기술의 사회적 역할 확대를 위해 사회문제 해결 연구가 국가 주도로 확대 추진되고 일부 연구과제에서 제품·서비스 개발 및 실증 초기 단계에 해당하는 성과 창출이 있었으나, 전반적으로 사회적 인지도가 낮아 실제로 체감되는 사회문제 해결 성과는 미흡하다는 평가를 받았다[2]. 이에 따라 정부와 공공기관은 문제 발굴 체계를 구축하고 지속적으로 성과 창출을 독려하고 있지만, 연구개발과 사회적 수요 간의 연계가 이루어지지 않아 R&D 그 자체에 그치면서 실질적인 성과 창출에 있어서 한계가 있다[3].

사회문제 해결 연구는 문제발굴과 정의(발굴·기획), 문제해결을 위한 대안 개발(운영 관리 및 평가), 대안을 적용해서 문제를 해결하는데 필요한 조직 구성으로 연구 과정이 구성되어 있다. 사회문제 해결 연구의 목표는 사회문제 해결을 통한 사회적 가치창출이므로 실질적인 성과 창출을 위해서 실수요자에게 필요한 문제발굴 과정이 중요하다[2]. 즉 사회문제 해결 연구가 사회적 가치를 창출하기 위해서는 현존하는 사회문제와 이슈를 명확히 판단하여 시민사회가 체감할 수 있는 연구 목적과 목표를 설정해야 한다.

현존하는 사회문제와 이에 대한 연구성과를 명확히 판단하기 위해서는 사회문제 관련 당사자들과의 협의를 통한 문제 파악과 구체화가 요구된다. 이를 위해서 기존 사회문제 해결 연구보고서를 통해 사회문제와 해결방안을 통합하여 구조화할 필요가 있다[2].

사회문제 해결 연구보고서의 경우 국가적 연구개발 전략에 기반하여 사회적 이슈 해결을 위한 R&D 수행의 결과물로서 Table 1과 같이 연구의 목표, 사회 이슈, 연구 문제점, 해결방안 등이 구체적이고 체계적으로 기술된 최고 품질의 전문적 정보 자원이다. 또한 연구보고서의 메타데이터뿐만 아니라 PDF 본문 자체도 공개되어 자유롭게 분석할 수 있는 공공재(Public Resource)적 성격을 띤다.

Table 1. Example of Research Key Sentences in a Research Report

Categories	Example
Need for Research	• 사건 사고 시 정확한 위치를 파악하지 못하여 경찰 현장 출동 후 즉각적인 조치에 애로사항이 있으며, 골든타임 확보는 국민의 안전과 직결됨
Research Objective	• 앱 설치 없이 신고자의 실시간 영상과 위치를 획득하여 빠른 신고 접수와 효율적인 의사 결정을 지원하는 '보이는 112 긴급 신고 지원 시스템' 개발 및 구축
Significance of Study	• 실제 사건에 활용할 수 있도록 기존 112 운영시스템과 상호 연계 구축 • 기존 상황실, 태블릿, 모바일에서 구동할 수 있도록 프로그램 연계 개발 • 안전 사회 구현 및 범죄로 인한 사회적 비용 감소

하지만 현재 국내 사회문제 해결 연구보고서의 본문은 신속한 정보추출이 어려운 비정형 텍스트 형태를 취한다. 그러므로 연구의 목적, 문제 제기, 그리고 해결책 등 연구 핵심정보에 대한 식별 및 분류를 수작업에 의존하게 하여 사회적 문제와 해결방안을 구조화하는 데에 많은 시간과 비용을 소모하게 한다.

이러한 문제를 개선하기 위해서는 연구보고서와 같은 기존의 연구성과물에서 사회문제, 연구 목적, 해결방안, 성과 등을 자동으로 추출할 수 있는 모델이 필요하지만, 현재 국내의 경우 자동추출 모델을 구성하기 위한 기계가독형으로 구조화된 데이터가 부족하다. 따라서 본 연구는 사회문제 해결 연구보고서에서 연구 핵심 문장을 자동으로 식별하는 모델의 구성을 목표로 하는 문장 의미식별 데이터를 구축하였다.

사회이슈 탐지를 위한 문장 의미식별 데이터는 연구보고서 내 연구개발의 내용 및 성과를 체계적으로 구조화한 데이터로, 연구보고서 본문에서 해당 연구의 "목적(Purpose)", "문제점(Problem)", "해결책(Solution)" 표현 문장을 자동으로 식별하는 데이터이다. 해당 데이터를 구축하기 위하여 기존에 수행된 과학기술 기반 사회문제 해결 R&D 보고서를 수집하고 수집된 연구보고서를 대상으로 본문 영역 내에서 연구의 핵심 내용에 해당하는 연구 목적, 연구 방법, 연구 대상 데이터, 연구 결과와 관련된 문장을 추출하였다. 추출된 문장에는 보고서 내 문장이 의도하는 역할(연구 목적, 연구 방법, 연구 결과)을 구분하는 의미영역 태그를 부착하여 각 문장의 역할을 분류하였다.

그리고, 구축된 데이터의 정합성과 일관성 등을 간접적으로 평가하기 위하여 한국어 대용량 사전학습 언어모델 4개를 활용하여 준거 모델인 문장 의미식별 모델을 구축하였다. 문장 의미식별 모델은 연구보고서 본문 내에서 문장의 역할을 자동으로 구분하는 모델로 실험 결과 KLUE-BERT 사전학습 언어모델을 미세조정한 문장 의미식별 모델이 정확도 73.318, F1 스코어 62.231로 가장 높은 성능을 보였다.

2. 관련 연구

2.1 사회문제 해결 연구

공공연구개발사업으로 추진되는 사회문제 해결 연구의 최종목표는 사회문제 해결을 통한 사회적 가치의 창출이다. 사회문제 해결 연구는 과학기술을 기반한 사회문제 해결을 통해 사회적 가치 창출과 삶의 질 개선을 목표로 한다는 점에서 기존 기술 공급 중심의 국가 연구 과제와 다른 성격을 가지고 있다.

Table 2는 기존 국가연구 과제와 사회문제 해결 연구의 차이점을 나타내는 표이다. 기존 국가연구 과제의 목적과 목

Table 2. Comparison of Existing Research and Social Problem Solving Research[1]

Concepts	General Research	Research on Social Problem Solving
Purpose	National strategy or Economic growth	Improvement of living convenience
	R&D, R&BD → R&SD, R&SBD	
Target	Securing competitiveness in science and technology	Social problem solving
Feature	Expert-oriented R&D	User Participation R&D
Agent	User Participation R&D	Collaboration between R&D department and policy department
Result	These, Patent, Technology	New social services

표는 국가 전략 또는 경제성장을 위한 과학·기술 경쟁력 확보이지만 사회문제 해결 R&D는 사회문제 해결을 통한 시민의 삶의 질 향상에 연구의 주안점을 둔다. 그리고 기존 국가연구과제는 논문·특허 형태의 연구 산출물을 결과로 하지만 사회문제 해결 R&D는 시민사회의 실생활 문제를 해결할 수 있는 구체적인 제품과 서비스를 도출한다[1].

위와 같이 새로운 관점에서 사회문제 해결 연구가 진행되고 있으나, 기존 연구 개발과 차별화되는 특성을 충분히 파악하지 못하는 상황이다. 이에 [2]는 사회문제 해결형 연구개발이 가지는 특성을 탐색하기 위하여 기존에 수행된 연구개발 사업 중에서 성공을 거둔 사례를 바탕으로 사례 연구를 진행하였다.

[2]는 사례 연구의 대상으로 과학기술정보통신부가 추진한 '사회문제 해결형 기술개발사업'에서 최우수를 받은 과제 중 '야간 작업자의 사고 예방을 위한 자가발전 기술 기반 융합형 안전장비 제작 및 실증' 프로젝트를 선정하였다. 위 프로젝트는 '사회문제 해결형 기술개발사업'에서 최우수(S) 등급을 받았으며, 사업 추진과정에서 사회문제 해결형 연구개발의 가이드라인을 구현하고자 하였다.

[2]는 위 사례를 1) 문제해결을 위한 사회/기술 기획, 2) 참여형 기술개발, 3) 사회적 효과 실현을 위한 법 제도/전달체계 구성, 4) 새로운 연구개발 방식의 확장이라는 4개의 측면에서 분석하고 정책 방안을 제시하였다.

[4]는 현재 진행되고 있는 사회문제 해결 연구개발 사업의 중장기적 발전과 성과제고 등을 위하여 추진되어온 기존 사업의 행위 주체에 관한 실증 연구를 현장 인터뷰를 통하여 진행하였다. 이를 통하여 연구개발 정책의 개선 과제를 도출하였으며, 결과를 토대로 개선된 형태의 사업추진절차 및 추진체계를 제시하였다.

이처럼 사회문제 해결 연구개발사업은 아직 초기단계로 기존 국가 연구과제와의 차별점과 특성을 파악하고 이를 통하여 개선방안을 도출하고자 하는 연구가 다양하게 진행되고 있다. 위 연구에서 공통된 제안은 연구의 실질적인 성과를 위해서는 실수요자가 필요한 사회문제를 명확히 파악하는 것이 중요하다. 이에 본 연구에서는 기존에 수행된 연구의 성과물 중 연구보고서에서 신속한 연구성과를 파악할 수 있도록 하는 모델을 학습할 수 있는 기계가독형 데이터를 구축하고자 한다.

2.2 연구 논문 및 보고서 대상 기계가독형 데이터 구축 연구

연구성과물 중 연구논문과 보고서는 보통 정제되지 않는 비정형 텍스트로 제출된다. 이 정제되지 않는 비정형 텍스트는 본문 내 개념과 이론을 파악하고 구조를 갖춘 데이터로 만드는 데에 많은 시간과 비용이 소모된다. 이에 텍스트 구조의 본문에서 연구의 핵심정보를 추출하는 텍스트 마이닝 기술의 필요성이 대두되고 있으며, 이와 관련된 기계가독형 데이터를 구축하는 연구 또한 활발히 진행되고 있다[5-12].

창원대학교와 KISTI(한국과학기술연구원)는 과학기술 분야의 논문을 대상으로 정보의 효율적인 검색과 정보추출을 위하여 논문 본문 문장에 대하여 논문의 의미 구조를 반영하는 수사학적 태그를 자동으로 부착하는 분류 모델을 구축하였다[5]. [5]는 모델의 구축과 실험을 위하여 과학기술 분야의 논문 3,000 건을 수집하였고, 논문내 전체 문장을 수동으로 태깅하였다. 논문내 문장은 논문의 의미구조를 반영하여 연구 목적, 연구 방법, 연구 결과로 구성된 3가지 대분류를 기본으로 각각의 대분류 아래에 9가지 유형의 세부 분류 태그로 태깅을 수행하였다.

[6]은 의료분야 문헌의 초록에서 절차적 지식을 추출하는 방법론에 관한 연구를 수행하였다. 절차적 지식은 연구 목적과 목표를 달성하기 위하여 수행하는 방법이나 기술에 대한 지식이다. [6]은 의료분야 문헌의 초록에서 절차적 지식을 추출하기 위하여 전문의와 함께 의료 문헌의 초록을 분석하여 의료문서에서의 절차적 지식을 모델링하고 텍스트 마이닝 기법을 활용하여 절차적 지식을 추출하였다.

[6]은 실험을 위해 전문의와 함께 위암과 척추질환에 대한 1,309개 문서에 절차적 지식 태깅 작업을 수행하였고, 이 문서 집합을 기반으로 목적/해법 추출, 절차의 개체 추출, 단위절차 구성, 그리고 개체 간 관계 추출 등의 실험을 수행하였고 단계별로 62%에서 82%의 F1스코어 값을 도출하였다.

[7]은 비구조화 텍스트로 이루어진 과학기술 분야 도메인에서 딥러닝 기반 정보추출 연구를 수행하였다. 이를 위하여 [7]은 식품 분야 학술지를 대상으로 개체명 분류 체계와 학습데이터를 구축하였다. 학습데이터 구축에 사용된 원문 대상 데이터는 동아사이언스(875건) 및 국내 식품 도메인 학술지 10종(11,433건)이다. 최종 구축된 과학기술 분야 개체명 인식 데이터는 18,705건(문장)이다.

[8]은 데이터 구축 과정을 효율적으로 수행할 수 있도록 연구 및 문헌 검토 프로세스를 자동화하여 과학 기술 분야 논문이나 학술 기사에서 해당 문헌의 메타데이터를 자동으로 추

출하여 과학 저널 텍스트를 효율적으로 분석하고 태그를 할당할 수 있는 주석 도구를 개발하였다. 개발된 주석 도구는 기존에 수작업으로 문헌을 분석하고 태깅하였을 때보다 사람의 실수를 줄이는 결과를 도출하였으며, Brat과 같은 기존 주석 도구보다 효율성이 증대하고 시간적 비용이 감소하는 결과가 도출되었다.

서술형 텍스트로 구성된 기존 연구성과를 기계가독형으로 변환하는 연구 다수는 과학기술 분야 보고서와 논문을 대상으로 수행되었다. 과학기술 분야 연구논문의 경우 기존 국가 R&D와 마찬가지로 연구 목표가 시민사회의 문제해결보다는 과학기술의 경쟁력 확보를 중점으로 한다.

이에 본 연구는 사회문제 해결 R&D의 성과물 중 하나인 연구보고서를 대상으로 비정형화된 구조로 고립되어 있던 연구 성과를 추출하고 이를 기계가독형 데이터로 변환하고자 한다. 그리고 다른 연구과제에서 데이터를 활용할 수 있도록 추출된 문장의 의도와 역할을 태깅하여 구분하였다.

3. 데이터 구축

사회이슈 탐지를 위한 문장 의미식별 데이터를 구축하는 과정은 Fig. 1과 같다. 우선 사회문제 해결 R&D보고서를 공공 및 연구기관 사이트에서 수집하였다. 수집된 연구보고서에서 연구 핵심 문장을 선별하여 연구 목적, 연구 방법, 연구 결과의 각 대분류 아래 11가지 세부 유형으로 정의된 의미 태그 기준에 따라 해당 역할을 하는 문장을 태깅하였다. 데이터 구축은 2022년 4월부터 10월까지 수행되었으며, 동원된 인력은 석사급 인력 2명과 5명의 연구 보조원(학부 과정)이다.

3.1 사회문제 해결 R&D 연구보고서 수집

본 연구에서는 보고서 수집을 위해 다음과 같은 기준을 마련하였다. 첫 번째 기준은 사회문제 해결의 성격과 특징을 충족하는 연구일 것, 두 번째 기준은 국가 R&D 과제 중 사회문제 해결의 성격과 특징을 가지고 있는 연구일 것으로 하였다.

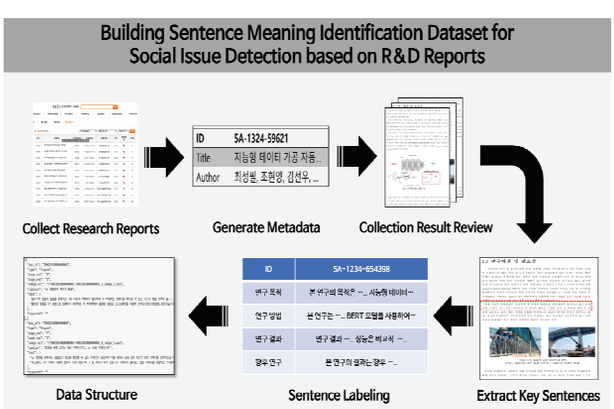


Fig. 1. Sentence Semantic Identification Data Construction Process for Social Issue Detection

Table 3. Top 10 R&D Areas for Solving Social Problems[1]

Categories	Detail
Health	Chronic disease, Intractable disease, Addicted, Degenerative nerve/brain disease, Mental illness/ Physical disability, etc.
Pollution	Household waste, Indoor air pollution, Water pollution, Endocrine disruptor, Industrial waste, Fine dust, Microplastic, etc.
Culture	Cultural alienation, Lack of leisure/cultural space, etc.
Public safety	Sexual crime, Food safety, Online crime, Negligent accident, White-collar crime, Invasion of privacy, etc.
Disaster	Meteorological disaster, Chemical accident, Infectious disease, Radioactive contamination, Quake, Conflagration, etc.
Energy	Electric supply, Energy poverty, etc.
House & Traffic	Decrepit house and Slum, Traffic congestion, Traffic safety, etc.
Family	Elderly suicide, Domestic violence, Low birth rate
Education	Educational gap, School violence
Social integration	Medical gap, Digital divide, Vulnerable social group, Discomfort of Living, etc.

두 번째 기준을 별도로 설정한 이유는 현재 사회문제 해결 연구가 본격적으로 시작된 지 약 7년 정도에 불과하여 연구보고서의 숫자가 풍부하지 않기 때문이다. 이에 많은 연구보고서를 수집하기 위하여 국가 R&D 보고서도 수집할 수 있도록 별도 수집 기준을 마련하였다. 세 번째 기준은 사회문제 해결 방안이 과학기술기반 방법일 것으로 하였다. 사례 연구와 현상 분석 연구의 경우 사회문제에 대한 분석과 방법 제안은 가능하지만, 제안된 방법에 대한 실증적 실험 및 결과가 없는 한계가 존재하기 때문이다.

수집 대상인 보고서는 제 2차 과학기술 기반 국민생활(사회)문제 해결 종합계획(2018~2022)에서 제시한 10대 분야 중 6개 분야를 대상으로 보고서를 수집하였다. Table 3은 사회문제 해결 R&D 주요 10대 분야의 개요이다. 수집의 대상으로 하는 6개 분야는 '건강', '환경', '문화여가', '생활안전', '재난재해', '에너지'를 대상으로 하였다. 나머지 분야인 '주거교통', '가족', '교육', '사회통합'의 경우 사전에 수집한 보고서를 분석한 결과 사회문제 해결을 위한 연구 방법이 대체로 사례연구 및 현황 분석으로 수행되어 본 연구의 수집 기준과 다르므로 제외하였다.

사회문제 해결 R&D 보고서 수집은 보고서 목록을 제공하는 3개 기관을 대상으로 수집하였다. 3개 기관은 Table 4와 같다. 공공기관 경영정보 공개시스템(All Public Information In-One, 이하 ALIO)은 기획재정부에서 운영하는 시스템으로, 공공기관의 경영과 관련된 주요정보를 종합 및 고지

Table 4. Institutions and Systems Subject to Research Report Collection

Institution and System	Description
All Public Information In-One (ALIO)	• Provides key information related to the management of public institutions • Collection of R&D data from public institutions and research institutes
National Science & Technology Information Service (NTIS)	• Information disclosure service for national R&D projects such as projects, tasks, and researchers • Collected from ScienceOn after collecting assignment information
National Balanced-Development Information System (NABIS)	• A system that discloses data such as policies, projects, information, and education conducted by national and regional organizations • Collection of reports published by national research institutes and local provincial research institutes

하여 이용자가 쉽게 문서에 접근하고 파악할 수 있도록 한다. 본 연구는 ALIO에서 제공하는 연구보고서를 Python으로 제작한 웹크롤러를 활용하여 연구보고서와 메타데이터를 수집하였다. 수집된 연구보고서 중 위에서 설정된 기준에 부합하지 않는 연구보고서는 제외하였다.

두 번째 수집 대상 시스템은 국가과학기술지식정보서비스(National Science & Technology Information Service, 이하 NTIS)이다. NTIS는 한국과학기술정보연구원에서 운영하는 서비스로 국가 및 공공기관의 사업, 과제, 연구자 등 국가연구개발 사업에 대한 정보를 한 곳에서 서비스하는 국가R&D지식 포털이다. NTIS의 경우 연구보고서 원문을 제공하지 않아, 과제 정보를 확인한 후 연구보고서를 한국과학기술정보연구원에서 운영하는 ScienceON에서 연구보고서와 관련 메타데이터를 수작업으로 수집하였다.

마지막 수집 대상 시스템은 국가균형발전종합정보시스템(National Balanced-Development Information System, 이하 NABIS)이다. NABIS는 국가균형발전위원회와 한국산업기술평가관리원(균형발전평가센터)에서 운영하는 시스템으로 국가 및 지역 기관에서 수행하는 정책, 사업, 정보, 교육 등의 자료를 공시하는 시스템이다. 본 연구는 NABIS에서 공시한 국책연구원 및 각 지역 시도 연구원에서 발간된 연구보고서를 6개 연구 분야에 맞춰 수작업으로 연구보고서와 연구보고서의 메타데이터를 수집하였다.

수집된 연구보고서는 Table 5와 같이 700건이다. 제일 많이 수집된 분야는 '건강'으로 총 239건이며, 제일 적게 수집된 분야는 '문화여가' 분야이다. 수집대상기관 중 'ALIO'에서 가장 많은 보고서가 수집되었다.

3.2 연구 핵심 문장 선정 및 태깅

사회이슈 탐지를 위한 문장 의미식별 데이터를 구축하기 위해 Fig. 2와 같이 수집된 사회문제 해결 R&D 보고서에서 주요 서술 영역을 분리하고, 영역 내에서 문장을 선별하여 태깅하였다. 먼저 주요 서술 영역을 분리하기 위하여 연구보고서 내 모든 텍스트 중 본문에 해당하는 영역의 텍스트를 Python 외부 라이브러리인 PDFminer.six[13]을 활용하여 추출하였다.

Table 5. Research Report Collection Status

Categories	ALIO	NTIS	NABIS	Total
Health	225	12	2	239
Pollution	90	68	33	191
Culture	12	1	16	29
Public safety	28	23	29	80
Disaster	71	13	12	96
Energy	50	2	13	65
Total	476	119	105	700

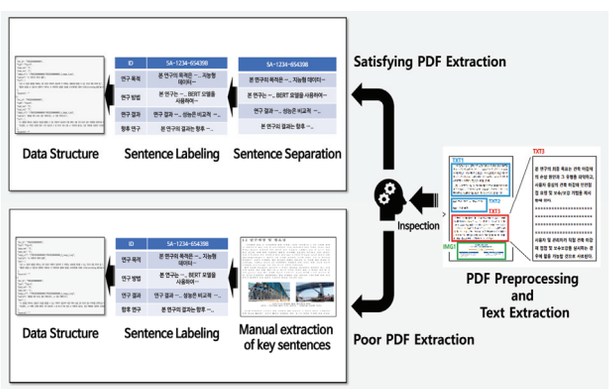


Fig. 2. Selection and Tagging Process of Research Key Sentences

PDF 파일에서 텍스트 영역이 제대로 추출되었는지 확인하기 위하여 2명의 석사급 연구원과 5명의 보조 연구원이 검수하였다. 검수 결과 제대로 추출되지 않은 연구보고서 PDF 파일은 별도 보관하여 연구 참여자들이 PDF 파일에서 직접 연구보고서 내 본문에서 연구 핵심 문장을 추출하여 선정하였다.

검수 결과 제대로 추출된 연구보고서는 연구내용에 해당하지 않은 텍스트 영역을 수작업으로 제외하고 연구내용이 담긴 텍스트 영역을 문장 분리를 수행하였다. 문장 분리는 Python 라이브러리인 KSS(Korean Sentence Splitter)[14]를 사용하였다. 문장 분리기로 분리된 연구문장은 연구핵심(연구 목적, 연구 방법, 연구 결과 등)의 내용을 담은 문장과 기타 문장으로 구분하였다.

Table 6. Types and Classification of Sentence Tagging

Main Categories	Subcategories	Sentence Meaning Tagging Criteria
Necessity and purpose of research	Need for Research	• Sentences suggesting areas to be improved in the research topic • Reason statements for conducting research
	Purpose of Research	• Sentences about the purpose and direction of the presented problems and topics
	Research Hypothesis	• A hypothesis about the relationship between the variables raised in the research question
	Contribution of Research	• Contribution to problem-solving through research
Research Method	Proposals for Research	• A method or theory to solve the research goal • Proposal of a characteristic method different from previous studies • Overview of study design or study methods
	Definitions	• Definition of the method or theory presented in the study • Sentences that define technical terms
	Subject of Experiment	• Definition of the data or experimental group that is the main subject of the study • A description of the collection method or source of the test subject\
	Method of Analysis	• Methods for measuring and analyzing target data
Resarch Result	Performance and Effect	• Sentences that quantify the results of the experiment • Analysis and interpretation of experimental results
	Limit of Research	• Sentences that presents the limitations of the study
	Follow-up Research	• Proposal for research expansion and utilization

선정된 연구핵심 문장은 의미별로 Table 6의 기준과 같이 분류하여 태그작업을 수행하였다. 문장 의미 태그는 연구보고서의 본문에 대한 의미 구조를 반영하여 '연구 필요성 및 목적', '연구 방법', '연구 결과'를 대분류로 정의하였다. 세부 분류는 각 대분류 아래 Table 6과 같이 11가지 세부유형으로 정의하였다. 세부 유형에는 '연구 필요성', '연구 목적', '연구 가설', '연구 기여', '제안 방법', '기술 정의', '실험 대상', '분석 방법', '성능 효과', '연구 한계', '후속 연구'이다.

데이터 구축 과정은 구축-검토-검수로 이루어진 프로세스를 거쳤다. 먼저 구축 단계는 앞서 말한 바와 같이 연구 핵심 문장을 추출 및 선별한 후 각 문장 역할에 해당하는 라벨 태깅을 수행하는 과정이다. 작업자가 구축한 데이터는 검토 단계를 거친다. 검토 단계는 태깅 기준에 의해 구축된 데이터를 수정 및 보완하는 단계이다. 검토까지 완료된 데이터는 검수 단계를 거친다. 검수 단계는 태깅 기준에 맞추어 검토가 완료된 데이터를 완료/재검토/폐기 결정을 수행하는 단계이다. 이처럼 각각 단계별로 작업자 상호 간 피드백을 통하여 데이터 품질을 향상하는 과정을 수행하였다.

Table 7. Data construction status

Main Categories	Subcategories	Count of papers
Necessity and Purpose of Research	Need for Research	2,288
	Purpose of Research	1,979
	Research Hypothesis	277
	Contribution of Research	1,653
Research Method	Proposals for Research	1,127
	Definitions	552
	Subject of Experiment	1,719
	Method of Analysis	3,509
Resarch Result	Performance and Effect	9,678
	Limit of Research	380
	Follow-up Research	860
Total		24,022

4. 데이터 특징 및 분석

기계 학습은 결과의 예측을 위하여 데이터 내 패턴을 파악, 이용한다. 이에 본 장에서는 구축된 데이터가 일관적인 패턴을 가져 모델 학습에 적합한지 판단하기 위하여 데이터의 특징을 분석하고 그 결과를 도출하였다.

최종 구축된 사회 이슈 탐지를 위한 문장 의미 식별 데이터는 Table 7과 같이 총 24,022개이다. 구축된 데이터 중에서 '연구 결과'를 의미하는 문장이 10,918건으로 가장 많았

Table 8. Data Example 1 (Necessity and Purpose of Research)

Subcategories	Examples
Need for Research	• "한편 만성 정신질환자에 대한 치료의 궁극적인 목표가 그들의 삶의 질 향상에 있고, 이를 위해 정신사회적 손상을 보완하여야 한다는 필요성이 제기되고 있다." • "공동 연구과제로 진행 중인 안정화 처리의 대안 공법인 적극적(열탈착, 토양세척)을 이용한 처리 기술 개발 추진이 필요하다."
Purpose of Research	• "따라서 본 연구진은 치료 전 후의 환자 시료 내의 대사체와 단백체 분석을 통해 PD-1 차단 면역 항암제의 반응 예측 평가 방법을 구축하고자 함." • "본 연구에서는 블록체인에 대한 개념, 특성, 기술현황, 국내외 동향, 의료분야 적용사례 등을 종합적으로 살펴보고 의료 분야에 블록체인의 기술이 도입되기 위한 블록체인의 특성을 살펴보고자 한다."
Research Hypothesis	• "둘째, 탐색적 연구의 특성을 고려하여 완충요인들이 없이 작업 부하요인, 안전조건요인, 조직적 요인, 개인적 요인 모두가 심리적 부담에 영향을 미치고 그 결과 휴먼에러나 안전사고로 이어지는 경로를 두 번째 가설로 설정을 하였다. (그림 9)" • "가설: 진행성 위암 중 일부 초기 그룹, 즉 stage IB/IIA/IIB 위암에서 D1 + 림프절 절제가 종양학적으로 충분하며, 환자의 삶의 질 향상, 합병증 감소, 의료비의 감소를 자져올 수 있음"
Contribution of Research	• "그러므로 산업적 제품적 특징이 반영된 농산물 온라인 직거래의 물류서비스품질에 관한 연구가 이루어진다면 현장에 있는 중소교무 농민들에게 의미 있는 시사점을 줄 수 있을 것이다." • "이를 통하여 암환자의 생존율 개선을 위한 새로운 치료기전 발굴 및 치료법 개발 연구임. 두경부 암환자의 사망률을 낮추고 생존율을 향상시켜 국민건강에 기여하는 연구 중심적 암치료 개발의 사례가 될 것임."

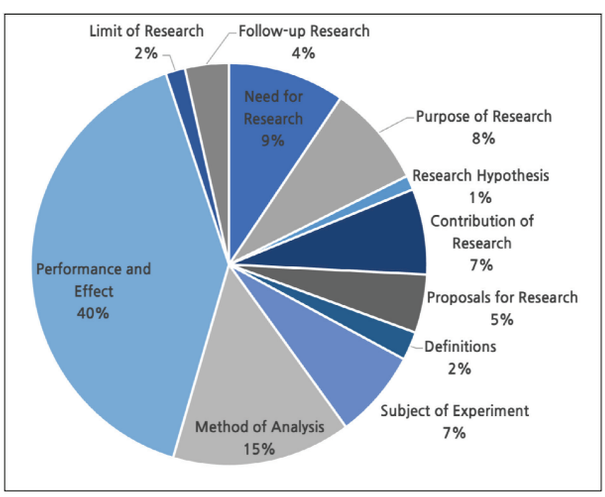


Fig. 3. Ratio of Sentence Semantic Identification Data Type for Social Issue Detection

으며, 다음으로 '연구 방법' 6,907건, '연구 필요성 및 목적' 6,197건 순으로 구축되었다. 세부 유형 기준은 Table 7과 Fig. 3과 같이 '성능 효과'가 9,678(40%)로 가장 많았으며, 다음으로 '분석 방법'이 3,509건(15%), '연구 필요성'이 2,288건 (9%) 순으로 많았다. 가장 적게 구축된 세부 유형은 '연구 가설'로 277건(1%)이 분류되었다.

Table 8은 사회 이슈 탐지를 위한 문장 의미식별 데이터 대분류 유형 중 "연구 필요성 및 목적"에 해당하는 문장의 예시이다. 먼저 "연구 필요성"에 해당하는 문장의 경우 연구 배경과 달리 연구과제가 해결하고자 하는 구체적인 문제 상황과 필요성을 설명하고 있다. 다음으로, "연구 목적" 유형에 해당하는 문장들은 연구 필요성에서 제기된 문제의 해결을 위하여 연구가 하고자 하는 목적과 목표를 명확하게 제시하고 있다. "연구 가설" 유형에 해당하는 문장은 연구과제에서

Table 9. Characteristics of Phrases by Sub Category 1 (Necessity and Purpose of Research)

Sub-categories	Examples
Need for Research	• "A와 같은 문제점이 존재하므로 B와 같은 대안이 필요하다." • "C의 필요성이 제기되고 있다."
Purpose of Research	• "본 연구는 A를 이루는 것을 목표로 한다." • "본 연구는 B라는 결과를 내려고 한다."
Research Hypothesis	• "본 연구는 A와 같은 결과를 낼 것이다." • "B와 C등의 것들을 종합할 때 D라는 결과를 낼 수 있다."
Contribution of Research	• "A와 같은 성과가 있을 것이다." • "본 연구의 결과는 다음과 같은 부분에 기여할 수 있을 것이다."

확인하고자 하는 잠정적 결론 또는 추측을 제시하는 역할을 수행한다. 논문과 달리 연구보고서의 경우 Table 8의 예시와 같이 가설을 명확하게 제시하는 특징을 보인다. 마지막으로 "연구 기여"는 연구를 통해 도출된 결과가 연구 주제에 기여하는 내용을 설명하는 역할을 한다. 이 유형에 해당하는 문장은 기대 결과에 대해서 예측하는 형태로 문장이 작성된다. 이 분류에서 자주 나타나는 문구는 Table 9와 같다.

Table 10은 대분류 유형 중 "연구 방법"에 해당하는 문장의 예시이다. 먼저 세부 분류 중 "제안 방법"에 해당하는 문장은 연구에서 설정한 주요 문제나 목표를 해결하기 위한 방법이나 이론을 제시하는 역할을 한다. 다음으로 "기술 정의" 유형에 분류된 문장들은 "제안 방법"이나 "분석 방법"에서 사용된 기술이나 방법에 대한 개념을 정의하는 역할을 한다. 마지막으로 "실험 대상"과 "분석 방법" 유형에 해당하는 문장은 각각 연구과제가 수행하고자 하는 실험의 대상과 실험의 구

Table 10. Data Example 2 (Method of Research)

Subcategories	Examples
Proposals for Research	• “가뭄상황 판단 및 예측을 위해 관측자료 기반 복류수 취수 지점의 가뭄판단방법론을 제시하기 위해 Markov Chain Model과 Hidden Markov Chain Model(HMM) 기반의 시계열자료 모의기법 기반 Gaussian Mixture Non-homogeneous Hidden Markov Chain Model(GM-NHMM)을 제시하였다.” • “이 절에서는 LSTM 신경망을 이용하여 센서가 감지한 데이터의 오결측을 정제하는 시뮬레이터를 제시한다”
Definitions	• “XRD 분석은 물질의 결정 구조를 분석하는데 가장 널리 사용되는 방법으로 X 선이 Bragg법칙을 만족할 때 얻어지는 X-선 회절패턴을 측정하여 물질의 결정구조, Impurity phase의 존재여부 등을 분석한다.” • “SWOT 분석은 조직 내부의 강점(Strength), 약점(Weakness), 외부환경의 기회(Opportunities), 위협(Threats)을 파악하여 조직의 강점과 매력적인 환경의 기회를 살리는 동시에 조직의 약점을 극복하거나 없애고, 위협을 최소화하는 전략을 수립하는 방법이다.”
Subject of Experiment	• “현재 진행 중인 갑상선암 환자 500명과 정상 대조군 1000명에 대한 전장유전체연관 분석을 통해 candidate gene을 발굴” • “연구대상은 의료종사자 중 의사, 간호사, 전문직 의료 종사자, 병원전산정보담당자로 구분하였다.”
Method of Analysis	• “UPLC-TQ-MS/MS 분석에서 얻은 스펙트럼을 통해 다변량 분석을 수행하여 정상과 구강암 환자의 상이한 대사체 패턴을 보이는 것을 확인함” • “혈액을 15분간 실온에서 3,000 rpm으로 원심분리한 뒤 혈장(혈액의 액체성분, plasma)을 채취한 후 혈장 속에 포함된 항체(Immunoglobulin) 수준을 sandwich ELISA(Enzyme-Linked Immunosorbent Assay) 방법으로 정량 분석하였다.”

Table 11. Characteristics of Phrases by Sub Category 2 (Method of Research)

Subcategories	Examples
Proposals for Research	• “본 보고서는 A에 대한 연구를 달성하기 위하여 B와 같은 방식으로 실험을 진행하였다.” • “본 보고서는 다음과 같은 순서로 연구를 진행한다.”
Definitions	• “본 연구는 소기의 목표를 달성하기 위하여 A를 투입하였는데, A의 기술적 특징은 다음과 같다.” • “B분석은 C, D, 그리고 E를 차례로 수행하여 F라는 계산 결과를 도출하는 공식을 근간으로 한다.”
Subject of Experiment	• “A치료법의 효능을 검증하기 위해서 모집한 지원자의 신체적 특징은 다음과 같다.” • “B를 검증하기 위하여 실험군 C를 수집하였는데, 수집 조건은 다음과 같다.”
Method of Analysis	• “실험은 다음과 같이 진행되었다.” • “실험의 진행을 위하여 A방법론을 채택하여 B와 같은 결과를 도출하였다.”

Table 12. Data Example 3 (Result)

Subcategories	Examples
Performance and Effect	• “TRIX 분석을 통해 유입수의 수질을 평가한 결과 52개 정점 중 지수의 범위가 0~4(수질 high / 영양 단계 low)에 속하는 지역은 존재하지 않았으며, 8개 정점에서 4~5(수질 good / 영양단계 medium)의 범위를 나타내고 있다.” • “탄소나노튜브의 겉보기 밀도는 제품마다 달랐다. 본 연구에서 측정한 겉보기 밀도와 제조자가 제시하는 겉보기 밀도는 조금 차이가 나기는 하나 대체적으로 유사하였다.”
Limit of Research	• “현재까지 확실한 항암면역치료의 반응성에 대한 바이오 마커가 부재하고 본 연구에 참여하는 환자수가 제한적이라는 점은 본 연구의 가장 큰 단점일 수 있음.” • “본 연구의 제한점으로 MPPD 모델은 그 기본 원리가 수학적 공식을 통한 추정치의 제공이므로 실제 침착 분율과 차이가 있을 수 있다.”
Follow-up Research	• “현재 갖추어진 “국립암센터 GEM facility”의 시스템과 인력을 유지하기 위해서는 지속적인 기관의 지원이 필요한 상황이며, 후속 연구과제를 수행함으로서 현 시스템을 유지 발전시키고자 함.“ • “자연정화시설 등의 침전 슬러지 처리사업을 수행함에 있어 사업계획 수립 및 비용산출, 장치 현장 설치 운영, 폐기물처리 등 업무 수행 전 단계에 대한 업무절차를 표준화하고 운영지침을 제시하였으며, 체계적 사업 도입(‘19년 3건, ’20년 4건)을 통해 수질정화시설의 슬러지 처리에 소요되는 유지관리 비용절감에 기여할 것으로 기대된다.”

체적인 방법을 제시하는 역할을 한다. 이 두 유형에 해당하는 문장들은 대상과 방법을 명확하게 제시하고 있는 경우가 대다수이다. 이들 분류에서 자주 나타나는 문구는 Table 11과 같다.

Table 12는 사회이슈 탐지를 위한 문장 의미식별 데이터 대분류 유형 중 “연구 결과”에 해당하는 문장의 예시이다. 먼저 “실험 결과”로 분류된 문장들은 사회문제해결 연구 과제에서 수행된 실험 결과를 설명하는 역할이다.

Table 13. Characteristics of Phrases by Sub Category 3 (Result)

Subcategories	Examples
Performance and Effect	• “이번 실험의 시사점은 다음과 같다.” • “이번 실험을 통하여 A와같은 성과를 거둘 수 있었다.”
Limit of Research	• “하지만 A라는 결과는 이번 실험의 명확한 한계점으로 작용한다.” • “B라는 경우는 연구의 구조상 진행이 어려웠다.”
Follow-up Research	• “이번 연구의 결과는 다음과 같이 상용화될 수 있을 것으로 사료된다.” • “본 연구의 미흡점은 다음과 같은 후속연구에서 보충될 것으로 보인다.”

이에 이 분류에 속한 문장들은 실험 결과에서 도출된 수치적 결과를 나타내거나 실험 결과에 대한 연구자의 분석과 견해를 보여준다. 그래서 이 유형으로 태깅된 문장 대부분은 공통적인 문구를 나타내기보다는 실험 결과의 수치에 대한 표현과 연구를 수행한 연구자들이 실험 결과에 대해서 분석한 내용들이 많이 나타난다. 다음으로 “연구 한계”유형에 해당하는 문장들은 연구과제에서 수행한 방식의 한계점을 설명하고 있다. 이 유형에 속한 대다수의 문장들은 대체적으로 연구보고서 결론에서 선정되었다. 마지막으로 “후속 연구”에 해당하는 유형의 문장은 향후 연구가 진행되어야 하는 방향이나 확장에 관한 내용을 담고 있다. 이 유형에 속한 문장들은 연구 기여와 유사한 특징을 보이나 연구적 한계를 보완하는 방법에 대해서 언급한다는 점에서 차이점을 가지고 있다. 이 분류에서 자주 나타나는 문구는 Table 13과 같다.

5. 실험 및 결과

본 연구는 구축된 데이터에 대한 정합성, 일관성 등을 간접적으로 평가하기 위하여 한국어 대용량 사전학습 언어모델 4개를 활용하여 연구보고서 문장 수사학적 분류 모델을 구축하고 성능에 대한 비교 실험을 진행하였다.

5.1 실험데이터

본 연구는 모델의 성능을 좀 더 객관적으로 평가하기 위하여 연구보고서 내 연구 핵심 내용을 표현하지 않은 문장을 ‘기타’유형으로 1,152개 추가하였다. 추가된 데이터셋은 모델 학습을 위해 9:1 비율로 학습데이터와 실험 데이터로 분할하였다. 데이터에 대한 통계는 Table 14와 같다.

5.2 실험 환경

본 연구에서 구축한 분류 모델에 사용된 한국어 대용량 사전학습 언어모델은 KLUE-BERT[15], KcBERT[16], KcELECTRA[17], KoELECTRA[18]이며, 각각 사전학습 언어모델의 Vocab, Layer, Feed Forward Network(FFN), Dimension, Hidden dimension(HD)의 크기는 Table 15와 같다. 분류 모델의 학습은 총 10번의 Epoch에서 진행되었으며, 2번의 Epoch에서 성능 개선이 없는 경우 학습을 종료하였다. Optimizer는 AdamW를 사용하였고 Learning rate는 5e-5로 설정하였다. 그리고, 과적합을 막기 위하여 Dropout은 0.1로 설정하였다. 모든 실험은 Intel® Core™ i9-9990X CPU 3.5GHz와 NVIDIA TITAN RTX 4개가 장착된 PC에서 이루어졌다.

Table 14. Model Training Data Statistics

Main categories	Sub categories	Train	Test	Total
Necessity and Purpose of Research	Need for Research	2,039	249	2.288
	Purpose of Research	1,791	188	1.979
	Research Hypothesis	252	25	277
	Contribution of Research	1,512	141	1.653
Research Method	Proposals for Research	1,014	113	1.127
	Definitions	500	52	552
	Subject of Experiment	1,560	159	1.719
	Method of Analysis	3,157	352	3.509
Research Result	Performance and Effect	8,677	1,001	9.678
	Limit of Research	347	33	380
	Follow-up Research	772	88	860
Etc		1,035	117	1.152
Total		22,656	2,518	25,174

Table 15. Parameters of Pre-trained Language Model

Model	Vocab	Layer	FFN	HD
KLUE-BERT	32,000	12	3072	768
KcBERT	30,000	12	3072	768
KcELECTRA	50,135	12	3072	768
KoELECTRA	35,000	12	3072	768

5.3 실험 결과

본 연구에서 구축한 데이터로 학습한 분류 모델은 정확도, F1 스코어를 이용하여 평가하였다. 정확도(Accuracy)는 실제 데이터와 모델이 예측한 데이터가 얼마나 같은지를 판단하는 지표이다. F1 스코어는 정밀도(Precision)와 재현율(Recall)에 동일한 가중치를 주어 계산한 조화평균으로, 데이

Table 16. Performances of Pre-trained Language Model

Model	Accuracy	F1 (Macro)	Recall (Macro)	Precision (Macro)
KLUE-BERT	75.853	63.503	63.187	64.751
KcBERT	73.551	60.826	58.405	67.616
KcELECTRA	74.265	59.564	59.341	60.404
KoELECTRA	75.059	60.698	61.544	60.416

터의 클래스가 불균형할 경우 사용되는 측정 지표이다. 정밀도는 모델이 Positive로 예측한 결과 중 실제로 Positive인 비율로 예측 결과가 얼마나 정확한지 판단하는 지표이다. 재현율은 실제 정답 중 올바르게 예측한 정답의 비율을 판단하는 지표이다. F1 스코어, 정밀도, 재현율은 클래스 별 수치에 별도의 가중치 없이 평균을 내는 방식인 Macro 평균을 활용해 평균을 산출하였다. 정확도, F1 스코어, 정밀도, 재현율의 수식은 아래와 같다.

$$\begin{aligned} Precision &= \frac{True\ \ Positives}{True\ \ Positives + False\ \ Positives} \\ Recall &= \frac{True\ \ Positives}{True\ \ Positives + False\ \ Nagatives} \\ F1 &= 2 \times \frac{Precision \times Recall}{Precision + Recall} \\ ACC &= \frac{TP + TN}{TP + FN + FP + TN} \end{aligned} \quad (1)$$

실험 데이터를 통해 측정한 사전학습 언어모델별 성능은 Table 16과 같다. KLUE-BERT는 정확도 75.853%, F1 스코어 63.187%로 4개 사전학습 언어모델 중 가장 높은 성능을 보여주었다. 정확도의 경우 75.059%로 KoELECTRA 사전학습 언어모델이 두 번째로 높은 성능을 보인 반면, F1 스코어의 경우 KcBERT 모델이 60.826%로 두 번째 높은 성능을 보였다. 이는 KcBERT 모델이 정밀도가 67.616%로 KoELECTRA 모델보다 약 7% 높기 때문이다. 4개 사전학습 언어모델의 정확도 평균은 74.682이고 F1 스코어의 평균은 61.147이다.

실험 집합에 대한 성능을 자세히 분석하기 위하여 실험 데이터에서 성능이 좋았던 KLUE-BERT 사전학습 언어모델을 미세조정한 분류 모델의 혼동 행렬(Confusion Matrix)를 도출하였다. 혼동 행렬의 결과는 Table 17과 같다.

Table 17을 보면 전반적인 예측과 정답 값의 분포가 '성능 효과'에 집중되어 있다. 심지어 모델은 다른 태그가 정답인데도 불구하고 '성능 효과'로 예측된 결과가 더욱 많음을 확인할 수 있다. 이러한 현상은 전체적인 데이터가 '성능 효과'에 편중되어 있기 때문으로 보인다. 한편, Table 17에서 추가로 확인할 수 있는 사항은 '연구 가설'에 대한 예측값이 전혀 존재하지 않는 점이다. 이 현상 또한 학습데이터 중 '연구 가설'에 매핑된 데이터가 매우 적기 때문이다. 그럼에도 '기타'에 해당하는 노이즈 데이터는 비교적 높은 성능으로 분류할 수 있음을 확인하였다.

그리고 본 연구는 모델에 안정적인 학습이 이루어지고 있는지 확인하기 위하여 학습 Epoch 횟수에 따른 성능 변화도

Table 17. Confusion Matrix of Sentence Rhetorical Classification Model(KLUE-BERT)

		Predict											
		Etc	Need for Research	Purpose of Research	Research Hypothesis	Contribution of Research	Proposals for Research	Definitions	Subject of Experiment	Method of Analysis	Performance and Effect	Limit of Research	Follow-up Research
Label	Etc	113	1	0	0	0	0	0	1	2	0	0	0
	Need for Research	0	203	3	0	11	1	0	0	0	26	2	3
	Purpose of Research	0	4	144	0	15	11	1	1	3	9	0	0
	Research Hypothesis	0	2	1	0	13	1	0	1	1	6	0	0
	Contribution of Research	0	5	11	0	107	4	0	1	2	11	0	0
	Proposals for Research	0	5	21	0	2	43	1	5	22	13	0	1
	Definitions	0	5	0	0	0	1	38	1	2	5	0	0
	Subject of Experiment	1	0	2	0	1	5	7	89	40	14	0	0
	Method of Analysis	0	0	6	0	0	18	2	16	269	41	0	0
	Performance and Effect	1	45	13	0	43	11	1	2	12	859	8	6
	Limit of Research	0	6	0	0	0	0	0	0	0	6	20	1
	Follow-up Research	0	10	2	0	33	5	0	1	0	15	1	21

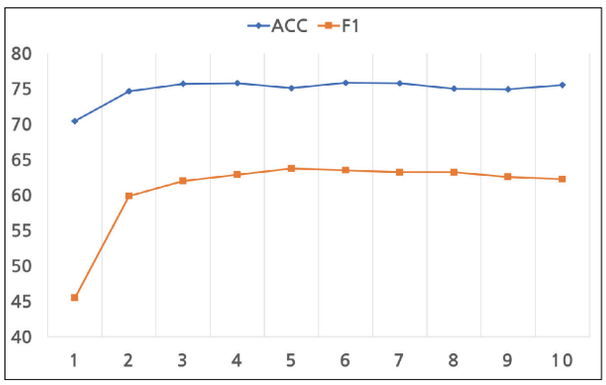


Fig. 4. Performance Change According to the Number of Epochs Fine-tuning the KLUE-BERT Pre-trained Language Model

확인하였다. Fig. 4는 KLUE-BERT 언어 모델을 미세조정 Epoch 횟수에 따른 성능 변화를 그린 그래프이다. 전반적으로 학습 Epoch 6까지는 성능이 상승하나 이후 과적합으로 인하여 성능이 점차 하락하는 것을 확인할 수 있다. 정확도의 경우 학습 Epoch: 6 이후 성능이 떨어지기 시작하여 마지막 Epoch에서 성능이 상승하였다. 한편 F1 스코어의 경우 Epoch1에서 2로 진행되면서 급격한 성능 향상이 있었지만, 학습 Epoch 6 이후로 성능이 점차 떨어짐을 확인하였다. 이를 통해 모델의 학습이 안정적으로 이루어지고 있음을 확인하였다.

본 연구는 추가로 [5]에서 구축한 논문 문장 의미식별 데이터를 통하여 학습한 모델의 실험 결과와 비교분석을 수행하였다. [5]연구와 본 연구의 실험 결과를 직접 비교하기 어려우나, 데이터를 구성하기 위해 사용한 논문과 본 연구에서 사용한 보고서가 유사한 의미구조인 것으로 판단하여 비교 대상 연구로 선정하였다.

[5]에서 구축한 데이터는 Table 18과 같다. [5]는 논문 문장 수사학적 분류 모델을 구축하기 위하여 데이터를 총 289,432건을 구축하였으며, 이전 문장의 정보를 현재 문장에 반영하는 다섯 가지 방법을 제안하였다. 제안된 방법 중 가장 높은 성능을 보인 방법은 Encoder Concat Model이다. Encoder Concat Model은 이전 문장의 임베딩 벡터와 본 문장의 임베딩 벡터를 Concatenation하여 분류를 수행하는 모델이다. 이 모델의 성능 결과는 F1 스코어 71.91, 정확도 88.77 이다.

본 연구와 [5]에서 구축한 데이터의 원문과 의미 태깅에 구조적 차이는 존재하나, [5]에서 구축한 데이터가 약 20만건 이상임에도 불구하고 본 연구에서 실험한 결과의 정확도와 F1 스코어에 큰 차이가 없다. 특히 [5]의 경우 대다수 데이터가 '기타'로 분류되어 있어서 데이터 간 불균형이 심하게 나타난다. 본 연구는 좀더 정확한 실험을 위하여 '기타'의 데이터를 다른 데이터와 균형이 맞게 구축하여 실험을 수행하였다.

Table 18. Dissertation Sentence Rhetorical Classification Data Statistics[5]

	train	validation	test
Hypothesis	1,413	189	184
Definition of technology	2,093	269	229
Etc	178,483	22,232	22,204
Target data	5,491	541	519
Data processing	6,236	658	619
Problem definition	3,187	286	319
Performance / Effect	17,026	1,691	1,770
Theory / Model	2,861	298	258
Proposal	8,379	862	871
Follow-up research	8,379	916	969
Total	233,548	27,942	27,942

Table 19. Error Result Sample

정답	예측	문장
Proposals for Research	Method of Analysis	• "이에 본 연구에서는 낙동강 권역 주요 하천에서의 담수생물다양성 및 서식환경을 분석하고, 주요 담수생물종의 생존 범위에 대한 분석을 수행하였으며, 국내 미적용 평가법의 국내 적용 적합성을 검토하고 보완된 방법을 적용하여 시범 분석 결과를 도출하였다."
Contribution of Research	Follow-up Research	• "또한, 기존에 omics 연구에서 도출되었으나 그 기능이 전혀 알려지지 않아 연구의 우선순위에서 밀려난 다른 hypothetical gene들의 탐색을 위한 모델을 제공할 것으로 기대된다."
Follow-up Research	Contribution of Research	• "소동물 방사선조사 모델을 통하여 방사선 유도 임상부작용 연구를 위한 표준 동물 모델 개발, 방사선 유도 부작용의 예방 및 치료방법 개발, 신치료 기법 및 검증 모델 설입이 가능함."

그리고, 본 연구에서 구축한 사회이슈 탐지를 위한 문장 의미식별 데이터의 정합성과 일관성을 평가하기 위한 실험결과 정확도 평균이 70%대에 그쳐 [5]에 비해 낮았다. 이는 학습 데이터의 절대적인 양 때문에 나타나는 한계점이기도 하나, 연구보고서가 가지고 있는 구조적 한계점도 일부 영향을 미친 것으로 사료된다.

학술논문의 경우 대다수 논문이 주어와 술어가 완전한 서술형 문장의 형태를 가지고 있다. 하지만 연구 보고서의 경우 주변의 서술적 맥락을 고려하지 않은 경우가 다수 존재하며, 단순히 명사 어구가 나열된 경우도 존재하였다. 일부 보고서의 경우 연구 결과에 수치만 나열하고 분석된 내용이 없는 경우도 존재하였다. 또한, 학술논문과 연구보고서에서 나타나는 질적 차이도 존재한다. 학술논문은 각 분야의 전문가가 두 명 이상 참가하는 심사 과정에서 연구내용과 구조를 평가함으로써 일정 이상의 질적 수준을 유지한다. 하지만 별도의 심사 과정을 거치지 않은 연구보고서의 경우 연구 과정과 연구 성과에 대해 명확한 설명이 없는 경우도 존재하였다. 이러한 한계로 인해 검토 및 검수를 거침에도 불구하고 데이터상에 오류가 존재하기도 한다.

Table 19는 성능이 제일 좋은 KLUE-BERT 사전학습 언어모델로 미세조정한 모델이 잘못 예측한 결과이다. 이는 세부 분류 중 문장의 표현법과 구조가 비슷한 경우에 가장 많이 발생하였다. 첫 번째 예시의 경우 “제안 방법”은 연구의 기본적인 틀에 대한 설명, “분석 방법”은 전체 연구를 구성하는 한 실험에 대한 설명으로 서술방식과 보고서 내 위상이 구체적/포괄적이라는 부분에서 다르나, 연구 혹은 실험의 진행을 설명한다는 점에서 문장의 구조가 유사하여 오류가 발생한 것으로 보인다. 두 번째 예시에서 “연구 기여”는 연구 결과에 대한 예측을 기본적인 기준으로 한다. 예측이 가설에 불과한 경우도 있지만, 결론에서 예측이 올바른 것으로 드러나는 경우도 존재한다. 이는 “후속 연구”가 의미하는 연구 결과에 대한 활용 부분에서 비슷한 맥락을 이루어 모델의 혼동을 불러일으킨 것으로 사료된다. 해당 부분은 실험 데이터의 한계점이다. 이를 보완하기 위해서는 향후 연구에서 해당 세부 분류에 속하는 데이터의 수를 지속하여 늘려야 할 것이다. 해당 분류에 속하는 데이터의 크기를 늘리면 각 분류별 문장 패턴이 더 나은 구조를 취할 수 있을 것으로 보인다.

6. 결론 및 향후 연구

현재 과학기술기반 사회문제 해결 연구과제는 실질적인 성과 창출이 미약한 것으로 평가되고 있다. 연구 개발의 실질적인 성과 창출이 시민들에게 전달되기 위해서는 현존하는 사회문제와 사회 이슈를 연구자들이 명확히 파악하여야 한다. 이를 위해서는 기존 연구보고서에서 사회문제와 해결방안을 자동으로 식별할 필요가 있다. 이에 본 연구는 연구보고서에서 사회문제와 해결방안을 자동으로 식별하는 모델을 학습하기 위한 데이터셋인 사회이슈 탐지를 위한 문장 의미식별 데이터를 구축하였다.

사회이슈 탐지를 위한 문장 의미식별 데이터를 구축하기 위하여 기존에 수행된 과학기술 기반 사회문제 해결 R&D 연구보고서를 3곳의 기관 및 시스템에서 총 700건을 수집하였다. 수집된 연구보고서를 대상으로 본문 영역 내에서 연구의 핵심 내용인 연구 목적, 연구 방법, 연구 대상 데이터, 연구 결과 등에 해당하는 문장을 추출 및 선정하였다. 그리고 추출된 문장에는 연구보고서 내 문장이 의도하는 역할을 구분하는 의미 태그를 부착하여 각 문장의 역할을 분류하였다. 위와 같은 방법으로 최종 구축된 데이터의 규모는 24,022 건이다.

구축된 사회이슈 탐지를 위한 문장 의미식별 데이터에 대한 정합성, 일관성 등을 간접적으로 평가하기 위하여 한국어 사전학습 언어모델을 활용하여 문장 의미 식별 모델을 구축하였다. 문장 의미 식별 모델은 연구보고서 본문 내 문장이 의도하는 역할을 자동으로 구분하는 모델로 실험 결과 KLUE-BERT 사전학습 언어모델을 미세조정한 모델이 정확도 75.853, F1 스코어 63.503로 가장 높은 성능을 보였다.

위 성능은 수집된 연구보고서의 문장에 한정한 성능이며, 데이터 증강에 사용되기에는 부족하여 다른 방법론 혹은 연구 주제 분야 확대 등을 통한 추가 실험의 여지가 있다. 향후 연구에는 수집된 연구보고서의 규모를 확장하거나 다른 주제 분야의 보고서를 추가적으로 수집하여 데이터셋 규모를 확장하고 데이터셋의 이용 가치를 좀 더 높이고자 한다. 또한 데이터 증강 모델을 별도로 구축하여 데이터 구축에 소모되는 비용을 줄이는 연구도 병행하여 진행하고자 한다.

References

[1] J. H. Kim, K. J. Lee, S. Y. Kim, and S. K. Lee, “Trends and characteristics of 2021 social problem-solving R&D investment,” Sa.gwa.plus(사.과.플러스), KISTEP(Korea Institute of Science & Technology Evaluation & Planning), No.4, pp.3-29, 2021.

[2] W. Song and J. Song “How is the Social problem-Solving R&D Done?,” *Journal of Science & Technology Studies*, Vol.18, No.3, pp.255-288, 2018.

[3] Korea Institute of Science & Technology Evaluation and Planning, “Social problem-solving R&D guidelines for spreading field application,” Ministry of Science and ICT, 2021.

[4] W. J. Kim and S. Y. Lee, “A study on the social problem-solving R&D policy improvement” in *Proceedings of the Korean Institute of Communication Sciences Conference*, pp.228-229, 2018.

[5] S. J. Seong, S. C. Kim, S. W. Lee, and W. J. Cha, "Rhetorical sentence classification using context information," in *Proceedings of the 33th Annual Conference on Human and Language Technology*, pp.316-319, 2021.
[6] S. K. Song, Y. S. Choi, S. P. Choi, H. S. Oh, S. H. Myaeng, H. W. Chun, and C. H. Jeong, "Procedural Knowledge Extraction on Medical Documents," *Journal of KIISE*, Vol. 18, No.2, pp.123-127, 2012.
[7] J. S Yun, G. Y. Kim, Y. C. Jeong, H. S. Oh, and D. Suh, "Development of deep learning technologies and applications for the information extraction of S&T open texts," *Korea Institute of Science and Technology Information*, 2017, Report num.: TRKO201700000489.
[8] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Tagging assistant for scientific articles," *INTAP 2018: Intelligent Technologies and Application*, Vol.932, pp.351-362, 2019.
[9] T. H. Lee, Y. M. Kim, E.i Jeong, and S. O. Na, "Query intent detection for medical advice: Training data construction and intent classification," *Journal of KIISE*, Vol.48, No.8, pp.878-884, 2021.
[10] G. H. Lee, Y. H. Park, and K. J. Lee, "Building a Korean text summarization dataset using news articles of social media," *KIPS Transactions on Software and Data Engineering*, Vol.9, No.8, pp.251-258, 2020.
[11] H. Kong, H. Yoon, M. Hyun, H. Lee, and J. Seol, "KorSciQA 2.0: Question answering dataset for machine reading comprehension of Korean papers in science & Technology domain," *Journal of KIISE*, Vol.49, No.9. pp.686-695, 2022.
[12] S. Park et al., "Klue: Korean language understanding evaluation," *arXiv preprint arXiv:2105.09680*, 2021.
[13] pdfminer.six [Internet], https://github.com/pdfminer/pdfminer.six
[14] Sang-Kil Park, Byeng Il Ko, Korean-Sentence-Splitter [Internet], https://github.com/likejazz/korean-sentence-splitter
[15] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park and Chisung Song et al. KLUE BERT base [Internet], https://huggingface.co/klue/bert-base
[16] Junbum Lee. KcBERT: Korean Comments BERT [Internet], https://github.com/Beomi/KcBERT
[17] Junbum Lee. KcELECTRA: Korean Comments ELECTRA [Internet], https://github.com/Beomi/KcELECTRA
[18] Jangwon Park. KoELECTRA: Pretrained ELECTRA Model for Korean [Internet], https://github.com/monologg/KoELECTRA

신 현 호

https://orcid.org/0000-0001-7130-152X
e-mail : shinhh9554@gmail.com
2020년 경기대학교 문헌정보학과(학사)
2022년 경기대학교 문헌정보학과(석사)
관심분야 : 기계독해, 정보추출, 문장생성, 자연어처리, 딥러닝

정 선 기

https://orcid.org/0000-0002-3229-3368
e-mail : jsk1610@kyonggi.ac.kr
2022년 경기대학교 문헌정보학과 (학사)
2022년 ~ 현 재 경기대학교 문헌정보학 석사과정
관심분야 : 정보추출, 컴퓨터비전, 딥러닝

전 홍 우

https://orcid.org/0000-0002-9584-7065
e-mail : hw.chun@kisti.re.kr
2002년 고려대학교 컴퓨터학과(학사)
2004년 고려대학교 컴퓨터학과(석사)
2007년 도쿄대학교 컴퓨터과학과(박사)
2022년 ~ 현 재 한국과학기술정보연구원 미래기술분석센터장
관심분야 : 자연어처리기술기반 사회문제예측 및 해결방안 분석

권 이 남

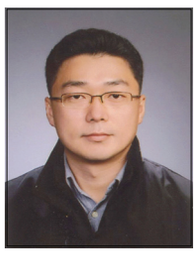
https://orcid.org/0000-0002-3503-7669
e-mail : ynkwon@kisti.re.kr
1996년 한국외국어대학교 MIS(석사)
2022년 충남대학교 컴퓨터공학과(박사)
1996년 ~ 현 재 한국과학기술정보연구원 기술지능연구팀 책임연구원
관심분야 : 데이터 및 SW공학, 치매 조기 예측, 빅데이터기반 증강분석, 사회이슈 탐지, 원부자재 공급망 사전감지

이 재 민

https://orcid.org/0000-0002-4011-987X
e-mail : jmlee@kisti.re.kr
2001년 서울대학교 물리학과(학사)
2008년 서울대학교 물리학과(박사)
2008년 ~ 2010년 삼성전자 반도체연구소 책임연구원
2010년 ~ 현 재 한국과학기술정보연구원 기술지능연구팀 팀장
관심분야 : 기계학습, 과학계량학(Scientometrics), 기술인텔리전스

박 강 희

https://orcid.org/0000-0002-1553-1942
e-mail : can17@kisti.re.kr
2008년 아주대학교 산업공학과(학사)
2014년 아주대학교 산업공학과(박사)
2014년 ~ 2017년 IBK 기업은행 경제연구소
2017년 ~ 현 재 한국과학기술정보연구원 기술지능연구팀 선임연구원
관심분야 : 경제-금융 시계열 예측, 원부자재 공급망 사전감지, 기업분석 알고리즘, 중소기업 지원모델, 핀테크



최 성 필

https://orcid.org/0000-0002-2153-3792
e-mail : spchoi@kgu.ac.kr
1996년 부산대학교 전자계산학과(학사)
1998년 부산대학교 전자계산학과(석사)
2012년 한국과학기술원 정보통신공학과 (박사)
2001년 ~ 2014년 한국과학기술정보연구원 선임연구원
2014년 ~ 현 재 경기대학교 문헌정보학과 부교수
관심분야 : 정보추출, 대화시스템, 텍스트마이닝, 자연어처리, 딥러닝