보안 과제( ), 일반 과제(O) / 공개(O), 비공개( )

과학기술혁신정책지원사업 제1차 연도 최종 보고서

발간등록번호
11-1721000-000845-01

혁신정책 / 2023-010

AI 기반 국가연구개발사업 평가지원체계 구축방안 연구

2024. 02. 10.

한국과학기술기획평가원

과학기술정보통신부

제 출 문

과학기술정보통신부 장관 귀하

‘AI 기반 국가연구개발사업 평가지원체계 구축방안 연구’(연구개발기간 : 2023.02.11~2024.02.10.) 과제의 최종보고서를 제출합니다.

2024. 02.10.

주관연구개발기관명 한국과학기술기획평가원
공동연구개발기관명 연세대학교 산학협력단

주관연구책임자 부연구위원 유지은
주관연구기관 참여연구원 연구원 원정연
연구위원 이일환
연구위원 김상일

공동연구책임자 교수 송민
공동연구기관 참여연구원 고영수
안윤주
강주영
정민화

외부 연구진 서울테크노파크 이은아
㈜광개토연구소 강지훈
한국과학기술연구원 김다경
연세대학교 승태진

국가연구개발혁신법 시행령 제35조에 따라 최종보고서 열람에 동의합니다.

최종보고서

보안등급
일반[√], 보안[ ]

중앙행정기관명	과학기술정보통신부	사업명	사업명	과학기술혁신정책지원사업	
전문기관명 (해당 시 작성)			내역사업명 (해당 시 작성)		
공고번호		총괄연구개발 식별번호 (해당 시 작성)			
		연구개발과제번호		CI23025	

기술분류	국가과학기술 표준분류	SB1207. 과학기술	85%	SB1102. 정책결정/집행	10%	SB0899. 달리 분류되지 않는 행정관리	5%	
	부처기술분류 (해당 시 작성)	1순위 소분류 코드명	%	2순위 소분류 코드명	%	3순위 소분류 코드명	%	

총괄연구개발명 (해당 시 작성)	국문	
	영문	
연구개발과제명	국문	AI 기반 국가연구개발사업 평가지원체계 구축방안 연구
	영문	A study on the establishment of an AI-based national R&D project evaluation support system

주관연구개발기관	기관명	한국과학기술기획평가원	사업자등록번호	229-82-01678
	주소	충북 음성군 맹동면 원중로 1339	법인등록번호	110271-0004210

연구책임자	성명		유지은	직위	부연구위원
	연락처	직장전화	043-750-2515	휴대전화	010-6402-2047
		전자우편	youje8@kistep.re.kr	국가연구자번호	1178 2133
연구개발기간	2023.2.11. - 2024.2.10. (12 개월)				

연구개발비 (단위: 천원)	정부지원 연구개발비	기관부담 연구개발비		그 외 기관 등의 지원금				합계			연구개발비 외 지원금
				지방자치단체		기타( )					
	현금	현금	현물	현금	현물	현금	현물	현금	현물	합계	
	90,000							90,000			

공동연구개발기관 등 (해당 시 작성)	기관명	책임자	직위	휴대전화	전자우편	비고	
						역할	기관유형
공동연구개발기관	연세대학교 산학협력단	송민	교수	010-5535 - 6647	min.song@yonsei.ac.kr	책임자	대학
위탁연구개발기관							
연구개발기관 외 기관							

연구개발담당자 실무담당자	성명		원정연	직위	연구원
	연락처	직장전화	043-750-2673	휴대전화	010-9130-9387
		전자우편	jywon0210@kistep.re.kr	국가연구자번호	1291 6389

이 최종보고서에 기재된 내용이 사실임을 확인하며, 만약 사실이 아닌 경우 관련 법령 및 규정에 따라 제재처분 등의 불이익도 감수하겠습니다.

2024 년 2 월 10 일

연구책임자: 유 지 은 (인)

주관연구개발기관의 장: 한국과학기술기획평가원장 정병선 (직인)
공동연구개발기관의 장: 연세대 산학협력단장 홍종일 (직인)

중앙행정기관의 장 귀하

〈 요 약 서 〉

<table>
<tr><td colspan="2">사업명</td><td colspan="3">과학기술혁신정책지원사업</td><td colspan="2">총괄연구개발 식별번호
(해당 시 작성)</td><td colspan="2"></td></tr>
<tr><td colspan="2">내역사업명
(해당 시 작성)</td><td colspan="3"></td><td colspan="2">연구개발과제번호</td><td colspan="2">CI23025</td></tr>
<tr><td rowspan="2">기술분류</td><td>국가과학기술
표준분류</td><td>SB1207. 과학기술</td><td>85%</td><td>SB1102. 정책결정/집행</td><td>10%</td><td>SB0899. 달리 분류되지 않는 행정관리</td><td>5%</td></tr>
<tr><td>부처기술분류
(해당 시 작성)</td><td>1순위 소분류 코드명</td><td>%</td><td>2순위 소분류 코드명</td><td>%</td><td>3순위 소분류 코드명</td><td>%</td></tr>
<tr><td colspan="2">총괄연구개발명
(해당 시 작성)</td><td colspan="6"></td></tr>
<tr><td colspan="2">연구개발과제명</td><td colspan="6">AI 기반 국가연구개발사업 평가지원체계 구축방안 연구</td></tr>
<tr><td colspan="2">전체 연구개발기간</td><td colspan="6">2023. 2. 11. - 2024. 2. 10. (12 개월)</td></tr>
<tr><td colspan="2">총 연구개발비</td><td colspan="6">총 90,000천원
(정부지원연구개발비: 90,000천원)</td></tr>
<tr><td colspan="2">연구개발단계</td><td colspan="6">기초[ ] 응용[ ] 개발[ ] 기타(위 3가지에 해당되지 않는 경우)[√]</td></tr>
<tr><td rowspan="6">연구개발
목표 및 내용</td><td colspan="2">최종 목표</td><td colspan="5">○ 기존 평가 데이터를 AI가 활용할 수 있는 형태로 정제하고, 이를 활용하여 평가 프로세스에 적용할 수 있는 AI활용 기법 개발</td></tr>
<tr><td colspan="2">전체 내용</td><td colspan="5">○ 기존에 수집 보관하고 있는 국가연구개발사업 평가 관련 데이터에 대한 정제, 저장 등 디지털전환 기반 강화
- 최근 몇 년간의 다양한 형태로 확보한 국가연구개발사업 평가 과정의 산출물을 저장 가능한 방식으로 통일하고, 이를 활용 가능한 방식으로 저장한 후 AI 모델을 통해 학습
○ AI 방법론의 실제 국가연구개발사업 평가프로세스 적용 방안 마련
- 부처별 자체평가에 대한 정성의견과 등급간 뉘앙스 분석, 전략계획서의 수립 및 점검 지원을 위한 사업특성에 맞는 성과지표 추천 기능 등</td></tr>
<tr><td rowspan="2">1단계
(해당 시 작성)</td><td>목표</td><td colspan="5"></td></tr>
<tr><td>내용</td><td colspan="5"></td></tr>
<tr><td rowspan="2">n단계
(해당 시 작성)</td><td>목표</td><td colspan="5"></td></tr>
<tr><td>내용</td><td colspan="5"></td></tr>
</table>

연구개발성과	○ 자연어처리의 정책적 활용 관련 문헌연구 결과 및 시사점 ○ 국가연구개발사업 평가 데이터셋 2종 - 성과목표지표 및 평가계획 관련 데이터셋 - 중간평가 자체평가 결과 관련 데이터셋 ○ AI 방법론의 실제 국가연구개발사업 평가프로세스 적용 방안 ○ BERT모델을 활용한 관련사업 분석 및 성과지표추천 알고리즘(안) ○ 머신러닝 및 딥모델을 활용한 자체평가 등급 설정 지원 알고리즘(안)
연구개발성과 활용계획 및	○ 성과평가정보시스템과 범부처 연구지원시스템의 연계 등 시스템 개선에 데이터와 AI를 활용한 지원체계 마련

<table>
<tr><td>기대 효과</td><td colspan="12">○ 국가연구개발사업 평가 결과의 신뢰성을 제고하고, 전략계획서 수립의 완성도 향상 및 점검에 대한 신뢰성 강화</td></tr>
<tr><td>연구개발성과의 비공개여부 및 사유</td><td colspan="12"></td></tr>
<tr><td rowspan="2">연구개발성과의 등록·기탁 건수</td><td rowspan="2">논문</td><td rowspan="2">특허</td><td rowspan="2">보고서 원문</td><td rowspan="2">연구 시설 ·장비</td><td rowspan="2">기술 요약 정보</td><td rowspan="2">소프트 웨어</td><td rowspan="2">표준</td><td colspan="2">생명자원</td><td rowspan="2">화합물</td><td colspan="2">신품종</td></tr>
<tr><td>생명 정보</td><td>생물 자원</td><td>정보</td><td>실물</td></tr>
<tr><td></td><td></td><td></td><td>√</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="3">세부 정량적 연구개발성과 건수</td><td colspan="3">과학적 성과</td><td colspan="8">사회적 성과</td><td rowspan="2">기타</td></tr>
<tr><td>논문 게재</td><td>학술 회의 발표</td><td>보고서 원문</td><td>법령 반영</td><td>정책 활용</td><td>안건 상정</td><td>제도 개선</td><td>다른 연구에 활용</td><td>국제 협력</td><td>(정책) 홍보</td><td>포상 ·수상</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>√</td></tr>
<tr><td>국문핵심어 (5개 이내)</td><td colspan="3">국가연구개발 사업 평가</td><td colspan="3">인공지능</td><td colspan="3">성과지표</td><td colspan="2">전략계획서</td><td colspan="2">성과평가 정보시스템</td></tr>
<tr><td>영문핵심어 (5개 이내)</td><td colspan="3">National R&D Program Evaluation</td><td colspan="3">Artificial intelligence</td><td colspan="3">R&D Evaluation indicator</td><td colspan="2">R&D Program Strategic plan</td><td colspan="2">R&D Performance & Evaluation Information System</td></tr>
</table>

요 약 문

1. 연구개발과제의 개요

ㅇ ChatGPT 등 기술의 발전으로, 인공지능이 공공 분야에서 더욱 광범위하게 활용되고 정책적 의사결정과 행정 효율화에 핵심적인 역할을 수행할 것으로 기대됨

ㅇ 기존 수치데이터의 계량적 분석에서 더 나아가 문서형태로 보관되고 있던 텍스트 공공데이터에 대한 활용가치와 분석 필요성이 증대되는 가운데, 국내외를 비롯해 다양한 부문에서 공공데이터의 인공지능 학습 및 활용을 통한 정책 서비스 부가가치 제고에 관한 연구가 이어지고 있는 추세임

ㅇ 한편, 국가연구개발 사업평가를 수행하는 과정에서 행정부담과 반복적 업무로 인한 비효율에 대하여 연구현장의 목소리가 제기되고 있음

ㅇ 본 연구는 위와 같은 비효율을 해소하기 위해 인공지능을 활용하여 평가에 필요한 정보를 효율적으로 제공함으로써 업무 효율성을 제고할 수 있는 방안을 마련하고자 함

ㅇ 이를 위해 기존 평가 데이터를 AI가 활용할 수 있는 형태로 정제하고, 이를 활용하여 평가 프로세스에 적용할 수 있는 AI 활용 기법 개발

과업목표 1) 국가연구개발사업 전략계획서 상 성과목표별 성과지표 작성 시 유사 사례를 추천해주는 인공지능 모델 탐색

과업목표 2) 각 부처가 수행한 자체평가 결과의 등급분류 및 예측을 자동화하여 평가의견과 등급 간 정합성을 제고할 수 있는 인공지능 모델 탐색

ㅇ 연구결과를 활용하여 인공지능을 활용한 국가연구개발사업 평가 프로세스 개선방안을 마련하여 실질적으로 평가업무의 효율성 증대에 기여할 것으로 기대되며, 장기적 관점에서 사업평가의 효율성과 신뢰성을 제고할 수 있는 체계 마련의 기반이 될 것으로 기대

2. 연구개발과제의 수행 과정 및 수행 내용

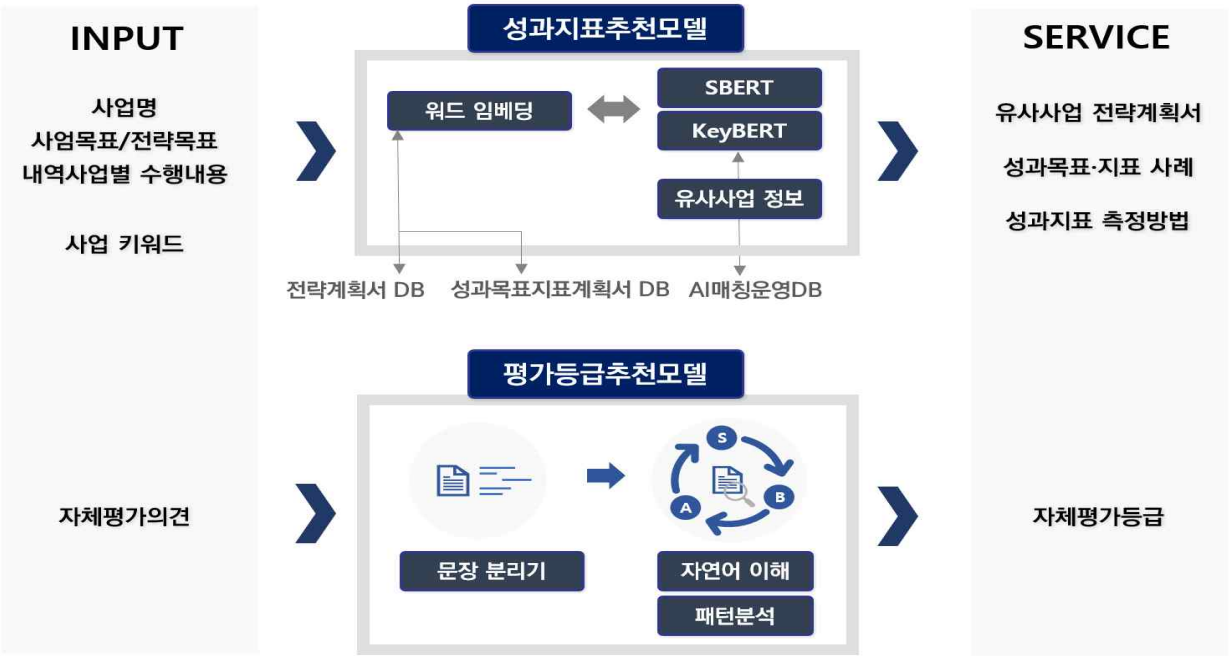
[표] 연구개발과제 추진체계

연구개발과제명	
AI 기반 국가연구개발사업 평가지원체계 구축 방안 연구	
과기정통부 성과평가정책과	
(주관) 한국과학기술기획평가원	
역할	· **문헌연구** – 인공지능의 국가연구개발사업 평가 프로세스 적용 관련 문헌연구 및 선행연구 · **데이터셋 구축** – 전략계획서, 성과목표지표계획서, 자체평가 보고서 · **모델 검토** – 모델 분석결과의 검토 및 피드백 · **모델 검증** – 평가 담당자 및 전문가 평가를 통한 모델 검증 · **적용방안 마련** – AI 기반 평가지원체계 구축 방안(안) 마련
(공동) 연세대학교 산학협력단	
역할	· **문헌연구** – 활용 가능한 자연어 처리 및 인공지능 모델 관련 문헌연구 – 형태소 분석기 관련 문헌연구 · **데이터 전처리** – 말뭉치 구성, 불용어 사전 구축 등 데이터 전처리 · **모델 탐색** – SBERT, KeyBERT, 평가의견의 오피니언 마이닝 및 등급 다중분류가 가능한 딥러닝/머신러닝 모델 탐색 및 분석 · **모델 구축** – 성과지표 설정사례 및 자체평가 등급 추천이 가능한 모델 구축

- o 인공지능을 활용한 국가연구개발사업 평가지원 서비스 구축을 위하여 KISTEP-연세대 공동연구 추진
- o 평가자, 피평가자 및 기술분야 전문가로 구성된 검증단과 과제 자문위원을 구성하여 모델 분석결과에 대한 검증·평가 및 과제 추진방향 보완 등 현장의 목소리를 반영
- o 연구개발계획을 토대로 세부 추진계획을 수립하여 부처와 협의를 거쳐 추진방향 결정
- o 문헌연구 및 선행연구 고찰 → 데이터 구축 → 모델 구축 및 보완 → 전문가 및 사용자 검증 → 최종모델 도출의 5단계로 나누어 추진

3. 연구개발과제의 수행 결과 및 목표 달성 정도

- o 문헌연구와 선행 사례 연구를 바탕으로 본 연구는 아래와 같이 AI 기반 평가지원체계 마련을 위한 기반연구로서 사업평가 프로세스 효율화가 필요한 세부 업무 프로세스를 선정하고, 인공지능을 활용하여 부가적인 정보를 제공하고, 의사결정을 지원



[그림] 인공지능 기반 국가연구개발사업 평가지원체계 개념(안)

o 연구결과와 그 질적 수준은 아래와 같음

① (평가등급 설정지원) 평균 80%의 정확도로 평가등급 추천이 가능한 모델 구축

- KoBERT 기반 모델은 최대 90%, 최소 73%의 등급 부여 정확도를 보였으며, 이는 선행연구(최대 88%, 최소 56%) 및 오피니언 마이닝 관련 연구(최대 89.3%, 최소 46.8%) 대비 제고된 성능에 해당

- 전문가가 부여한 평가등급과 비교를 통해 연구 결과의 활용 가능성 검증 완료

② (성과지표 설정지원) 검색항목, 유사도/파라미터 기준치 등 조건별 관련 사업 검색이 가능한 모델 구축

- 사업명, 사업목적, 전략목표, 세부 내용, 통합내용 등 5개 검색항목과 유사도 기준치(60~99%), 파라미터값(0.1~1.0)을 사용자 수요에 맞게 선택하여 유사 사업 검색 가능

- 전문가의 유사성(43% 유사), 유용성(68% 유용) 평가를 통해 AI 추천 결과의 활용 가능성 검증 완료

(1) 세부 정량적 연구개발성과 (과학적 성과-보고서 원문)

연도	보고서 구분	발간일	등록 번호
2023	최종보고서	2024.2.10.	-

(2) 목표 달성 수준

연구목표	완수여부			연구실적
	완수	진행	미완수	
◦ 문헌연구 및 선행연구 고찰	○			
- 선행연구의 시사점과 한계점 도출	○			자연어 처리 개념/활용 관련 문헌연구 국가 R&D 평가지원 개선 관련 선행연구 고찰
- 자연어 처리의 공공 활용 관련 연구	○			증거 기반 정책기획, 소통, 평가 등 4개 공공 활용사례 및 동향 관련 문헌연구
- 활용 모델의 원리/장단점 등 이론연구	○			활용 가능한 딥러닝(2개), 머신러닝(4개) 모델의 개요, 장단점, 활용사례(연구 동향) 등 문헌연구
◦ 국가연구개발사업 평가 데이터셋 구축	○			
- 모델 구축 및 분석용 데이터셋 구축	○			(자체평가보고서) 모델 구축용 데이터 646건 ('22-'23) 구축 완료
	○			(전략계획서) 모델 구축용 데이터 1,005건 ('20-'21) 구축 완료
- 검증 및 평가용 데이터셋 구축	○			(자체평가보고서) 검증 · 평가용 데이터 1,240건 ('19-'21) 구축 완료
	○			(성과목표 · 지표계획서) 검증 · 평가용 데이터 752건 ('16-'21) 구축 완료
◦ 자체평가 등급설정지원 서비스 기획	○			
- 세부 추진계획 수립	○			데이터, 모델, 평가 기준(안) 등 세부 계획 마련 후 부처 보고 완료('23.5월/11월)
- 데이터 전처리	○			한글, 영문, 숫자 등 외 기초 전처리(1,886건) 우수성/핵심성/성과와 3개 부문/ 8개 지표별 말뭉치 및 불용어 사전 구축(3차, '23.6월/9월/10월)
- 자체평가보고서 모델 탐색/성능 비교	○			문맥 이해 및 등급 다중 분류(예측) 가능한 자연어 처리모델 탐색 및 선별 머신러닝(랜덤포레스트, SVM, 나이브베이즈, XG Boost 복합), 딥러닝(KoBERT, RoBERTa) 모델별 분석 결과 및 모델 성능 비교·검증
- 최종모델 테스트/요구사항 검증	○			검증용 데이터 등급 예측 결과 모델 성능 및 정확도 비교 ('22-'23 20개 평가의견 예측 결과) 모델 등급 예측 결과 및 정확도 비교 ('19-'21 1,234개 전수 평가의견 예측 결과) ※ (방법) 중간평가 상위 점검 담당자 7인, 기술 분야별 전문기관 사업평가 담당자 9인이 부여한 평가등급과 구축모델별 등급 예측 결과 비교
◦전략계획서 성과지표설정지원 서비스 기획	○			
- 세부 추진계획 수립	○			데이터, 모델, 평가 기준(안) 등 세부 계획 마련 후 부처 보고 완료('23.5월/11월)
- 전략계획서 데이터 전처리	○			한글, 영문, 숫자 등 외 기초 전처리(1,757건) 사업 세부 내용, 목적, 전략목표 등 유효 문장/키워드 전처리 완료('23.6월)

연구목표	완수여부			연구실적
	완수	진행	미완수	
모델 탐색 및 성능보완	○			SBERT 기반 검색항목 도출(사업명, 사업목적, 전략목표, 사업 세부 내용 등) 유사 사업 검색실험 및 성능보완을 위한 불용어 제거, 파라미터 조정, 검색 결과 가중치 부여 등 키워드 및 문장 검색실험 및 알고리즘 보완 (4차)
- 최종모델 테스트/요구사항 검증	○			전략계획서 통합내용(사업명, 목표, 전략목표, 세부 내용 종합) 기준 유사 사업 추천 결과에 대한 유사성, 유용성 점수 비교 ※ (방법) 전략계획서 담당 간사 6인, 기술 분야별 전문기관 사업평가 담당자 9인이 모델 유사 사업 추천 결과의 유사성, 유용성 평가 특정 키워드 검색에 대한 모델의 추천 결과와 전문가 특정 결과 비교 ※ (방법) 전략계획서 담당 간사 6인, 기술 분야별 전문기관 사업평가 담당자 9인의 유사 사업 특정 결과와 비교 · 평가 JAVA 환경(NTIS 성과평가정보시스템) 개발을 고려한 사전테스트 및 보완사항 도출 완료

4. 연구개발성과의 관련 분야에 대한 기여 정도

- o (문헌연구) 자연어 처리 모델에 대한 이론적 이해를 바탕으로 관련 선행연구를 고찰하고, 이를 통해 본 연구의 차별성을 고려하여 과업 범위 및 연구 목표 설정
- o (데이터) 인공지능 활용을 위해 R&D 사업평가 산출물의 텍스트 데이터셋 구축 및 전처리, 평가의견 특성을 고려한 단어사전 및 불용어 사전 구축
- o (모델 구축) 문헌연구를 통해 활용에 적합한 방법론을 식별하고, 모델 전수에 대한 실험 및 검증을 통해 성능과 한계를 비교하여 최종모델 선정
- o (결과검증) 전문가와 AI 추천 결과의 비교를 통해 AI 기반 평가지원체계 구축 결과의 실용성과 한계 등을 실질적으로 검증

5. 연구개발성과의 관리 및 활용 계획

- o 본 연구결과를 토대로 향후 데이터 및 알고리즘에 대한 보완이 이루어진다면, NTIS 성과평가정보시스템 등과의 연계를 통해 실제 활용 가능한 서비스 구축에도 기여가 가능할 것으로 기대됨

< 별첨 자료 >

중앙행정기관 요구사항	별첨 자료
1. 보고서 원문 제출	1) 최종보고서 원문

목 차

Ⅰ. 서론 ····· 1

제1절 연구배경 ····· 1

제2절 연구목표 및 내용 ····· 2

제3절 연구방법 ····· 3

Ⅱ. 문헌연구 및 선행연구에 대한 고찰 ····· 6

제1절 자연어 처리를 활용한 텍스트 분석 ····· 6

제2절 자연어 처리의 공공부문 활용 ····· 16

제3절 데이터·인공지능 기반 국가R&D 프로세스 개선 사례 ····· 23

제4절 소결 : 국가연구개발사업 평가에의 AI 적용 가능성과 이슈 ····· 38

Ⅲ. 전략계획서 성과지표 설정 지원 서비스 기획 연구 ····· 42

제1절 데이터 ····· 42

제2절 모델 탐색 및 보완 ····· 45

제3절 평가 및 활용성 검증 ····· 78

Ⅳ. 중간평가 등급 설정 지원 서비스 기획 연구 ····· 81

제1절 데이터 ····· 81

제2절 모델 탐색 및 보완 ····· 86

제3절 평가 및 활용성 검증 ····· 109

Ⅴ. 결론 및 시사점 ····· 112

제1절 결론 ····· 112

제2절 시사점 ····· 115

[부록] AI 평가지원서비스 개념설계(안) ····· 117

[부록] 참고문헌 ····· 121

Ⅰ. 서론

제1절 연구배경

ㅇ ChatGPT 등 기술의 발전으로, 인공지능이 공공 분야에서 더욱 광범위하게 활용되고 정책적 의사결정과 행정 효율화에 핵심적인 역할을 수행할 것으로 기대됨

ㅇ 기존 수치데이터의 계량적 분석에서 더 나아가 문서형태로 보관되고 있던 텍스트 공공데이터에 대한 활용가치와 분석 필요성이 증대되는 가운데, 국내외를 비롯해 다양한 부문에서 공공데이터의 인공지능 학습 및 활용을 통한 정책 서비스 부가가치 제고에 관한 연구가 이어지고 있는 추세임

ㅇ 한편, 국가연구개발 사업평가를 수행하는 과정에서 행정부담과 반복적 업무로 인한 비효율에 대하여 연구현장의 목소리가 제기되고 있음

- 구체적으로 전략계획서 작성 시에는 관련사업의 성과지표 사례를 찾기 위해 사업 담당자 및 KISTEP 간사가 수기로 검색을 해야하는 비효율이 반복되고 있음

- 자체평가보고서 작성 및 상위점검 시에는 평가의견과 맞지 않는 등급이 부여되어 기술상의 오류나 의견-등급 간 정합성을 검토하는 데 불필요한 에너지를 소모하는 일이 빈번함

ㅇ 본 연구는 위와 같은 비효율을 해소하기 위해 인공지능을 활용하여 평가에 필요한 정보를 효율적으로 제공함으로써 업무 효율성을 제고할 수 있는 방안을 마련하고자 함

- 국정과제 '74. 국가혁신을 위한 과학기술 시스템 재설계' 중 'AI기반 평가지원체계 마련'을 위한 탐색 연구를 통해 국가연구개발 평가에 있어서도 데이터에 기반한 행정의 가능성이 향상되고 있음

- 국가연구개발사업 평가 데이터를 AI가 학습할 수 있는 형태로 축적하는 등 데이터 활용 체계를 마련하고, 관련 의사결정에 직접 도입하여 적용할 수 있는 모델들을 검토함으로써 향후 활용 가능성과 한계점 등을 진단하고자 함

제2절 연구목표 및 내용

1. 연구목표

ㅇ 국가연구개발사업 평가 및 전략계획서 수립·점검 시 보다 나은 정보에 입각한 의사 결정과 효과적인 자원 할당을 통해 사업평가 및 전략계획서 수립·점검의 효율성과 신뢰성 제고

ㅇ 이를 위해 기존 평가 데이터를 AI가 활용할 수 있는 형태로 정제하고, 이를 활용하여 평가 프로세스에 적용할 수 있는 AI 활용 기법 개발

과업목표 1) 국가연구개발사업 전략계획서 상 성과목표별 성과지표 작성 시 유사 사례를 추천해주는 인공지능 모델 탐색

과업목표 2) 각 부처가 수행한 자체평가 결과의 등급분류 및 예측을 자동화하여 평가의견과 등급 간 정합성을 제고할 수 있는 인공지능 모델 탐색

2. 연구내용 및 기대효과

ㅇ 정책적 활용에 있어 인공지능의 기능을 탐색하고, 공공부문 활용에 관한 사례연구를 통해 국가연구개발사업 평가에의 인공지능 활용 가능성과 기대효과, 적용한계 등을 검토

ㅇ 인공지능 기반 국가 R&D사업 평가 지원을 위해 기존 평가데이터의 정제·저장·관리, 활용 가능한 모델 탐색 및 검증 추진

- 기존에 수집 보관하고 있는 국가연구개발사업 평가 관련 데이터에 대한 정제, 저장 등 디지털전환 기반을 가오하하고 데이터의 정확성, 신뢰성, AI 알고리즘과의 호환성 등을 함께 고려하여 데이터 정제 수행

- 관련사업 분석을 위한 인공지능 모델과 자체평가 등급 분류가 가능한 자연어 처리 모델을 탐색하고 성능을 비교·검토하여 활용에 적합한 최적의 모델 식별

- 탐색한 인공지능 모델의 예측·분류 결과와 평가를 수행하는 담당자 및 전문가의 예측·분류 결과를 비교함으로써 모델 성능 및 유용성을 검증

ㅇ 연구결과를 활용하여 인공지능을 활용한 국가연구개발사업 평가 프로세스 개선방안을 마련하여 실질적으로 평가업무의 효율성 증대에 기여할 것으로 기대되며, 장기적 관점에서 사업평가의 효율성과 신뢰성을 제고할 수 있는 체계 마련의 기반이 될 것으로 기대

제3절 연구방법

1. 연구방법 및 추진체계

[표 1-1] 연구개발과제 추진체계

연구개발과제명	
AI 기반 국가연구개발사업 평가지원체계 구축 방안 연구	
과기정통부 성과평가정책과	
(주관) 한국과학기술기획평가원	
역할	**· 문헌연구** – 인공지능의 국가연구개발사업 평가 프로세스 적용 관련 문헌연구 및 선행연구 **· 데이터셋 구축** – 전략계획서, 성과목표지표계획서, 자체평가보고서 **· 모델 검토** – 모델 분석결과의 검토 및 피드백 **· 모델 검증** – 평가 담당자 및 전문가 평가를 통한 모델 검증 **· 적용방안 마련** – AI 기반 평가지원체계 구축 방안(안) 마련
(공동) 연세대학교 산학협력단	
역할	**· 문헌연구** – 활용 가능한 자연어 처리 및 인공지능 모델 관련 문헌연구 – 형태소 분석기 관련 문헌연구 **· 데이터 전처리** – 말뭉치 구성, 불용어 사전 구축 등 데이터 전처리 **· 모델 탐색** – SBERT, KeyBERT, 평가의견의 오피니언 마이닝 및 등급 다중분류가 가능한 딥러닝/머신러닝 모델 탐색 및 분석 **· 모델 구축** – 성과지표 설정사례 및 자체평가 등급 추천이 가능한 모델 구축

o 인공지능을 활용한 국가연구개발사업 평가지원 서비스 구축을 위하여 KISTEP-연세대 공동연구 추진

- 국가연구개발사업 평가에 장기간 노하우와 경험이 있는 KISTEP이 주관연구개발기관으로 참여하여 서비스 구축의 지향점과 목표를 설정

- 자연어 처리 및 오피니언 마이닝 인공지능 모델에 관한 연구 경력과 서비스 구축 경험이 풍부한 연세대 연구팀을 공동연구진으로 구성하여 과제 수행

- 연구진 월례 회의를 통해 과제기획 – 데이터 정제 및 전처리 – 모델 선정 – 분석결과 검토 등 기획 및 진행 전과정에서 기관의 전문성을 고려한 역할 분담

- 연구진행 과정에서 혁신본부 성과평가정책과 중간보고 및 검토 등 의견교환을 통한 과제 추진방향 보완

o 평가자, 피평가자 및 기술분야 전문가로 구성된 검증단과 과제 자문위원을 구성하여 모델 분석결과에 대한 검증·평가 및 과제 추진방향 보완 등 현장의 목소리를 반영

2. 추진일정 및 과정

o 연구개발계획을 토대로 세부 추진계획을 수립하여 부처와 협의를 거쳐 추진방향 결정

o 데이터 구축 → 모델 구축 및 보완 → 전문가 및 사용자 검증 → 최종모델 도출의 4단계로 나누어 [표]의 일정과 같이 추진

[표 1-2] 연구개발과제 추진일정

수행내용		추진일정 ('23.2월 ~ '24.2월)													산출물
		2	3	4	5	6	7	8	9	10	11	12	1	2	
선행연구 등 문헌연구															최종 보고서
자체평가등급설정지원	세부 추진계획 수립														보고자료
	데이터셋 구축														원시, 원천, 라벨링 데이터
	모델 탐색 / 성능비교														모델 특징 및 성능 비교결과
	모델 검증·평가														전문가 검증결과
	부처 결과보고														보고자료
	최종모델 성능개선														예측모델
	최종보고서 작성														최종 보고서
전략계획서성과지표설정지원	세부 추진계획 수립														보고자료
	데이터셋 구축														데이터
	SBERT														알고리즘
	KeyBERT														알고리즘
	모델 검증·평가														전문가 검증결과
	부처 결과보고														보고자료
	최종모델 성능개선														분류모델
	최종보고서 작성														최종보고서

○ 연세대-KISTEP 공동연구의 원활한 추진을 위하여 연구진 월례회의를 개최하고, 분석결과에 대한 검토 및 추진방향 공유 등을 수행함

[표 1-3] 연구진회의 추진일정 및 개요

일시	회의	개요
23.2.18.	1차 월례회의	연구계획 및 예산, 일정 등 논의, 탐색모델 1차 선별
23.3.30.	2차 월례회의(킥오프)	데이터 1차 검토결과 논의, 22년 연구결과 검토
23.4.17. 23.4.19.	3차 월례회의	성과지표 관련사업 분석 샘플테스트 결과 논의, 알고리즘 검토 22, 23년 자체평가보고서 raw data 검토결과 논의, 데이터 수집 및 전처리 가이드 초안 검토
23.5.3.	4차 월례회의	성과지표 설정지원 서비스 2차 샘플테스트 결과 검토 자체평가지침 비교를 통한 데이터 수집 가이드라인(안) 최종검토
23.5.22.	5차 월례회의	자체평가 데이터 수집결과 검토 및 보완 가이드
23.6.24. 23.7.12. 23.8.19.	6차 월례회의	형태소분석기 검토, 불용어 검토 및 제거 관련사업 KeyBERT 샘플테스트 결과 검토
23.9.25.	7차 월례회의	자체평가데이터 머신러닝 모델 실험결과 검토 KeyBERT모델 알고리즘 보완사항 검토
23.10.31.	8차 월례회의	불용어 재검토 및 2차 제거 KeyBERT 단어-문서간 유사도 실험 결과 검토
23.11.30.	9차 월례회의	자체평가데이터 딥러닝/머신러닝 모델 실험결과 비교·검토 SBERT/KeyBERT 평가검증용 분석결과 검토
23.12.4.	10차 월례회의	전문가 평가 검증 결과 공유

○ 과제 추진과 관련하여 데이터 구축, 모델 성능 검토, 최종결과 도출 시 담당부처인 과기정통부 과기혁신본부 성과평가정책과 보고를 통해 추진방향을 조정함

[표 1-4] 소관 부처 보고 및 조정사항 개요

일시	보고	개요
23.3.1.	연구개발계획 보고 (서면)	최종 연구개발계획 보고
23.5.24.	데이터 구축/전처리 결과 보고	데이터 구축 및 전처리 결과 보고 향후 추진계획 보고
23.11.6.	중간결과 및 모델성능검토 추진계획 보고 (서면)	중간평가 모델별 분석결과 및 최종 모델 관련 보고 성과지표 관련사업 분석 검증결과 보고
23.12.20.	연구결과 보고	데이터, 모델 구축, 평가 및 검증 관련 연구내용 최종 결과(안) 보고 및 피드백

Ⅱ. 문헌연구 및 선행연구에 대한 고찰

제1절 자연어 처리를 활용한 텍스트 분석

1. 개요

○ 자연어 처리(Natural Language Processing, NLP)는 컴퓨터가 인간의 언어를 이해하고 해석하여 조작하도록 돕는 인공지능(AI)의 한 분야[1]로서, 텍스트를 자동 분석하고 후속 분석에 의미있는 정보를 추출하는 방법에 대한 연구가 주를 이룸

- 언어를 보다 짧은 기본 요소로 분해하고, 각 요소 간 관계와 상호작용을 통해 어떻게 의미를 이루는지를 탐구하는 과정으로 컴퓨터 과학이나 전산학 등 뿐만 아니라, 언어학 등 많은 분야가 동원되는 다학제적 연구로 발전하고 있음

- 가장 궁극적인 목표는 가공되지 않은 언어 그대로를 입력한 후 언어학적 지식과 컴퓨터의 알고리즘을 활용해 텍스트를 보다 분석에 가치있는 형식으로 전환하거나 보강하는 것임

○ 본 연구는 국가연구개발 사업평가 과정에서 산출되는 문서들을 이해하고 분석할 수 있는 자연어 처리 방법을 연구함으로써 실질적인 지원과제를 도출해내고자 자연어 처리를 활용한 텍스트 분석에 대해 자세히 살펴봄

- 자연어 처리 방법은 다양하게 분류할 수 있으나, 정책과정에의 활용을 고려한 Jin, Z. & Mihalcea, R. (2022)의 기준에 따라 분류, 토픽모델링, 이벤트 추출, 점수예측 등 4가지로 구분하여 살펴봄[2]

[표 2-1] 자연어 처리를 통해 추출된 정보의 유형 및 적용 예

자연어처리 방법	추출 정보	목적 및 적용 예시
분류	텍스트 카테고리(분류)	감정분석(오피니언 마이닝), 입장(stance) 분석
토픽모델링	텍스트 내 주요 토픽	어젠다 발굴, 주제 요약
이벤트 추출	사건의 리스트	뉴스 사건, 국제정세 동향 및 분쟁사건 리스트업
점수 예측	척도	텍스트 스케일링

자료: Jin, Z., & Mihalcea, R. (2022)를 재구성하여 작성

1) 자연어 처리(NLP), SAS 인사이트, https://www.sas.com/ko_kr/insights/analytics/what-is-natural-language-processing-nlp.html (최종접속일: 23.7.2.)

2) Jin, Z., & Mihalcea, R. (2022). Natural Language Processing for Policymaking. In Handbook of Computational Social Science for Policy (pp. 141-162). Cham: Springer International Publishing

2. 분류(Classification)[3)]

ㅇ 분류는 텍스트를 읽고 범주(category)를 예측하는 방법으로 주어진 텍스트의 특성과 패턴을 학습하여 새 정보에 대한 분류 예측을 수행하는 방법을 일컬음(Kowsari et al., 2019)

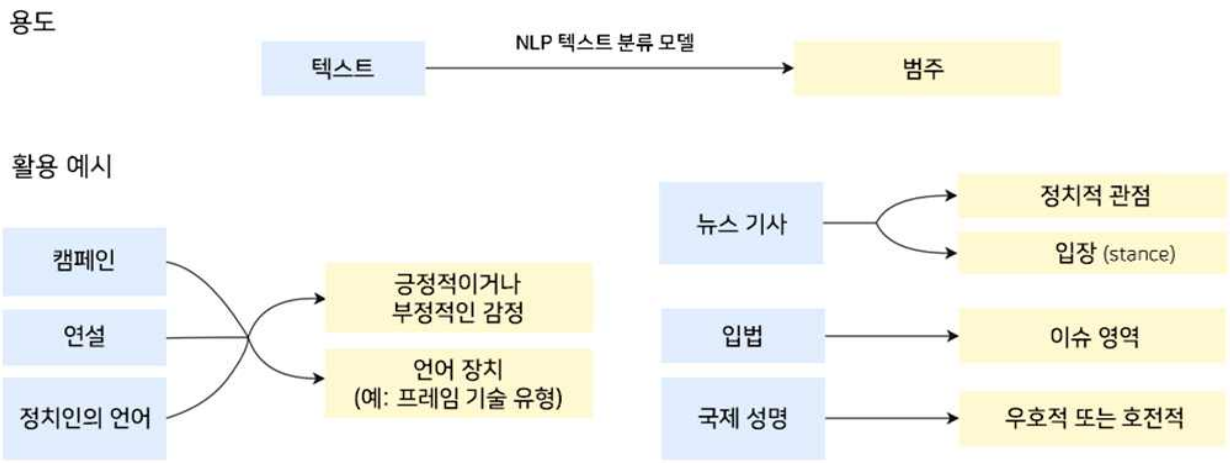
- 분류 예측을 수행하기 위해 먼저 특성 추출(feature extraction)을 통해 텍스트를 단어, 구절 또는 문장의 특징으로 표현하여 벡터 형태로 변환 이를 위해 단어를 밀집 벡터로 표현하고 의미 정보를 보존하는 방식인 단어 임베딩을 활용함

- 다음으로 훈련과 예측을 반복하여 모델 학습을 완료하고, 해당 모델을 사용하여 새로운 텍스트의 분류를 예측함

ㅇ 체계적이지 않고 정리되지 않은 날 것 그대로의(raw) 텍스트를 신속하고 경제적인 방법으로 구조화하여 얻고자 하는 정보를 얻어내고 적용이 가능한 모델로 구현하는 것이 기술의 핵심이라고 할 수 있음

ㅇ 주로 지도학습을 사용하며 나이브 베이즈, 로지스틱 회귀, 랜덤포레스트를 포함한 의사결정트리, 서포트 벡터 머신, 신경망 등이 대표적인 알고리즘에 해당함

ㅇ 감정분석(sentiment analysis) 또는 오피니언 마이닝이라고 불리는 감정분류(Sentiment classification)가 가장 잘 알려진 텍스트 분류의 하위 작업 중 하나이며, 텍스트에서 긍정 또는 부정과 같은 주관적인 정보를 구분해냄



자료: Jin, Z., & Mihalcea, R. (2022)를 재구성하여 작성

[그림 2-1] 텍스트 분류 모델의 범주 예측 활용 예시

3) 텍스트 분류는 아래와 같이 파이썬 프로그래밍의 다양한 라이브러리를 활용하여 수행 가능
① Natural Language Toolkit (NLTK): 파이썬 기반의 NLP 라이브러리로, 텍스트 분류에 필요한 다양한 기능과 알고리즘을 제공 (https://www.nltk.org/)
② Scikit-learn: 머신러닝 라이브러리로, 벡터화, 피처 추출, 분류 알고리즘 등을 포함하여 다양한 텍스트 분류 작업을 수행 (scikit-learn.org)
③ TensorFlow (tensorflow.org), PyTorch (https://pytorch.org/): 딥러닝 프레임워크로, 텍스트 분류에 적용할 수 있는 다양한 신경망 모델과 기능을 제공

- ㅇ 방법론 연구의 동향은 사전 훈련된 언어모델의 분류 성능을 개선하기 위하여 기술적으로 보완하는 연구에서 더 나아가 보다 광범위한 활용을 염두한 연구로 확장되고 있음
  - 기존에는 BERT (Devlin et al., 2018), GPT (Radford et al., 2018), RoBERTa (Liu et al., 2019)와 같은 언어 모델을 사용하여 텍스트 분류 성능을 향상시키는 연구를 중심으로 특징 추출, 클래스 불균형 문제 처리 등 보완 연구가 지속됨
  - 최근에는 텍스트 분류 모델의 결정 과정을 해석하고자 설명 가능한 모델 (Palatnik de Sousa et al., 2019), 텍스트 외 다른 형태의 데이터와 결합하여 분류 작업을 수행하는 멀티 모달 분류(Afridi et al., 2021) 등의 연구로 발전
- ㅇ 활용 연구는 주로 정치 캠페인, 연설, 뉴스기사 등의 텍스트에서 긍·부정 정서를 분류해내거나, 입법 이슈 영역을 분류하는 등 형태로 이루어지고 있음
  - 주로 뉴스 기사의 정치적 관점, 특정 주제에 대한 미디어의 입장, 캠페인이 긍정적 정서를 사용하는지 부정적 정서를 사용하는지 여부 등에 대한 연구가 진행되고 있음
  - Hausladen et al. (2020)은 미국 판사의 소속 정당(임명 당시의 여당 등)이 판례 결정에 어떤 영향을 미치는지 연구에서 분류 모델을 활용해 이념적 방향에 대한 분류를 예측함[4)]
  - Collingwood & Wilkerson(2012)는 법안을 19개의 주제로 분류하고 400,000개 이상의 법안 제목을 포함하는 데이터셋을 구축하여 의회 법안에 대한 정책 데이터셋을 효율적으로 구성하는 방식(적정 샘플크기, 알고리즘 등)에 대하여 탐색함[5)]
  - 그 외에도 뉴스 기사의 정치적 관점 분류 (Cabot et al., 2020), 특정 주제에 대한 언론사별 관점 분류 (Luo et al., 2020) , 정치 및 정책적 문제에 대한 감성 분류 (Ansolabehere & Iyengar, 1995; Schrodt, 2000; Schumacher et al., 2016), 미국 회로법원의 사건 결정에 대한 보수 및 진보 성향 분류 (Hausladen et al., 2020)등의 연구가 수행됨
- ㅇ 텍스트 분류는 정책 분야에서 많은 분석 작업을 자동화하는 강력한 도구로 효율적인 정보 추출 및 다양한 응용 분야로 활용될 수 있으나, 데이터의 다양성 및 복잡성으로 인한 분류 정확도 제고의 한계, 데이터의 종속성 및 인과관계 해석의 어려움 등이 존재함

4) Hausladen, Carina I., Marcel H. Schubert, and Elliott Ash. “Text classification of ideological direction in judicial opinions.” International Review of Law and Economics 62 (2020): 105903.

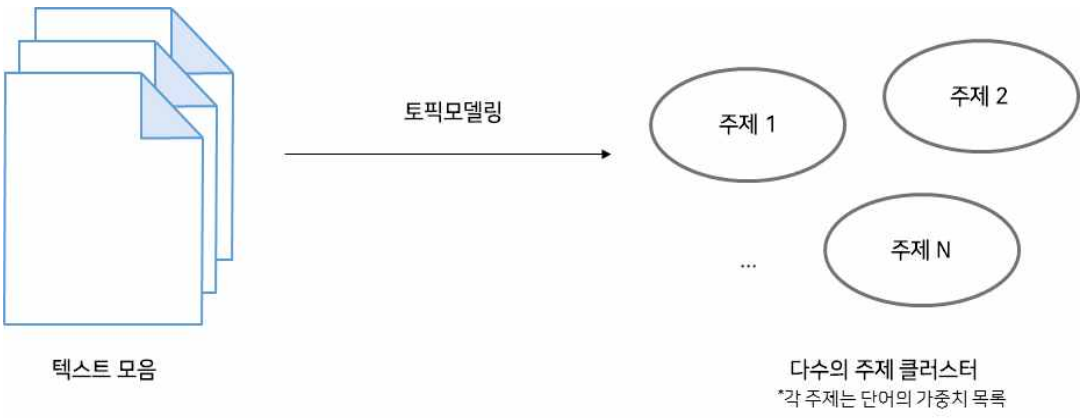
5) Collingwood, Loren, and John Wilkerson. “Tradeoffs in accuracy and efficiency in supervised learning methods.” Journal of Information Technology & Politics 9.3 (2012): 298-318.

ㅇ 텍스트 분류 모델로부터 정확하고 유의미한 결과를 얻기 위해서는 언어의 복잡성, 문맥 의존성, 주관성과 다의성, 그리고 도메인 특성 등 다양한 요인을 고려해 텍스트 분류의 정확성과 유의미한 결과 도출을 위해 고려해야 함

ㅇ 분류 및 적용 목적 등 목표 과제에 따른 사전훈련 모델의 파인튜닝, 어휘사전 구축, 파라미터 세팅, 모델 선택, 특징 선택 등 다양한 기술과 방법을 활용하여 텍스트 분류 모델을 개발하고 적절한 전처리와 평가 과정을 수행함으로써 문제점을 개선할 수 있음

3. 토픽 모델링(Topic Modeling)[6]

ㅇ 토픽 모델링은 말뭉치에서 자주 사용되는 주제 목록을 파악하는 방법으로 문서를 주제의 혼합으로 모델링하고, 사용된 단어의 확률적 분포를 토대로 토픽을 도출하는 방법(Vayansky & Kumar, 2020)으로 LDA, LSA, NMF 등이 있음



자료: Jin, Z., & Mihalcea, R. (2022)를 재구성하여 작성

[그림 2-2] 토픽 모델링을 통한 주제 클러스터 목록 생성 개념도

ㅇ Latent Dirichlet Allocation(LDA)는 텍스트 문서 집합에서 숨겨진(latent) 주제를 추론하는 확률 기반 모델로 주어진 문서 집합에서 빈도가 높은 주제 클러스터들을 생성하는 방법임

- 각 문서의 주제는 소수의 단어 혼합물로 구성되어 있으며, 각 단어의 출현은 문서의 주제에 의해 결정된다고 가정[7]하며, 이 특정 주제들의 집합으로 가정된 문서를 구성하는 단어들의 분포를 확률적으로 계산하여 결과값을 토대로 주제어들의 집합으로 추출함

6) 가장 대표적으로 활용되는 모델 중 하나는 LDA(Latent Dirichlet Allocation) 알고리즘 Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.으로 Python 패키지 NLTK 및 Gensim에서도 사용 가능하며, Python의 scikit-learn 및 R의 topicmodels 패키지 등을 활용하여 LDA 등의 토픽 모델링을 수행할 수 있음. 이외에도 Stanford Topic Modeling Toolbox (https://cs.stanford.edu/srcf_404), Mallet (https://mimno.github.io/Mallet/ 등 다양한 툴을 활용하여 토픽모델링뿐 아니라 상세 결과 및 시각화 등을 수행할 수 있음

7) 토픽 모델링에서는 단어가 서로 독립적이지 않다는 가정(Dirichlet Distribution)에 기초하여 단어 생성 조건에 따라 사후 확률을 추론하는 원리임

- 텍스트 문서 모음이 주어지면 LDA 모델은 주제를 정의할 수 있는 클러스터 목록을 생성하고 필요 시 가중치가 적용된(weighted) 주제 목록을 구성할 수 있음

- LDA 모델링은 사전확률을 고려하므로 분류 성공률이 높고 해석이 용이하지만, 연구자의 사전지식 및 경험에 기반하여 토픽수와 토픽 명이 결정되어야 한다는 어려움이 있음

ㅇ Latent Semantic Analysis(LSA)는 비지도학습 기반으로 행렬분해를 활용하여 텍스트의 잠재적 의미를 모델링하고 문서나 단어 간의 유사도를 계산하는 방법임(Dumais, 2004)

- 각 단어들이 특정 개념을 지향하고 있다고 가정하고 분석하는 것이 특징이나, 단어의 순서를 고려하지 않으며 문맥 정보를 완전히 반영하지 못하는 한계가 있음

- LDA보다는 단어의 의미를 반영하는 알고리즘으로 구성되어 있으나, 단어의 순서 정보는 고려하지 않고 빈도만 고려한다는 특징이 있음

ㅇ Non-negative Matrix Factorization(NMF)도 비지도학습 기반으로 주어진 텍스트 데이터를 토픽과 단어의 행렬로 분해하여 의미 있는 토픽을 추출하는 방법임(Xu et al., 2003)

- 행렬 분해를 통해 원본 행렬을 양수인 토픽과 단어의 행렬로 근사화하는 과정에서 원본 데이터의 음수 값이나 불필요한 잡음 요소를 제거하고, 양수 값을 갖는 요소들을 통해 의미 있는 토픽을 도출함

- LSA와 마찬가지로 의미를 고려하여 토픽을 생성하지만, 비음수성 제약으로 인해 단어간 상대적 중요도를 더 명확하게 반영할 수 있음

ㅇ 최근 토픽 모델링은 토픽의 동적 흐름 및 변동성을 고려한 모델링 (Bogdanowicz & Guan, 2022), 텍스트 임베딩 및 딥러닝 등을 활용한 모델링 (Dieng et al., 2020; Grootendorst, 2022), 다층 신경망을 사용한 기법 등의 연구로 발전하고 있음

ㅇ 또한 정치 및 정책 관련 텍스트를 기반으로 입법 연설의 주제(Quinn et al., 2006, 2010), 상원 보도자료(Senate press releases) (Grimmer, 2010), 선거 선언문(electoral manifestos) (Menini et al., 2017) 등에 적용된 바 있으며, 최근 동향 분석 정책 변화 분석 등 다양한 분야에서 활용되고 있음

- 권민지(2019)는 5년간의 뉴스 기사에서 빈번하게 출현하는 키워드들을 활용하여 토픽모델링을 적용해 서울시의 이슈 모니터링 방안을 연구함[8)]

- 한채연 등 (2021)은 재난과 관련된 국내외 학술지에 등재된 논문을 기반으로 재난 연구 주제를 유형화하고 그 동향을 비교분석하여 국내 재난 연구의 방향성을 제안함[9)]

8) 권민지. “토픽 모델링 기반 뉴스기사 분석을 통한 서울시 이슈 도출.” 한국방송미디어공학회 학술발표대회 논문집 (2019): 11-13.

- 유재호 등(2021)은 토픽 모델링으로 녹색성장 5개년 계획(제1차~제3차)의 핵심 내용 및 키워드 등을 분석하여 제4차 녹생성장 5개년 계획 수립 기초자료로 제시함[10)]

- 곽희종(2023)은 정부에서 발표한 보도자료를 통하여 그간 추진된 정책이슈와 동향을 살펴보는 토픽 모델링 분석을 실시함[11)]

o 토픽모델링은 텍스트 데이터의 주제를 추론하고 이해하는 데 유용한 도구이며, 사전에 주제에 대한 정보가 없어도 이를 추론할 수 있다는 장점이 있으나, 도출한 주제 해석에 임의성과 주관성이 포함되며, 모델 파라미터 설정의 어려움이 존재함

o 따라서 최종적으로 도출된 결과를 어떻게 의사결정 과정에서 해석하고 인사이트를 도출하느냐는 도메인 지식과 전문성, 여러 가지 정황적 배경 등에 의존할 수밖에 없음을 유의하여 활용하는 것이 필요함

4. 이벤트 추출(Event Extraction)

o 이벤트 추출은 지정된 텍스트에서 이벤트 목록을 추출하는 방법으로, 정보 추출이라고 불리는 더 큰 자연어 처리 영역의 하위 작업이라고 할 수 있음

- 텍스트에서 추출되는 개체(entity) 간의 관계를 찾아내고, 핵심 이벤트에 대한 정보를 추출해내는 기술로 텍스트에 포함된 이벤트의 중요성을 파악하고 해석하는 데 도움을 줄 수 있음 (Xiang & Wang, 2019)

o 이벤트 추출 모델은 스탠자(Stanza), 스파이시(spaCy), CoreNLP 등이 있으며, 맞춤형 이벤트 유형 세트가 필요한 경우 라이브러리 및 모델을 활용하여 이벤트 주석이 있는 텍스트 문서 모음에서 연구자가 별도로 모델을 훈련할 수 있음

- 스탠자는 Stanford NLP 그룹에서 만든 파이썬 자연어 분석 패키지로 총 66개 언어에 대하여 사전 훈련된 모델을 제공하며, 훈련된 데이터셋에 따라 UD(Universal Dependencies) 모델과 NER 모델로 구분됨[12)]

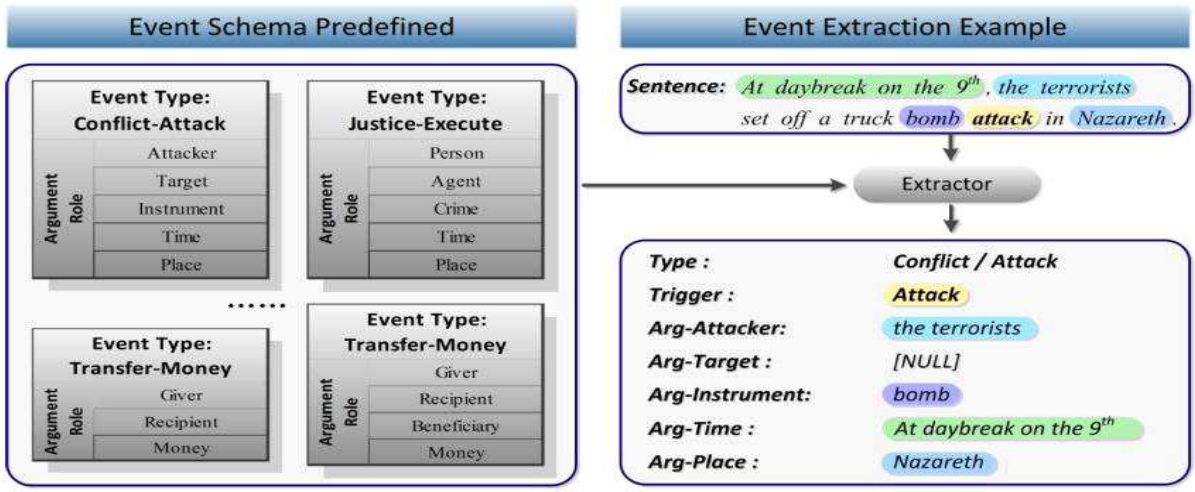
9) 한채연, 김우식, and 윤동근. “토픽모델링과 네트워크 분석을 활용한 국・내외 재난 연구 동향 분석.” 2. 한국방재학회 논문집 21.5 (2021): 79-88.

10) 유재호, 김하나, and 전의찬. “토픽모델링 기법을 활용한 녹색성장 정책 변화 분석.” 한국기후변화학회지 12.1 (2021): 67-75.

11) 곽희종. “토픽모델링을 활용한 도시재생정책 이슈 분석.” 대한국토도시계획학회지 [국토계획 58.2: 22-37.

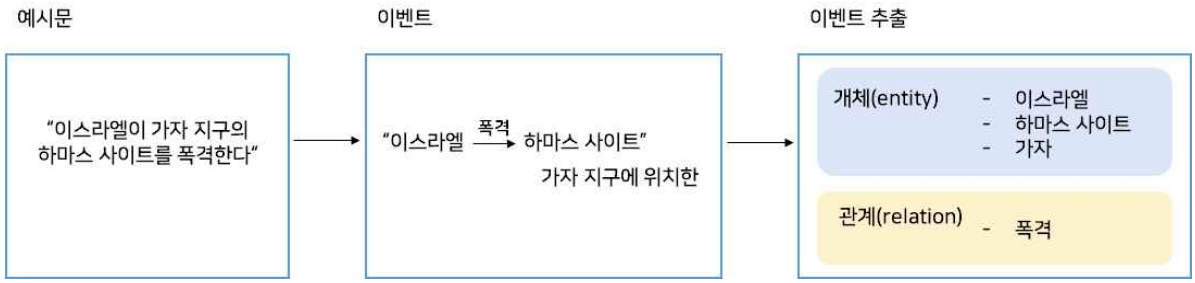
12) UD 모델은 UD 트리뱅크에서 학습하고 토큰화, 다중단어토큰(MWT) 확장, 원형 복원, 품사(POS) 및 형태학적 기능 태깅 및 종속성 구문 분석 등 기능을 포함하며, NER 모델은 8개 언어에 대하여 명명된 개체별 태깅을 지원하고 다양한 NER 데이터셋으로 훈련됨

- 스파이시는 독일 스타트업이 공개한 오픈소스 자연어 처리 라이브러리로 정보 추출 또는 자연어 이해 시스템을 구축하거나 딥러닝을 위한 텍스트 전처리 과정에 유용하며, CoreNLP는 품사 태깅, 개체명 인식에 적합함



자료: Xiang & Wang (2019)

[그림 2-3] 이벤트 추출 개념 및 원리



자료: Jin, Z., & Mihalcea, R. (2022)를 재구성하여 작성

[그림 2-4] 이벤트 추출 적용 예시

o 이벤트 추출의 최근 연구동향은 다양한 정치적 이벤트를 분석하는 코딩체계 개발, 수집된 텍스트의 유효성 평가 등으로 이루어지고 있음(McClelland, 1976; Merritt et al., 1993; Raleigh et al., 2010; Schrodt & Hall, 2006; Sundberg & Melander, 2013)

- 텍스트 데이터에서 이벤트를 식별하고, 인수를 추출하고, 서로 다른 이벤트 간 관계를 분석하는 모델을 훈련시키는 과정 등에 대한 연구가 이루어지고 있음[13)]

- 이유나 등(2022)은 개체명 인식과 이벤트 추출기법을 범죄사건 재구성에 활용하여 형사 판결문의 범죄(공소)사실에 대한 스토리라인을 시각적으로 확인할 수 있는 방법론을 제시함

13) Liu, Kang, et al.(2020) "Extracting events and their relations from texts: A survey on recent research progress and challenges." AI Open 1 (2020): 22-39.

- 국가간 분쟁, 비국가 분쟁 등을 시·공간적으로 세분화하는 이벤트 데이터 세트인 UCDP GED(UCDP Georeferenced Event Dataset)을 구축하는 등 이벤트 추출을 위한 데이터 수집 및 활용 관련 연구가 활발히 진행되고 있는 추세임[14)]

자료: Liu, Kang, et al.(2020)

[그림 2-5] 서울남부지방법원 판결문 개체명 인식 및 이벤트 추출을 통한 스토리라인 시각화[15)]

o 다만, 이벤트 추출은 언어의 다의성에 대한 분석 한계로 인해 오류가 발생할 수 있다는 치명적인 장애요인이 있음

- 문장 수준의 이벤트 추출과 달리 문서 수준의 이벤트 추출에서는 하나의 이벤트에 대한 인수가 하나의 문서에서 여러 문장으로 흩어지는 Arguments-scattering이 나타날 수 있어 이벤트 정보 간의 장거리(심지어 문장 간) 종속성을 캡처해야 하며, 문서에 혼합된 여러 이벤트가 있으면 문서 수준 이벤트 추출 시스템이 이를 구별하고 서로 다른 이벤트에 대한 해당 인수를 분할해야 하는 등의 어려움이 있음

- 최신 이벤트 추출 모델은 지도 학습 설정을 기반으로 하는 경우가 많은데 레이블이 지정된 학습 데이터는 이벤트 유형의 범위가 적고 크기가 제한되어 생성하는 데 비용이 많이 들기 때문에 대규모 이벤트를 추출하기 어렵다는 이슈가 있음

o 이벤트 추출 방법을 통해 정책 관련 이벤트를 정확히 식별하고 추출하기 위해서는 동일한 이벤트에 대해 일관된 추출 결과를 얻을 수 있도록 일관성을 확보해야 하며, 다양한 유형의 이벤트를 구조화하여 식별하고 추출할 수 있어야 함

14) Sundberg, Ralph, & Erik Melander. "Introducing the UCDP georeferenced event dataset." Journal of peace research 50.4 (2013): 523-532.

15) 이유나, 박성미, & 박노섭. "개체명 인식과 이벤트 추출을 통한판결문 범죄사실 구성요소 및 스토리라인시각화방안 연구." 한국정보처리학회 학술대회논문집 29.2 (2022): 490-492.

5. 점수 예측(Score Prediction)[16]

o 점수 예측은 주어진 텍스트의 특징을 토대로 점수를 예측하는 작업으로, 주로 정치 텍스트 스케일링과 같은 분야에서 텍스트에 대한 다양한 측면을 수치화하여 점수로 제시(Gennaro, Gloria and Elliott, 2022)[17]

- 정치 연설, 선언문, 소셜 미디어 등 주어진 광범위한 텍스트에 대해 좌우 이념, 감성, 정치적 통합 과정 또는 정책 등에 대한 다양한 태도를 예측하는 것을 목표로 함

o 원리는 데이터에 기반하여 특징을 추출하고, 예측 모델 훈련을 통해 새로운 텍스트의 기준별 점수를 수치화하여 예측하는 것임

- 예측된 점수는 텍스트 또는 문서의 특징을 요약하여 제공하는 중요 정보를 제공하며, 사전에 정의된 스케일에 따라 문서의 특징을 개괄적으로 파악할 수 있음

o 텍스트 스케일링을 위한 전통적인 모델로는 Wordscore와 WordFish가 있으며, 최근에는 텍스트를 고차원 벡터로 표현하여 신경망을 통해 점수를 예측하는 방식으로 학습하는 방식을 활용하고 있는 추세

- Wordscore 알고리즘은 데이터로 단어에 대응 분석을 적용한 것으로 볼 수 있으며, 텍스트를 이해하고 해석해야 할 담론이 아니라 단어 형태의 데이터로 취급하는 정치적 텍스트에서 정책 입장을 추출하는 새로운 방식을 제시[18]

- WordFish는 텍스트의 단어 빈도를 기반으로 정책 위치를 추정하기 위한 스케일링 알고리즘으로 시계열 파티 위치를 추정하는데 주로 활용[19][20]

- InstructGPT와 같은 기성 범용 모델을 적용하고 API에 스케일링 유형을 지정하는 프롬프트를 설계하거나 사전 훈련된 모델을 활용하여 유사하게 척도를 모방하여 예측하는 방법으로 모델을 학습하고 활용하는 것이 가능[21][22]

16) 점수 예측은 텍스트 분류와 동일한 라이브러리를 활용함. Scikit-learn은 파이썬 기반의 머신러닝 라이브러리로, 점수 예측 모델링에 널리 사용함. 선형 회귀, 로지스틱 회귀, 서포트 벡터 머신 등 다양한 알고리즘을 제공하며, 특징 선택, 교차 검증, 모델 평가 등의 기능을 포함하고 있음. Hugging Face Transformers(https://huggingface.co/)는 자연어 처리를 위한 라이브러리로 , 최신의 사전 훈련된 언어 모델을 활용하여 점수 예측 모델을 구축할 수 있으며, 사전 훈련된 모델을 활용하면 더욱 정확한 예측을 수행할 수 있다는 장점이 있음

17) Gennaro, Gloria, and Elliott Ash. (2022) "Emotion and reason in political language." The Economic Journal 132.643 (2022): 1037-1059.

18) Laver, Michael, Kenneth Benoit, and John Garry. "Extracting policy positions from political texts using words as data." American political science review 97.2 (2003): 311-331.

19) Slapin, Jonathan B., and Sven-Oliver Proksch. "A scaling model for estimating time-series party positions from texts." American Journal of Political Science 52.3 (2008): 705-722.

20) Nanni, Federico, et al. "Political text scaling meets computational semantics." ACM/IMS Transactions on Data Science (TDS) 2.4 (2022): 1-27.

21) Ouyang, Long, et al. (2022) "Training language models to follow instructions with human feedback." Advances in Neural Information Processing Systems 35 (2022): 27730-27744.

o 점수 예측은 정치 언어의 감정성을 측정하는 데에 활용하는 연구가 대부분이며, 이념, 특정 이슈에 대한 긍·부정 등의 특징을 점수를 통해 유추할 수 있도록 정보를 추출하는 연구가 주로 수행되고 있음

- 선행연구에서는 정치 관련 텍스트 데이터에 대하여 좌우 입장에 대한 정도를 스케일링함 Slapin & Proksch (2008)는 정치 관련 테스트 데이터에 대하여 좌우 입장에 대한 정도를 스케일링함

- Huang and Luk (2020)는 19년 간의 중국 신문을 토대로 경제 정책 불확실성에 대한 월간 지수를 개발하여 텍스트 스케일링 방안을 제시함

- Grimmer and Stewart (2013)는 신문 기사 내 정치 관련 텍스트에 대한 어조가 긍정적인지 부정적인지에 대한 정도를 스케일링 함

- Gennaro et al.(2022)은 1858년부터 2014년까지 미국 의회에서 열린 600만 개의 연설의 감정을 스케일링하고, 연설 주제, 정치인의 개인적 특성과 제도적 요인에 따른 감정성, 당파적 양극화와 감정성의 관계 등을 분석함

o 점수 예측은 텍스트에 대한 주관적인 해석을 넘어서 객관적이고 자동화 가능한 측정 지표를 통해 비교 분석이 가능하다는 장점이 있으나, 다른 텍스트 분석 및 자연어 처리와 마찬가지로 데이터에 대한 의존성, 모델의 복잡성 및 해석의 한계점이 존재함

o 따라서 데이터셋 및 활용 목적에 따라 텍스트 수집, 전처리, 및 점수화 분석 등 전체 프로세스에 대한 구성 및 상세 절차를 차별화하여 적용하는 것이 필요함

- 연구 시 연구 질문(Research Question)에 따라 수치화 대상이 주체, 특정 대상, 사건, 정치적 입장, 혹은 경향 등 다양한 형태로 정의하게 됨. 분석 대상으로 설정한 스케일링 타겟의 특성을 고려하여 방법론을 제시하는 것이 중요함

- 또한, 정치 텍스트 스케일링은 텍스트에 대하여 점수화 하고자 하는 유형에 대하여 사전에 정의해야 함. 이 때 유형의 구분 및 기준을 명확하게 확보해야 분석된 결과에 대한 신뢰도를 확보할 수 있음. 학습 모델을 적용할 경우 텍스트의 언어 다국성 등 자연어처리 이슈를 검토할 필요가 있음(Grimmer J. & Stewart BM., 2013)[22)]

22) Grimmer J, Stewart BM. (2013) Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis. 2013;21(3):267-297. doi:10.1093/pan/mps028

제2절 자연어 처리의 공공부문 활용

○ 근거 기반 정책 결정을 위한 데이터 분석, 국민과의 정책 커뮤니케이션 개선, 정책 효과 조사, 정치 현상 해석 등 정책 입안 과정에서 자연어 처리를 활용할 수 있음

- 정치 영역에는 의회 토론, 연설, 입법 텍스트, 정당 데이터베이스, 전문가 조사 데이터 등 분석 가능한 텍스트 데이터가 대량으로 존재하는데 자연어 처리를 통해 정치과정에서 산출되는 방대한 텍스트 데이터를 분석할 수 있음

자료: Jin, Z., & Mihalcea, R. (2022)를 재구성하여 작성

[그림 2-6] 정책 과정에서의 NLP 활용 개요[23)]

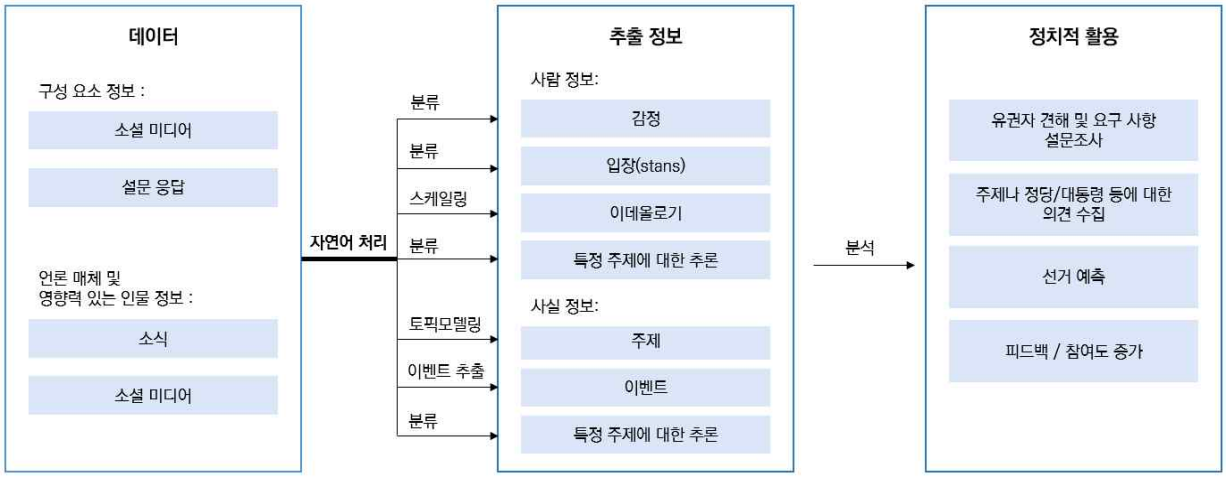
1. 정책수립

○ 연구자 또는 정책 입안자는 데이터를 분석하여 증거기반 정책을 수립하는 데 기반이 되는 핵심 정보를 추출해낼 수 있음

- 자연어 처리의 주요 용도는 많은 텍스트 모음에서 정보를 추출하는 것으로, 정책 대상자 등의 견해와 요구를 분석하므로써 정책 입안자들의 의사결정을 지원할 수 있음

- 일반적으로 빅데이터 기반 분석을 의사결정자가 시민과 사회로부터 더 많은 피드백을 수집할 수 있도록 지원하여 정책 입안자가 시민과 더 가까워지고 투명성과 정치적 문제에 대한 참여를 높일 수 있도록 함[24)]

23) Jin, Z., & Mihalcea, R. (2022). Natural Language Processing for Policymaking. In Handbook of Computational Social Science for Policy (pp. 141-162). Cham: Springer International Publishing.



자료: Jin, Z., & Mihalcea, R. (2022)를 재구성하여 작성

[그림 2-7] 증거 기반 정책 수립을 위한 데이터 분석용 자연어 처리[25)]

○ Hiware *et al.* (2020)은 재난 후 구호 조정 노력을 지원하기 위해 소셜미디어 게시물에서 크라우드 소싱 정보를 활용하는 반자동 플랫폼인 NARMADA를 제시[26)]

- 트위터와 같은 소셜 미디어 사이트의 가능한 텍스트 데이터를 수집하여 자동 분류할 수 있는 모니터링 모델을 개발하고, 카테고리 분류 프로세스를 자동화하여 여론 자동분류 시스템을 구축함

- 다만, 소셜 미디어를 통해 정책수요나 여론을 팡가하는 연구는 소셜 미디어 데이터 자체의 신뢰성에 따라 결과의 오류가 발생할 수 있기 때문에 활용 가능성과 진위성에 대한 추가적인 검토를 필요로 함

Tweet text (excerpts)	Resource	Location	Quantity	Source	Contact
Urgent need of analgesic,antibiotics, betadiene, swabs in kathmandu!! Call for help 98XXX-XXXXX #earthquake #Nepal #KTM (N)	analgesic, antibiotics, betadiene, swabs	kathmandu, ktm, nepal			98XXX-XXXXX
India sends 39 #NDRF team, 2 dogs and 3 tonnes equipment to Nepal Army for rescue operations: Indian Embassy in #Nepal (A)	NDRF team, dogs,	nepal	dogs - 2, NDRF team - 39	India	
Visiting Sindhupalchok devastating earthquake highly affected district . Delivery Women in a tent . No water no toilet (N)	tent, delivery women, water	Sindhupalchok			
Rajasthan Seva Samiti donates more than 800 tents to Nepal Earthquake victims (A)	tents		tents-800	Rajasthan Seva Samiti	

자료: Jin, Z., & Mihalcea, R. (2022)를 재구성하여 작성

[그림 2-8] Hiware el al(2020) 연구에서 제시하는 트위터 필요(N) 및 가용(A) 정보의 예[30)]

24) Arunachalam, Ravi, and Sandipan Sarkar. "The new eye of government: citizen sentiment analysis in social media." Proceedings of the IJCNLP 2013 workshop on natural language processing for social media (SocialNLP). 2013.

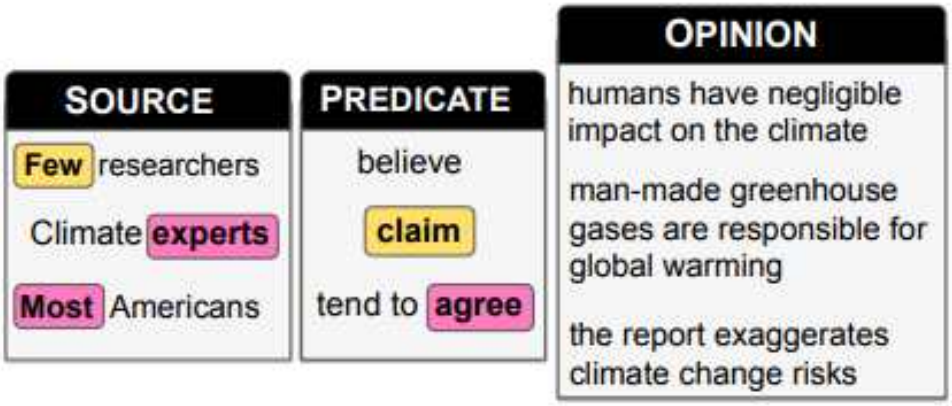
25) Jin, Z., & Mihalcea, R. (2022). Natural Language Processing for Policymaking. In Handbook of Computational Social Science for Policy (pp. 141-162). Cham: Springer International Publishing.

26) Hiware, Kaustubh, et al. "NARMADA: Need and available resource managing assistant for disasters and adversities." arXiv preprint arXiv:2005.13524 (2020).

ㅇ Luo *et al.*(2020)은 지구 온난화 관련 텍스트 데이터셋인 GWSD(Global Warming Stance Dataset)를 구축하여 지구 온난화 논쟁에서의 의견 프레이밍을 연구

- 지구 온난화 관련 입장을 연구하기 위해 뉴스 미디어에서 스탠스 레이블이 지정된 2,000개의 문장으로 GWSD를 구축하고, 500,000개 의견의 입장을 예측하기 위해 가중 BERT 모델을 훈련시켜 분석

- 분석 결과를 토대로 지구 온난화에 회의적인 입장을 가진 미디어는 더 많은 의구심을 제기하거나 반대 의견을 제기한 것으로 제시



자료: Luo, Yiwei, Dallas Card, and Dan Jurafsky(2020)

[그림 2-9] OPINION 구성 요소 및 구성 요소 내에서 프레이밍 장치를 확인하고 의심하는 예[27)]

- 본 연구는 기존 데이터셋 및 어휘집의 유용성을 제시하였을 뿐만 아니라, 이를 공개하여 후속 연구 및 텍스트 마이닝 작업 등에 시사하는 바가 있음

- 또한 연구를 통해 문장 수준에서 많은 항목이 모호할 수 있으며, 정당 가입 및 성별과 같은 인구 통계학적 특성이 사람들이 반응하는 방식에 영향을 미칠 수 있음을 제시하고 자연어 처리 모델 활용 시 훈련 데이터의 출처에 따라 모델이 결과적으로 어떻게 편중될 수 있는지를 연구자나 활용자가 인지해야 함을 강조

2. 정책해석

ㅇ 정책이 입안된 후, 정치학자들과 사회과학자들은 정치적 결정을 해석(interpret)하거나, 의제를 발굴하거나 정책 대응을 파악하기 위해 텍스트 데이터를 분석할 수 있음

- 정책의제 발굴은 텍스트 데이터를 활용하여 정책 입안자가 우선순위를 두는 주제, 정치적 사건 및 특정 주제에 대한 다양한 정책 행위자의 입장을 포함한 정책 의제를 추론할 수

27) Luo, Yiwei, Dallas Card, and Dan Jurafsky. "Detecting stance in media on global warming." arXiv preprint arXiv:2010.15149 (2020).

있으며, 그 과정에서 보도자료, 법률 및 선거 캠페인 자료, 여론조사 자료 등을 활용할 수 있음

- 정책 대응을 파악하는 것은 여론의 변화가 공공정책의 대응으로 이어지는 방식이나 다양한 요인에 대하여 정치나 정책이 대응하는 방식에 대한 연구로 발전할 수 있음

자료: Jin, Z., & Mihalcea, R. (2022)를 재구성하여 작성

[그림 2-10] 정치적 결정을 해석하는 자연어 처리 방법[28)]

ㅇ 정치 또는 정책적 주제와 우선순위를 분석하기 위한 연구로 좌우 이데올로기나 포퓰리즘의 정도 등 정치적 태도와 선호도를 잠재 공간의 입장으로 개념화한 정치 공간 모델 연구가 수행된 바 있음[29)]

- 단어 자체가 아닌 단어의 의미론적 표현을 기반으로 매우 다른 정치적 입장을 어휘화할 수 있게 하는 스케일링 알고리즘인 SemScale을 이용하였으며, 유럽 의회 웹사이트의 연설문을 수집하여 분석을 실시

- 특정 어휘 및 의미 정보를 제어하는 것이 강력한 예측으로 이어진다는 분석 결과가 도출되었으나, 텍스트가 긴 기간에 걸쳐 있을 경우 의미론적 스케일링은 적합하지 않다는 제한사항을 염두해야 함

ㅇ Jin *et al.* (2021)은 미국 주지사를 주제로 한 1,000만 개 이상의 COVID-19 관련 트위터 의견을 분석하여 주지사의 정치적 결정을 분석함[30)]

- 대중의 정서를 얻기 위해 분류 모델을 사용하여 대중의 정서가 어떻게 미국 주지사가 만든 COVID-19 정책의 정치적 결정으로 이어지는지 연구

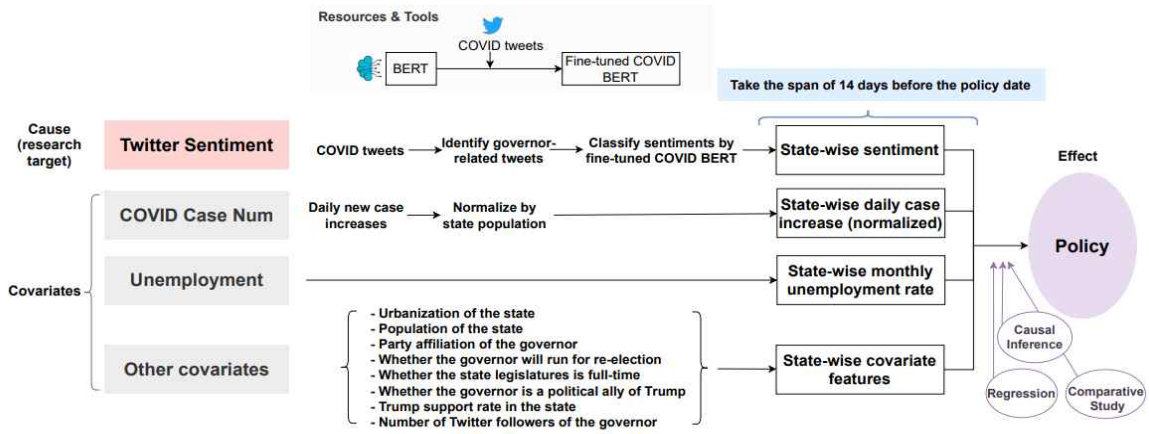
- 정책을 0(가장 느슨)부터 5(가장 엄격) 등급으로 나누어 미국 50개 주별로 COVID-19와 관련된 사회적 거리두기 정책을 설명하고, 이를 위해 관련 행정명령 데이터를 활용

28) Jin, Z., & Mihalcea, R. (2022). Natural Language Processing for Policymaking. In Handbook of Computational Social Science for Policy (pp. 141-162). Cham: Springer International Publishing.

29) Nanni, Federico, et al. (2022) "Political text scaling meets computational semantics." ACM/IMS Transactions on Data Science (TDS) 2.4 (2022): 1-27.

30) Jin, Zhijing, et al. "Mining the cause of political decision-making from social media: A case study of COVID-19 policies across the US states." Findings of the Association for Computational Linguistics: EMNLP 2021. 2021.

- 자연어 처리 방법을 통해 정책 관련 데이터를 조작적 정의하여 변수화하고, 실업률 및 COVID-19 감염자 수 등을 통제요인으로 추가하여 분석 진행
- 또한 소셜 미디어(트위터) 데이터를 BERT 모델을 활용해 분류하여 소셜 미디어 내 COVID-19 관련 정서와 정책 간 인과적 관계에 대한 다중회귀분석을 수행하여 여론이 정책 결정에 어떤 영향을 미치는지 설명
- 해당 연구는 설문에 근거한 정책 대응성(정치적 의사결정의 원인 분석) 연구의 패러다임을 바꾼 연구라고 할 수 있으며, 단기적인 결정으로 구성되는 정치 현상을 분석하고자 하는 연구에 자연어 처리 방법의 유용성을 시사함

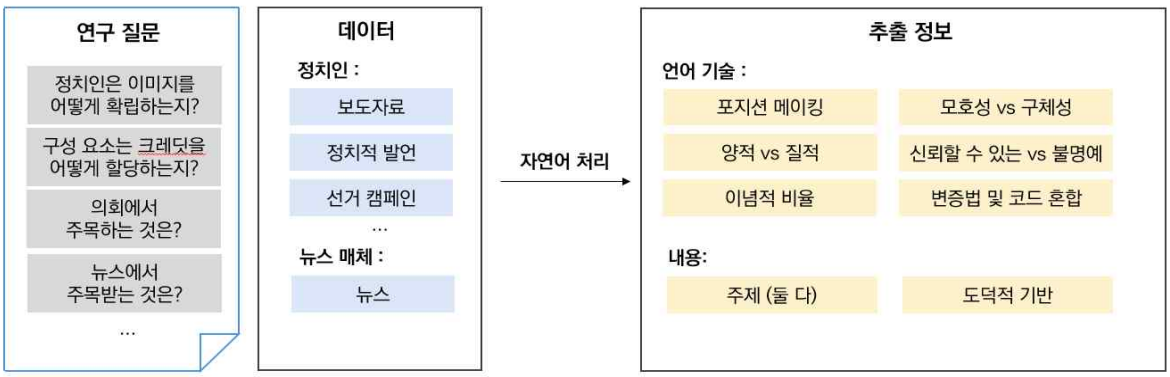


자료: Jin, Z., & Mihalcea, R. (2022)를 재구성하여 작성

[그림 2-11] 국가별 COVID 정책을 예측하기 위한 시스템의 데이터 수집 파이프라인 및 아키텍처[34)]

3. 정책소통

- 정책 소통은 주로 정책 입안자 또는 의사결정자가 어떻게 정책을 국민들에게 제시하는지 이해하기 위한 연구로, 정치인들이 어떻게 그들의 이미지를 확립하는지에 대한 연구가 이루어지고 있음



자료: Grimmer, Justin, (2013)를 재구성하여 작성

[그림 2-12] 정책 소통 분석을 위한 자연어 처리[31)]

- 정책소통의 언어 기술적 요소에 관한 연구는 주로 정치인들이 사용하는 언어 유형에 초점을 맞춤

- 정책소통의 내용적 측면에 관한 연구는 상원의원들이 원내 성명에서 논의하는 내용과 대통령들이 일상 연설에서 다루는 내용, 정치인들이 정치 트윗의 기초로 사용하는 도덕적 기반과 같은 정치적 성명의 주제를 포함하고 있음

- 구체적으로 언론이 특정 정치인의 정치적 메시지를 얼마나 자주 다루는지, 정책적 우선순위에 어떤 영향을 미치는지 등 정책과 정치인의 메시지 간 상호작용을 분석하는 등의 방법으로 텍스트 데이터 분석을 활용할 수 있음

- 기존 정치 텍스트의 언어를 분석하는 것 외에 사회 전반에 더 유익한 미래를 향하도록 정책 소통을 개선하는 방법도 중요하나 이에 대한 연구는 상대적으로 거의 없어 향후 연구의 확대가 필요

ㅇ Grimmer (2013)는 정치 텍스트의 어떤 부분이 입장 표명과 신용 주장인지를 분석하는 연구를 수행하여 집단 대표가 정책 토론 중에 정당이 표현하는 입장의 품질에 부정적인 영향을 미칠 수 있음을 보임[32)]

- 보도 자료의 토픽모델링을 통해 주제 레이블 라벨링을 실시하고, 상원 의원의 스타일 및 우선순위에 대해 2차원적 요약을 통해 각 상원의원에 대한 신용 청구 및 직책 취직에 할애하는 비율을 생성하여, 상원 의원이 매년 신용 주장과 직책 채택의 균형을 맞추는 방법과 정책 토론에 미치는 영향을 모델링 함

- 새로운 연구 의제의 필요성과 스타일을 도입함으로써 이전에는 불가능했던 가설 검정을 실시하였으나 입법자들이 어떻게 유권자들에게 그들의 정책 작업을 제시하는지 정보가 적어 추가 데이터 확보가 필요

ㅇ Kristen Johnson and Dan Goldwasser (2018)는 정치가가 문제에 대한 자신의 입장을 표현하는 데 사용하는 도덕적 기반을 예측하기 위해 언어와 정치인이 Twitter에서 문제를 구성하는 방법을 모델링함[33)]

31) Jin, Z., & Mihalcea, R. (2022). Natural Language Processing for Policymaking. In Handbook of Computational Social Science for Policy (pp. 141-162). Cham: Springer International Publishing.

32) Grimmer, Justin. "Appropriators not position takers: The distorting effects of electoral incentives on congressional representation." American Journal of Political Science 57.3 (2013): 624-642.

33) Johnson, Kristen, and Dan Goldwasser. "Classification of moral foundations in microblog political discourse." Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers). 2018.

- 미국 트위터에 정치인의 도덕적 기반을 분류하기 위한 공동 모델링 언어 및 정치적 프레이밍 기술을 탐색하고, 주석 지침에 대한 설명과 주석이 달린 2,050개의 데이터셋을 구축하여 트위터에 존재하는 도덕적 기반의 분류를 위한 문제에 쉽게 적응할 수 있는 계산 모델을 제안함

- 도덕적 기반과 정책 프레임에 대한 모델 개발의 초기 접근 방식을 제공하였으나, 트위터에는 충분한 정보가 제공되지 않아 보다 상세한 이데올로기적 성향 및 입장 예측 분석을 위한 프레임 개발을 위해서는 추가 작업이 필요

4. 정책평가

ㅇ 정책평가에 관한 연구는 정책 적용 이후 여론이나 피드백 데이터를 수집하거나, 정책 효과를 평가하기 위하여 민심의 변화, 경제적 상황의 변화 등에 대한 텍스트 데이터를 수집 분석하는 연구가 이뤄지고 있음

- 부가적으로 정책수립 시와 비슷한 패러다임을 가지고 정책 전후의 정서 분류 결과를 비교하여 분석할 수 있어 활발하게 진행되고 있음

ㅇ Calvo-González *et al.* (2018)은 장기적인 경제 성장에 해를 끼치는 정책 변동성의 부정적인 결과를 조사함[34)]

- 정책 변동성을 측정하기 위해 먼저 대통령 연설에 대한 주제 모델링을 통해 주요 주제를 얻은 후 주제의 중요성이 시간이 지남에 따라 어떻게 변화하는지 분석을 실시

- 라틴 아메리카 10개국과 스페인의 953개 대통령 연설의 새로운 데이터 세트에 LDA알고리즘을 사용하여 정책 변동성을 측정하였으며, 1940~2010년의 장기 성장과 대통령 연설에서 전달되는 우선 순위의 정보 내용을 활용하는 정책 변동성에 대한 프록시가 음의 상관관계가 있음을 보임

- 데이터로 텍스트를 활용하는 것이 경제적, 정치적 이해를 심화시키는데 도움이 될 수 있음을 보여줬으나 특정 연도의 연설이 누락된 국가의 수가 다수 있었으며, 원본의 품질이 제대로 디지털화하기에 너무 낮은 연설 등도 존재하여 데이터 품질 및 확보의 중요성이 강조됨

34) Calvo-González, Oscar, Axel Eizmendi, and Germán Reyes. "Winners never quit, quitters never grow: Using text mining to measure policy volatility and its link with long-term growth in latin America." World Bank Policy Research Working Paper 8310 (2018).

제3절 데이터·인공지능 기반 국가R&D 프로세스 개선 사례

1. 개요

ㅇ 공공데이터 활용의 중요성이 강조됨에 따라, 과학기술정책 기획, 국가R&D 투자 및 사업관리 등 전주기 연구개발시스템 개선에 데이터와 인공지능을 활용한 연구가 활발해지고 있는 추세임

ㅇ 본 연구의 취지 또한 평가 산출물과 인공지능을 활용하여 사업평가 업무 프로세스를 개선하고 피평가 부처를 포함한 이해관계자에게 효율적으로 정보를 제공하고자 함임

ㅇ 이에 기존 연구개발정보를 활용하거나 인공지능 모델링을 통해 업무 프로세스를 개선하거나 효율화한 사례들을 분석하여 AI 기반 R&D사업 평가지원체계 마련의 가능성과 이슈를 발굴하고 시사점을 도출하고자 함

ㅇ 분석대상 사례는 NTIS 성과평가정보시스템, 차별성 검토 서비스, 과학기술 표준분류 추천 서비스, KISTEP 국가연구개발사업 및 과제 데이터 지능형 검색 서비스 등 총 4가지임

- 첫 번째로, NTIS 성과평가정보시스템은 그동안 문서로 쌓여왔던 평가 산출물들을 데이터화하여 업로드하고, 관리하는 시스템으로 문서와 개별 파일로만 관리되던 성과평가 데이터의 체계적인 관리와 향후 활용의 기반이 되는 시스템임

- 두 번째로, NTIS 차별성검토 서비스는 API 연계를 통해 NTIS 내 쌓여있는 연구개발과제정보를 참조하여 과제정보의 입력 규칙을 학습하고 코사인 및 유클라디안 유사도 기반 타 과제와의 유사도를 산출해주는 방식으로 차별성 검토 프로세스를 지원함

- 세 번째로, NTIS 과학기술표준분류추천 서비스는 합성곱신경망 기법(TK_CNN)을 활용해 연구보고서 메타 데이터에 자주 등장하는 키워드르 범주화하여 학습 후 이에 따라 새로운 과제의 과학기술표준분류를 추천해주는 방식으로 연구과제 신청 및 연구계획 수립 프로세스를 지원함

2. NTIS 성과평가정보시스템[35)]

ㅇ 사업·기관 전주기 성과평가정보를 체계적으로 수집·관리·공개하고 평가 업무를 효율적으로 지원하기 위해 R&D PEIS(Performance & Evaluation Information System) 구축

- 연구성과평가법(2021. 12. 28., 개정, 2022. 6. 29., 시행) 개정에 따라 신설된 제22조(성과평가정보의 공개 등)에 근거하여 성과평가 통합관리시스템을 구축·운영함

35) KISTEP 국가연구개발 성과평가 정책 수립 및 성과평가 실시 최종 보고서(2021년, 2022년)의 내용을 재구성하여 작성

자료: NTIS 성과평가정보시스템(https://www.ntis.go.kr/rndeval/prog/othbc/intro/bzEvalInfoManageSysView.do, 최종접속 : 24.2.2.)

[그림 2-13] 성과평가 통합관리시스템 기능 및 기대 효과

(1) 데이터

o 성과평가정보시스템 상 국가연구개발사업 평가 전략계획서 입력 항목은 사업개요, 성과목표·지표, 지식재산권 창출 활동, 사업평가계획으로 구분됨

- 사업개요, 성과목표지표, 지식재산권 창출 활동, 사업평가계획은 아래의 [그림 2-15]와 같이 웹에서 데이터를 입력하며 데이터셋 형태로 다운로드 가능함

자료: NTIS 성과평가정보시스템(https://www.ntis.go.kr/rndeval/prog/ministry/writng/bzEvalWritngStep01Sn0Form.do, 최종접속 : 24.2.2.)

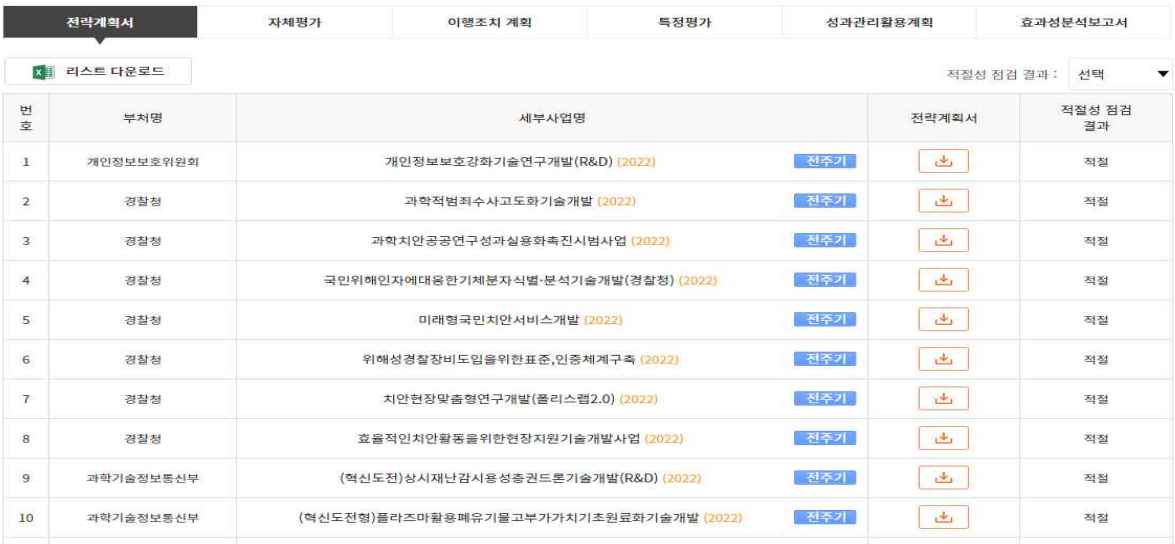
[그림 2-14] PEIS 전략계획서 입력 화면 예시

o 부처 및 성과목표지표 정보를 대시보드에 제시하여 예산, 단위사업 수, 세부사업 수 등 최근 3년간의 데이터 추이를 확인할 수 있음

자료: NTIS 성과평가정보시스템(https://www.ntis.go.kr/rndeval/prog/othbc/intro/bzEvalInfoManageSysView.do, 최종접속 : 24.2.2.)

[그림 2-15] 2022년도 사업평가 정보 대시보드

o 전략계획서, 자체평가, 이행조치 계획, 특정평가, 성과관리활용계획, 효과성분석보고서의 사업정보, 성과목표지표, 평가이력, 대표성과 및 사례를 데이터로 관리하고 있으며 필요시 원본 pdf 파일을 다운로드 하여 확인할 수 있도록 시스템 구축



번호	부처명	세부사업명		전략계획서	적절성 점검 결과
1	개인정보보호위원회	개인정보보호강화기술연구개발(R&D) (2022)	전주기		적절
2	경찰청	과학적범죄수사고도화기술개발 (2022)	전주기		적절
3	경찰청	과학치안공공연구성과실용화촉진시범사업 (2022)	전주기		적절
4	경찰청	국민위해인자에대응한기체분자식별·분석기술개발(경찰청) (2022)	전주기		적절
5	경찰청	미래형국민치안서비스개발 (2022)	전주기		적절
6	경찰청	위해성경찰장비도입을위한표준,인증체계구축 (2022)	전주기		적절
7	경찰청	치안현장맞춤형연구개발(폴리스랩2.0) (2022)	전주기		적절
8	경찰청	효율적인치안활동을위한현장지원기술개발사업 (2022)	전주기		적절
9	과학기술정보통신부	(혁신도전)상시재난감시용성층권드론기술개발(R&D) (2022)	전주기		적절
10	과학기술정보통신부	(혁신도전형)플라즈마활용폐유기물고부가가치기초원료화기술개발 (2022)	전주기		적절

자료: NTIS 성과평가정보시스템(https://www.ntis.go.kr/rndeval/prog/othbc/intro/bzEvalInfoManageSysView.do, 최종접속 : 24.2.2.)

[그림 2-16] PEIS 전략계획서 제공 현황(2022년 예시)

(2) 프로세스

- 사업 전 주기(기획-수행-종료-활용)에 걸친 성과평가정보를 부처 및 연구회가 등록·공개하고 과학기술혁신본부가 점검/모니터링
- R&D PEIS는 입력단계-관리단계-결과(공개) 단계로 구분하여 설계·운영되며, 총 4등급의 권한(부처&기관-부처총괄·연구회-과기혁신본부&KISTEP-시스템 설계자)으로 구성됨

[표 2-2] 성과평가 통합관리시스템 프로세스

평가 종류	단계	자료 목록	세부 내용	역할	
				부처/기관	혁신본부
사업평가	기획	사업기획보고서	• 보고서 전체를 평가자가 열람할 수 있게 파일로 등록	작성·등록	관리
		사업전략계획서	• 사업전략계획서 파일 등록 • 연차별 성과목표/지표는 모니터링을 위한 별도 프레임으로 공개 • 평가계획은 사업별 평가주기를 사업 현황정보에 구현		점검
	수행	사업추진현황 (연차별 실적)	• 각 사업의 성과지표 연차별 실적을 사업 현황정보로 구현 • 성과목표/지표명, 연차별 목표치는 사업전략계획서 수립 시 등록, 연차별 실적은 매년 3월까지 직전년도 실적을 등록		모니터링
		평가결과 (중간평가, 특정평가)	• 각 평가별 보고서 등록·공개 및 평가결과를 사업현황 정보에 공개		모니터링
	종료	성과활용계획서	• 성과활용계획서, 대표 성과 등 등록·공개		점검
	활용	효과성분석보고서	• 효과성분석보고서 등록·공개		점검
기관평가	기관운영계획서		• 계획서를 등록하고 점검-공개		점검
	기관운영평가		• 자체평가/상위평가 과정-결과 등록 및 공개		점검
	연구사업계획서		• 계획서를 등록하고 점검-공개		점검
	연구사업평가		• 자체평가/상위평가 과정-결과 등록 및 공개		점검

자료: NTIS 성과평가정보시스템(https://www.ntis.go.kr/rndeval/prog/othbc/intro/bzEvalInfoManageSysView.do, 최종접속 : 24.2.2.)

(3) 특징 및 시사점

- 성과평가정보시스템은 그동안 개별 파일이나 문서로 쌓여있어 오프라인으로 관리되던 데이터를 직접입력 및 업로드 방식으로 체계적인 관리가 가능하다는 장점이 있으며, 평가 담당자가 입력된 정보와 결과를 상시 확인할 수 있다는 점에서 업무 프로세스의 투명성을 높였다는 데 의의가 있음
- 다만, 국가과학기술지식정보서비스(NTIS)에 데이터를 수동으로 입력하여, 데이터 구축 단계에서 발생하는 오류가 있을 수 있음

o 간단한 인공지능 기술을 활용하여 국가 R&D정보에 대한 요약정보를 생성하여 제공하는 서비스를 제공하고 있으나, 대시보드 상에서 사업 수, 예산 총액 등 간단한 정보만 확인할 수 있고 부처 간 또는 관련사업 간의 데이터 비교에는 한계가 있음

3. NTIS 차별성 검토 서비스

o 연구개발과제 선정 시 전문기관 등의 과제 차별성 검토를 위한 참고자료를 제공하고자 하는 목적으로 축적된 연구개발과제 데이터를 토대로 유사성을 분석할 수 있는 가장 오래된 서비스라고 할 수 있음

(1) 데이터

o 데이터 기반 R&D 사업 유사도 검토를 위해, 내역사업 수준에서 정부 연구개발 사업의 특성을 나타내는 내역사업 profile DB를 구축하였으며, 데이터 마이닝의 다양한 기법 중 Lucene[36] 방법을 기반으로 한 Dataclustering algorithm 개발을 통해 해당 DB를 바탕으로 정부 연구개발사업간 유사도를 검토 할 수 있는 시스템을 구축(홍세호, 2013)[37]

- (과제단위의 유사분석 시스템) 문서 내 항목(field)에서 전체에서 추출된 키워드 대비 매칭된 키워드 수에 대한 비중산정 및 가중치 부여 후 이를 각 항목의 합으로 계산[38]

- (과제의 과학기술표준분류를 이용한 사업간 유사분석) 사업 내 과학기술표준분류의 출현여부와 빈도를 희소행렬(sparse matrix)[39]로 구성하여 코사인 기법[40]과 유클리디안 기법[41]으로 사업간 유사도를 측정

o 최소 글자수, 특수문자 사용, 동일문구 반복 기준 등으로 이루어진 과제정보 입력 규칙을 참조하여 엑셀 또는 웹브라우저 상에서 데이터를 업로드하여 대상과제를 등록

- 기준차별성, 기준연도, 연구과제명, 과제공개여부, 연구책임자명, 과제관리기관명, 연구목표, 연구내용, 기대효과, 한글 및 영문 키워드 등을 입력

36) Lucene은 텍스트 검색 및 정보 검색을 위한 오픈 소스 검색 엔진 라이브러리이며, 데이터 마이닝과는 조금 다르게, Lucene은 주로 텍스트 기반의 정보를 효과적으로 검색하기 위해 사용됨, https://lucene.apache.org/

37) 홍세호, 국가연구개발사업 유사·중복 검색 시스템 개발을 위한 실증연구, 2013

38) 유사도(%)를 기반으로 한 과제 단위 유사분석 서비스로 운영되었으나, 2023년 10월 이후 차별성 검토로 이름 변경

39) 희소행렬(sparse matrix)은 대부분의 요소가 0인 행렬을 의미하며, 자연어 처리 등의 분야에서는 대량의 데이터 중에서 대부분이 0인 경우가 많기 때문에 희소행렬을 사용하여 메모리를 효율적으로 관리할 수 있음

40) 두 벡터 간의 각도를 이용하여 유사성을 측정

41) 두 벡터 간의 직선 거리를 이용하여 유사성을 측정

- Open-API를 활용하여 논문, 특허, 연구보고서 등 개별 성과정보의 조회 기능 제공[42] (KISTI, 2014)

REST(URI)

http://roots.ntis.go.kr/RiSrchService?target=211&pjtId=1350018168

응답메시지

```
<?xml version="1.0" encoding="UTF-8"?>
<ResultList>
        <ResultInfo>
                <TotalCount>1</TotalCount>
        </ResultInfo>
        <Result>
                <RschRpt>
                    <ResutId>REP-2003-0012169493</ResutId>
                    <ReportTitle>근골격계 질환예방을 위한 여러 가지 들기작업에서의
인체심리학적·생리학적 연구</ReportTitle>
                    <Publisher>동아대학교</Publisher>
                    <PubDate>200405</PubDate>
                    <OpenYn>Y</OpenYn>
                </RschRpt>
        </Result>
</ResultList>
```

자료: KISTI(2014)

[그림 2-17] 과제의 최종보고서 원문 API 연계 소스 샘플

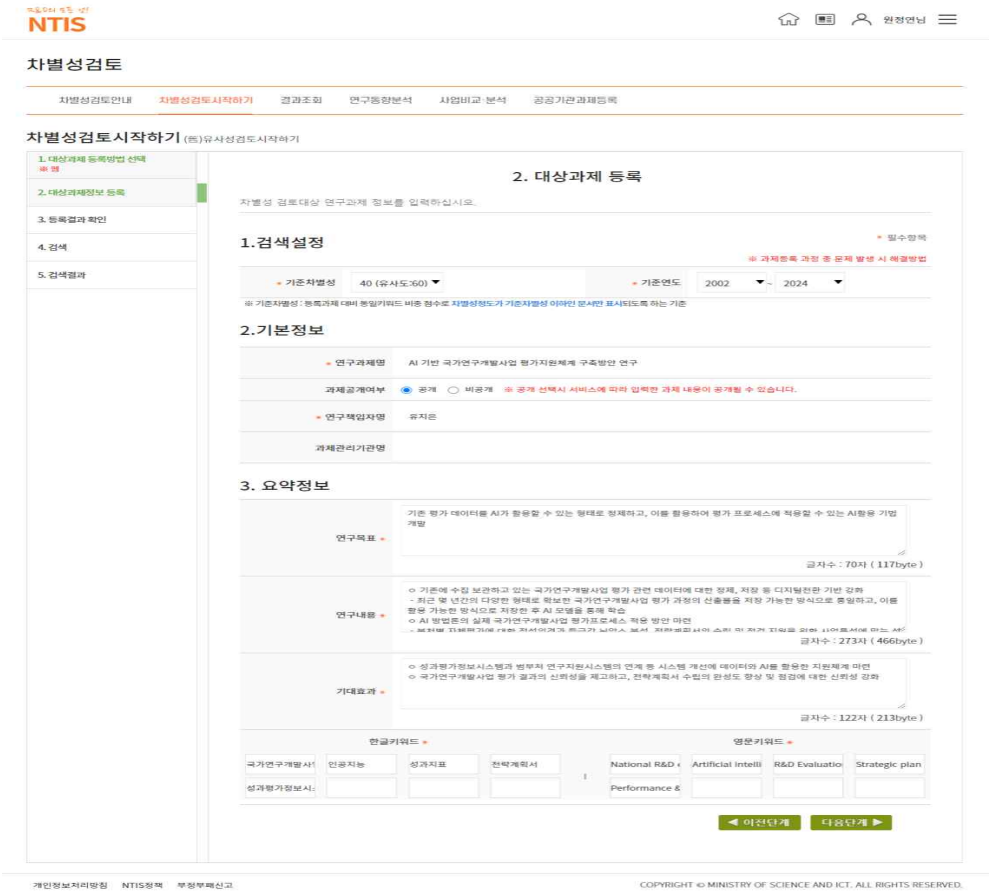
(2) 프로세스

o 차별성검토는 웹 입력과 엑셀 입력 두 가지 방식으로 진행할 수 있으며 주로 하나의 과제를 등록하는 웹 입력 방식과 여러 개의 과제를 등록하는 엑셀 입력 방식으로 구성됨

자료: NTIS 차별성검토 서비스 페이지(https://www.ntis.go.kr/yusa/ia/sp/main.do, 최종접속 : 24.2.5.)

[그림 2-18] 차별성검토 프로세스

42) 한국과학기술정보연구원(2014), 국가과학기술지식정보서비스 사업 최종보고서, 2014



자료: NTIS 차별성검토 서비스 페이지(https://www.ntis.go.kr/yusa/ia/sp/index.do#, 최종접속 : 24.2.5.)

[그림 2-19] 차별성검토 시작하기 웹 입력 화면 예시

○ 정상적으로 과제정보를 입력하면 1분 내외의 시간 내 관련 연구개발과제 차별성 검토 검색결과 화면이 출력되며, 검색결과 화면에서 관련 과제와 차별성 점수, 요약문, 과제 추가정보 등을 확인할 수 있음

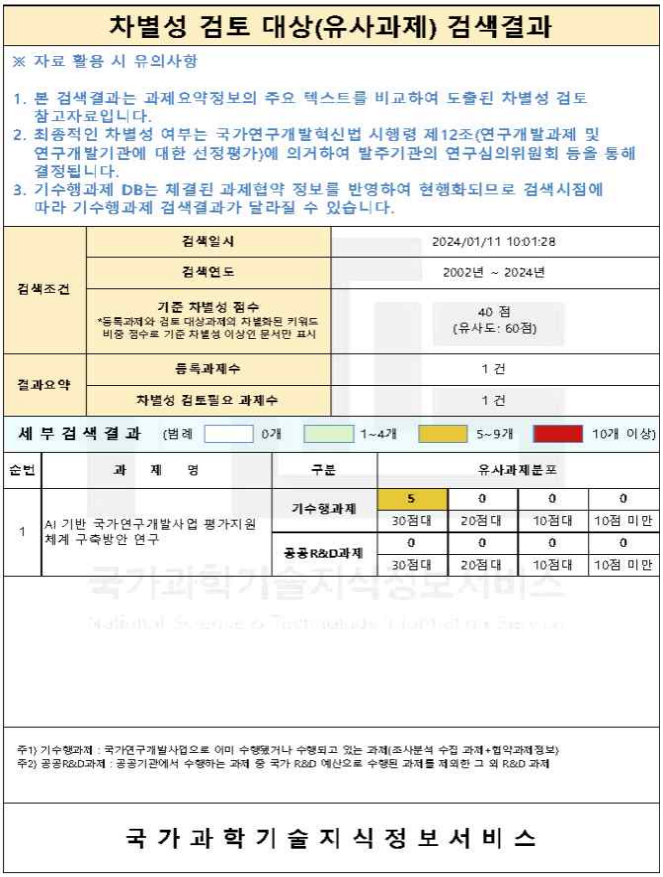
검색결과 5건

순번	수행연도	과제 고유번호	기관세부과제번호	세부과제명	연구책임자	차별성 점수	요약문 보기	과제 추가정보
과제명: AI 기반 국가연구개발사업 평가지원체계 구축방안 연구								
1	2006	1350006935	06GP10025	한국생산기술연구원의 성과목표 관리 정보시스템 구축사업	정규채	32.9점	보기	보기
2	2008	1395015162	2007-01-009	무리칭 정보화사업의 효율적 평가방법에 관한 연구	남은영	33.73점	보기	보기
3	2014	1711021217	AE14060	국가연구개발 성과평가계획 수립을 위한 평가체계 분석 및 발전방안 연구	이태근	35.8점	보기	보기
4	2015	1345242400	2015S1A5A2A02048559	요인분석실험방식을 통한 관광정보추천시스템의 잠재성 평가 분석	차정진	37.72점	보기	보기
5	2015	1711034024	CN15010	2015년도 미래창조과학부 직할 출연(연) 평가	신분봉	39.72점	보기	보기

자료: NTIS 차별성검토 서비스 페이지(https://www.ntis.go.kr/yusa/ia/sp/index.do#, 최종접속 : 24.2.5.)

[그림 2-20] 차별성검토 시작하기 검색결과 예시

o 마지막 단계에서 검색결과증을 pdf로 제공하여 추후 연구계획 발의서 등에 활용할 수 있음



차별성 검토 대상(유사과제) 검색결과

※ 자료 활용 시 유의사항

1. 본 검색결과는 과제요약정보의 주요 텍스트를 비교하여 도출된 차별성 검토 참고자료입니다.
2. 최종적인 차별성 여부는 국가연구개발혁신법 시행령 제12조(연구개발과제 및 연구개발기관에 대한 선정평가)에 의거하여 발주기관의 연구심의위원회 등을 통해 결정됩니다.
3. 기수행과제 DB는 체결된 과제협약 정보를 반영하여 현행화되므로 검색시점에 따라 기수행과제 검색결과가 달라질 수 있습니다.

검색조건	검색일시	2024/01/11 10:01:28
	검색연도	2002년 ~ 2024년
	기준 차별성 점수 *등록과제와 검토 대상과제의 차별화된 키워드 비중 점수로 기준 차별성 이상인 문서만 표시	40 점 (유사도: 60점)
결과요약	등록과제수	1 건
	차별성 검토필요 과제수	1 건

세 부 검 색 결 과 (범례 0개 1~4개 5~9개 10개 이상)

순번	과 제 명	구분	유사과제분포			
1	AI 기반 국가연구개발사업 평가지원체계 구축방안 연구	기수행과제	5	0	0	0
			30점대	20점대	10점대	10점 미만
		공공R&D과제	0	0	0	0
			30점대	20점대	10점대	10점 미만

주1) 기수행과제 : 국가연구개발사업으로 이미 수행됐거나 수행되고 있는 과제(조사분석 수집 과제+협약과제정보)
주2) 공공R&D과제 : 공공기관에서 수행하는 과제 중 국가 R&D 예산으로 수행된 과제를 제외한 그 외 R&D 과제

국 가 과 학 기 술 지 식 정 보 서 비 스

자료: NTIS 차별성검토 서비스 페이지(https://www.ntis.go.kr/yusa/ia/sp/index.do#, 최종접속 : 24.2.5.)

[그림 2-21] 차별성 검토 대상 검색결과증 예시

(3) 특징 및 시사점

o 차별성 검토는 당초 최다빈도 단어 기반의 머신러닝 모델로부터 유사한 단어 또는 키워드의 개수를 가지고 유사도를 판단하는 방식에서 현재는 과제정보 입력의 메커니즘을 학습하여 다차원의 유사도 분석이 가능하고, 현재까지 축적된 다량의 연구개발과제 시계열 데이터를 분석하여 직관적인 유사도(%) 결과를 반환할 수 있다는 점에서 그 장점이 있음

o 다만, 차별성 검토과제로 검색된 과제가 입력한 과제정보와 유사하지 않는 등 데이터의 오류가 있을 수 있다는 점에서 유사도 이상의 참고 및 활용에 한계가 있을 수 있음

- NTIS 차별성 검토 서비스는 입력한 과제정보에서 키워드를 추출하고, 기존의 과제정보와 키워드 비교를 통해 비교하기 때문에 연구내용이 다르더라도 같은 단어들이 함께 포함되어 있다면 차별성검토과제로 검출될 수 있음

- 따라서 차별성검토 서비스는 참고자료로만 활용하고, 최종적인 과제의 차별성 여부는 발주기관의 연구심의위원회 등에서 재검토가 필요함

ㅇ 또한 신청하려는 과제나 유사과제의 pool로 활용하고자 하는 과제가 보안과제에 해당할 경우 상세 내용을 열람하고 분석할 수 없기 때문에 산출된 유사도를 정확하게 활용하는 데에는 한계가 있을 수 있음

- 차별성검토서비스의 데이터는 여러 연구기관과 공공기관으로부터 입력받는 정보들을 이용하여 운영하고 있어, 기관에서 제공하는 원데이터의 상태에 따라 정보의 편차가 존재함

ㅇ 유사과제 데이터를 검토하는 방식은 기관에 국한된 자료에 근거하거나, 제한된 혹은 너무 광범위한 정보를 기반으로 판단해야하는 키워드 매칭 검색을 이용하므로 정확한 중복과제 파악이 어려움(홍세호, 2013)

4. NTIS 과학기술표준분류추천 서비스

ㅇ 국가R&D사업에 참여하는 연구자를 지원하기 위해 연구내용 및 키워드 등 입력 시, 그에 맞는 국가과학기술표준분류 및 부처 분류 2종 연구 분야, 각 소분류 최대 5개까지 적합한 연구분야를 추천함

(1) 데이터

ㅇ 딥러닝 기반의 연구보고서의 제목, 키워드를 중요한 요소로 이용한 TK_CNN(Title - Keyword Convolutional Neural Network)기법[43]을 통해 국가과학기술표준분류 자동분류 시스템을 구축(김윤정, 2021)

- 연구보고서의 메타 데이터들에서는 과제의 키워드와 과제명에 자주 나타나는 단어는 대부분 범주를 대표하는 경향을 보였으며 이러한 근거에 기반하여, 각 범주에 해당하는 대표적 단어들을 선별·분리 등의 과정을 거쳐 학습을 진행함(최종윤, 2020)

ㅇ 머신러닝 기반으로 사용자의 요청에 적합한 국가과학기술표준분류 연구분야[44]를 추천, 국가R&D 과제(2013년~), 보고서(인문사회 등 일부), 국립국어원 말뭉치를 학습에 활용함

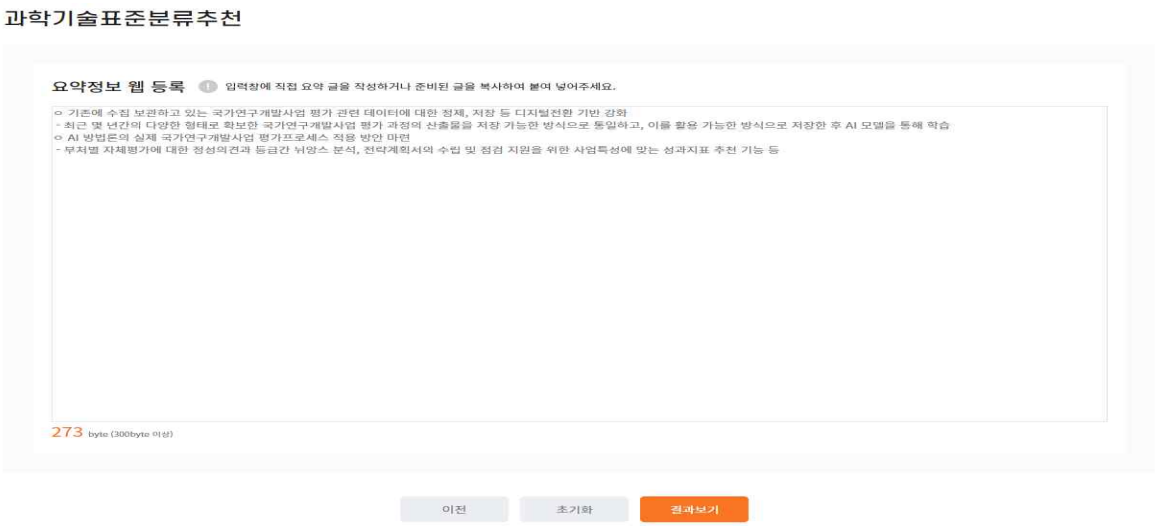
43) 핵심단어가 존재하는 제목과 키워드 필드(Title-Keyword)만을 입력으로 하는 문장의 단어 벡터에 대해서 임베딩된 데이터를 이용하여 학습을 진행하는 모델(CNN)
※ 출처: 최종윤 외, 국가 과학기술 표준분류 체계 기반 연구보고서 문서의 자동 분류 연구, P.172

44) 2023년부터 관련 규정의 개정으로 서비스 상 연구분야라는 용어로 지칭

- 2013년부터 2019년까지의 과제 데이터를 대상으로 Word2vec모델[45]에 파라미터(dimension, iteration, window size, negative sample)를 변경하면서 반복학습함
- 연구분야 추천 시스템은 입력되는 과제정보(sentence)에 대해 Word2vec 모델을 통하여 단어를 벡터화하고 과제정보의 단어 벡터들을 기반으로 Doc2Vec 모델[46]을 통해 문장 벡터를 생성함
- K-NN(K-Nearest Neighbor)[47]을 사용하여 가장 근접한 추천 후보 5개를 선정하고, Naive Bayes[48]를 이용하여 소분류에 대한 최종 점수를 계산함(김윤정, 2021)

(2) 프로세스

ㅇ 연구개발목표, 연구개발 내용, 활용계획 및 기대효과, 국문 핵심어, 영문 핵심어 중 일부 내용 입력 후 추천 결과 확인 가능



자료: NTIS 과학기술표준분류추천 서비스 페이지(https://www.ntis.go.kr/autoclass/AutoClsRqstIns.do, 최종접속 : 24.2.5.)

[그림 2-22] NTIS 과학기술표준분류추천 요약정보 등록화면 예시

45) Word2vec: 단일 알고리즘이 아닌, 대규모 데이터 세트에서 단어 임베딩을 학습하는 데 사용할 수 있는 자연어처리 모델. https://www.tensorflow.org/text/tutorials/word2vec

46) Doc2vec: 단어를 숫자 벡터로 표현하는 접근 방식인 Word2Vec의 확장으로 도입되었으며, 문서 임베딩을 학습하는 데 사용되는 자연어처리 모델. https://www.geeksforgeeks.org/doc2vec-in-nlp/

47) K-NN(K-Nearest Neighbor): 분류 문제를 해결하는 대표적인 방식. K값은 사용자가 지정해주는 값으로 분류하고자 하는 데이터와 가장 가까운 K개의 데이터를 찾아서 어떤 범주에 많이 속하는지를 분석해 입력된 데이터를 분류하는 방식. AI타임스(https://www.aitimes.com),

48) Naive Bayes: 속성값을 확률적으로 예측해 분류하는 알고리즘. KNN과 다르게 분류된 속성일지라도 나이브 베이스는 그룹에 속할 확률을 분석하기 때문에 같은 그룹으로 분류될 수 있음. https://ditoday.com/%ec%84%9c%eb%93%9c-%ed%8c%8c%ed%8b%b0-%ec%bf%a0%ed%82%a4%ec%9d%98-%ec%a0%9c%ed%95%9c-%eb%a6%ac%ed%83%80%ea%b9%83%ed%8c%85-%ea%b4%91%ea%b3%a0%eb%8a%94-%ec%82%ac%eb%9d%bc%ec%a7%88%ea%b9%8c/

o 사용자가 입력한 정보와 비슷한 연구내용의 연구과제 리스트를 제공하고, 과제를 클릭하면 해당 과제의 상세정보 확인 가능

자료: NTIS 과학기술표준분류추천 서비스 페이지(https://www.ntis.go.kr/autoclass/AutoClsRqstIns.do, 최종접속 : 24.2.5.)

[그림 2-23] NTIS 과학기술표준분류추천 결과(웹)

o 추천 결과를 대-중-세부영역(최대 5개)의 코드와 코드명, 정확도 값으로 제공하며 추천결과 정확도 1~5순위의 연구분야의 주요 용어를 워드 클라우드 및 표 형태로 제공

자료: NTIS 과학기술표준분류추천 서비스 페이지(https://www.ntis.go.kr/autoclass/AutoClsRqstIns.do, 최종접속 : 24.2.5.)

[그림 2-24] NTIS 과학기술표준분류추천 워드 클라우드

(3) 특징 및 시사점

- 과학기술표준분류 추천 서비스는 과제 내용의 요약과 키워드를 합성곱신경망 모델을 통해 학습하여 문서를 군집화하는 방식으로 가장 적합한 표준분류를 추천해줄 수 있으며, 연구개발과제 정보 축적에 있어서 주요 통계 구분자로 활용되는 과학기술표준분류의 정확도를 높이는 데 있어 휴먼 바이어스를 줄이고자 하는 데 의의가 있음

- 머신러닝 기반의 연구분야 추천 시스템은 정확도를 높이기 위해 추후 데이터셋의 주기적으로 반영이 필요함

- 추천된 표준분류의 정확도가 낮을 수 있다는 위험성이 있어 제공하는 결과에 대한 심층 분석 결과(시사점), 분석 결과에 대한 정책적 활용 등에 사용하기에는 부족하며, 대체로 정책 입안자가 아닌 일반사용자 관점의 시각화 서비스 기능 정도로 볼 수 있음(KISTEP, 2022)

5. KISTEP 국가연구개발사업 과제데이터 지능형 검색 서비스[49)]

ㅇ KISTEP에서 수행하는 과학기술혁신 정책기획 및 평가 업무의 디지털 전환 가속을 위한 방안으로 인공지능 지능형 분석 서비스를 개발하여 KISTEP 내부에 축적하고 있는 방대한 데이터의 효과적인 정보 추출, 연관성(관계성) 분석 방법을 통한 데이터(근거) 기반의 R&D 정책의사결정을 목표로 함

- 2017년~2022년 국가연구개발사업의 과제 데이터를 바탕으로 벡터 데이터베이스[50)]를 활용할 수 있는 오픈소스 AI 어플리케이션인 Milvus[51)]에 연동하여 과제 및 사업데이터의 검색 대시보드를 제공

(1) 데이터

ㅇ 도메인 특화 사전학습 언어모델인 KISTEP DAPT(Domain Adaptive Pre-Training) 모델을 사용

- (DeBERTa[52)]) 특정 도메인에서 빈번하게 사용되는 고유한 어휘와 전문용어를 특화된 어휘로 처리할 수 있는 도메인 특화 토크나이저(kobigbird)를 통해 전문용어, 축약어, 해당 도메인에 필요한 언어 스타일을 고려한 텍스트 처리가 가능

- (KorSciDeBERTa[53)]) DeBERTa 기반의 모델로서 논문, 연구보고서, 특허, 뉴스, 한국어 위키 말뭉치를 사전 학습한 모델이며 총 146GB의 Corpus를 이용하여 학습된 모델임

ㅇ 국가연구개발사업정보 지능형 검색 서비스는 2016년~2022년까지의 국가연구개발사업 데이터 1,703건을 활용하여 벡터 데이터베이스를 구축함

49) 본 서비스는 K2BASE 내 오픈 예정 서비스로, '24년 2월 기준 검토된 내용을 바탕으로 작성됨 (한국과학기술기획평가원, 2023년 과학기술정책의 과학화 기반구축 연구. 2024(예정))

50) 벡터 데이터베이스는 각종 자연어, 이미지, 비디오 등 비정형데이터를 AI를 이용한 embedding 모델의 형태로 변환하여 적재할 수 있음

51) Milvus는 벡터 데이터베이스를 활용할 수 있는 오픈소스 AI 어플리케이션으로 타 데이터베이스 대비 탐색 속도가 빠르다고 알려져 있음

52) Decoding-enhanced BERT with dis-entangled attention의 줄임말로써, Disentangled Attention Mechanism과 Enhanced Mask Decoder를 기반으로 기존에 있던 RoBERTa 보다 우수한 성능을 달성함(한국과학기술기획평가원, 2023년 과학기술정책의 과학화 기반구축 연구. 2024(예정)

53) KISTI에서는 DeBERTa 모델의 아키텍처를 기반으로 과학기술 영역의 코퍼스를 추가적으로 학습한 KorSciDeBERTa 모델을 2023년 8월에 발표함

[표 2-3] 국가연구개발사업정보 데이터(2016년~2022년)

전체	2022년	2021년	2020년	2019년	2018년	2017년	2016년
1,703건	524건	481건	205건	150건	129건	52건	162건

자료: 한국과학기술기획평가원, 2023년 과학기술정책의 과학화 기반구축 연구. 2024(예정)

- ㅇ 기준년도, 사업코드, 부처명, 내역사업명, 세부사업명, 사업목적, 전략목표, 최종성과목표 정보로 데이터셋 구성
  - 2016년~2020년은 성과목표지표계획서 데이터, 2021년~2022년은 전략계획서 데이터를 활용하여 사업정보와 성과평가 관련 정보항목을 구성

(2) 프로세스

- ㅇ 검색 결과를 얻기 위해 기존 키워드 방식으로 검색하는 것이 아닌, 긴 문장의 자연어 형태로 입력하여 시멘틱 검색을 하여 의도와 문맥을 이해하는 검색 결과를 얻을 수 있음
- ㅇ 관심있는 ①기준년도, ②부처명 드롭박스를 선택하고 ③유사도(1~99) 하한값을 지정한 후 ④검색문구에 검색할 쿼리를 입력하고 ⑤“Apply” 버튼을 클릭
- ㅇ 텍스트 변환 기능은 관심이 있는 쿼리가 쉼표나 세미콜론 기호가 포함된 긴 문장을 복사하여 붙여넣기 후 검색
  - ‘UAM 운용을 위한 교육체계를 구축하고, 검증 및 교육훈련기법 등 인력양성 기술 확보 등 단계적 추진’ → 사이트 우측 텍스트 변환 기능을 반드시 사용
  - 그대로 사용하는 경우 ‘UAM 운용을 위한 교육체계를 구축하고’, ‘검증 및 교육훈련기법 등 인력양성 기술 확보 등 단계적 추진’ 2개가 동시에 검색되는 오류가 발생함
- ㅇ 검색 결과는 우측 상단의 “ ⋮ ” 클릭 〉 ⑥“Download” 〉 “Export to.csv” 버튼을 클릭하여 다운로드 가능

자료: 한국과학기술기획평가원(2024), 2023년 과학기술정책의 과학화 기반구축 연구.

[그림 2-25] 국가연구개발사업 및 과제 데이터 지능형 검색 기능(예정)

(3) 특징 및 시사점

- o 딥러닝 모델을 기반으로 일반 자연어 문서와 도메인 지식과 전문용어를 포함한 특화 학습 모델을 활용하여 고도화된 과제 및 사업검색 서비스 제공이 가능하다는 점에서 의의가 있음
- o 다만, 국가연구개발사업 과제 데이터의 지속 가능한 자동 업데이트 및 재학습이 어려움
  - 현재 시스템에서 사용된 데이터는 'AI 평가지원체계 구축 방안 연구'를 위해 특수문자, 목차어 제거 등 전처리 과정을 거쳐 수집된 데이터로, 추후에 추가되는 데이터 또한 전체 수작업[54]이 필요함
- o 시스템 구축 초기 단계로, 현재 NTIS, IRIS(범부처 통합 연구지원시스템) 등 타 정보시스템과의 연계가 불가능하며, 그에 따라 상시 쌓이고 있는 연구개발과제와 사업 정보를 학습할 수 없다는 한계가 존재

54) 전략계획서는 정형화된 양식·항목으로 자동수집이 가능한 반면, 자체평가보고서는 전체 수작업을 요하며, 등급별 불균형 심각

제4절 소결 : 국가연구개발사업 평가에의 AI 적용 가능성과 이슈

□ **자연어처리 모델은 정책 지원 Tool 로서 정책 수립, 정책 해석, 정책 소통, 정책 평가 등 다양한 정책과정에서의 활용 가능성이 있으며, 국가연구개발사업 평가의 효율성, 신뢰성 및 객관성 제고라는 이점을 제공할 것으로 기대됨**

ㅇ 자연어처리 모델의 활용은 정책의 효과성을 증대시키고 당위성을 확보하게 해준다는 점에서 증거 기반 과학 기술 및 혁신(Science technology innovation, STI) 정책 결정 프로세스의 고도화에 기여할 수 있음

- 증거 기반 정책입안과 소통에 관한 연구는 활발한 반면 R&D사업의 기획, 평가, 관리, 예산의 전략적 투자 등에 관한 전주기 추진 프로세스에 도입된 사례는 거의 없으며, 전문지식과 노동 집약적 업무 프로세스로 인한 비효율이 반복되고 있음에도 이를 혁신적으로 경감시켜줄 수 있는 수단이 마련되어 있지 않아 주관성과 인지오류 등이 반복되고 있는 실정임

- 이는 국가 R&D 정책 지원 보조수단으로서 공공 데이터와 자연어 처리 모델을 활용할 경우 일부 해결할 수 있을 것으로 판단됨

- 예를 들어 정책 수립과 관련하여 도로분야 ITS 정책 이슈 탐색기법(오창석, 2016), 기후변화와 관련 대응 정책에 대한 인식 및 정서 연구(김영현, 2022), 일본 IT 신전략 연도별 정책내용 변화 연구 등 정책 동향 분석을 위한 자연어처리 모델이 다수 연구되고 있음

ㅇ 특히, 국가연구개발사업 평가에 자연어처리 모델 활용 시 평가자의 편견이나 주관성의 위험을 감소시킬 수 있을 뿐만 아니라, 반복 작업을 자동화하거나 효율성, 신뢰성 및 객관성 제고 가능하다는 점에서 적용을 검토해볼만한 가능성이 있다고 할 수 있음

- 국가연구개발사업의 전략계획서와 기획보고서, 예산요구서 간의 유사도 분석을 통한 사전 기획 내용의 사업반영 충실도 검토 작업의 효율성 향상 가능

- AI를 활용하여 서로 다른 평가 등급 간의 질적 의견과 뉘앙스 분석을 수행함으로써 평가 프로세스의 일관성과 공정성 향상 가능

- 새로운 연구과제의 성과지표 작성 시 내용 유사도와 핵심 키워드를 기반으로 기존 사업의 성과지표 사례를 추천해주는 시스템을 개발하여 성과지표 수립 및 점검 업무의 효율성 및 신뢰성을 향상시킬 수 있음

- AI의 주요 활용 분야 중 하나인 추천시스템 도입을 통하여 평가 과정에서 사용된 기준과 요소(항목)를 식별하기 쉽게 하고 장기간 데이터를 쌓아 학습시키게 되면 인공지능 알고리즘을 통해 공정하고 일관성 있는 평가에 기여할 것을 기대해볼 수 있음

□ **다만, 자연어처리 모델의 도입 시 데이터 수집 및 조화, 데이터 분석 등 도입 과정에 어려움과 부작용이 발생할 수 있으며, 특히 국가연구개발사업 평가에의 활용 시 신뢰성이나 기술적 문제 등으로 인한 제한이 있어 업무를 완전 대체하는 데에는 한계가 있을 것으로 예상됨**

o 자연어처리 모델을 통해 의견을 분류하거나 토픽모델링 등 유형화하여 단기에 많은 의견들과 트렌드를 검토하고, 이를 토대로 정책수립이나 의사결정에 활용할 수 있다는 장점이 있으나, 신뢰성이나 기술적 한계 등에 의한 제한점이 존재함

- 자연어처리 모델 개발을 위해서는 데이터 질적·양적 품질이 확보되어야 하므로 기존 데이터 정제 및 저장, 활용을 위한 기반구축이 필요함

  • 세계경제포럼(WEF)에서는 2019년 공공부문 AI 도입의 방해요인 중 효과적인 데이터 활용의 어려움을 꼽았으며, 공공부문의 데이터 형태와 AI 활용의 연계성이 아직 어렵고 데이터 이해를 위한 기능 및 거버넌스가 부족하다는 점을 지적

  • 서호준(2019)은 텍스트 네트워크 분석을 통해 우리나라 과학기술정책 50년의 주요 의제를 도출하였으나, 과학기술 50년사라는 단일 텍스트를 마이닝의 대상으로 삼았기 때문에 도출할 수 있는 키워드가 제한적이고 그 해석과 정책함의 도출에 일정한 한계가 있음을 제시하여 심층적이고 종합적인 분석을 위해서는 다각적인 차원의 텍스트 원천의 중요성을 시사[55)]

- 또한, 잘못된 과거의 데이터로 학습했을 경우 나타날 수 있는 학습된 모델의 오류와 이를 통해 예측된 결과의 오류로 인한 효과를 진단하기 어렵고, 동시에 인공지능과 데이터를 활용해 프로세스 개선의 효과가 뚜렷하게 측정되기도 어렵기 때문에 도입 시에 충분한 검토가 수반되어야 함

o 특히, 자연어처리 모델 개발에 있어서 반드시 텍스트 전처리 과정이 필요하여 비용 효율성의 이슈가 발생하고, 자연어처리 모델의 완성도는 학습한 데이터의 양적·질적 품질이 중요하나 국가연구개발사업의 평가에 있어서 데이터의 양과 품질이 자연어처리 모델 개발에 충분하지 않을 수 있음

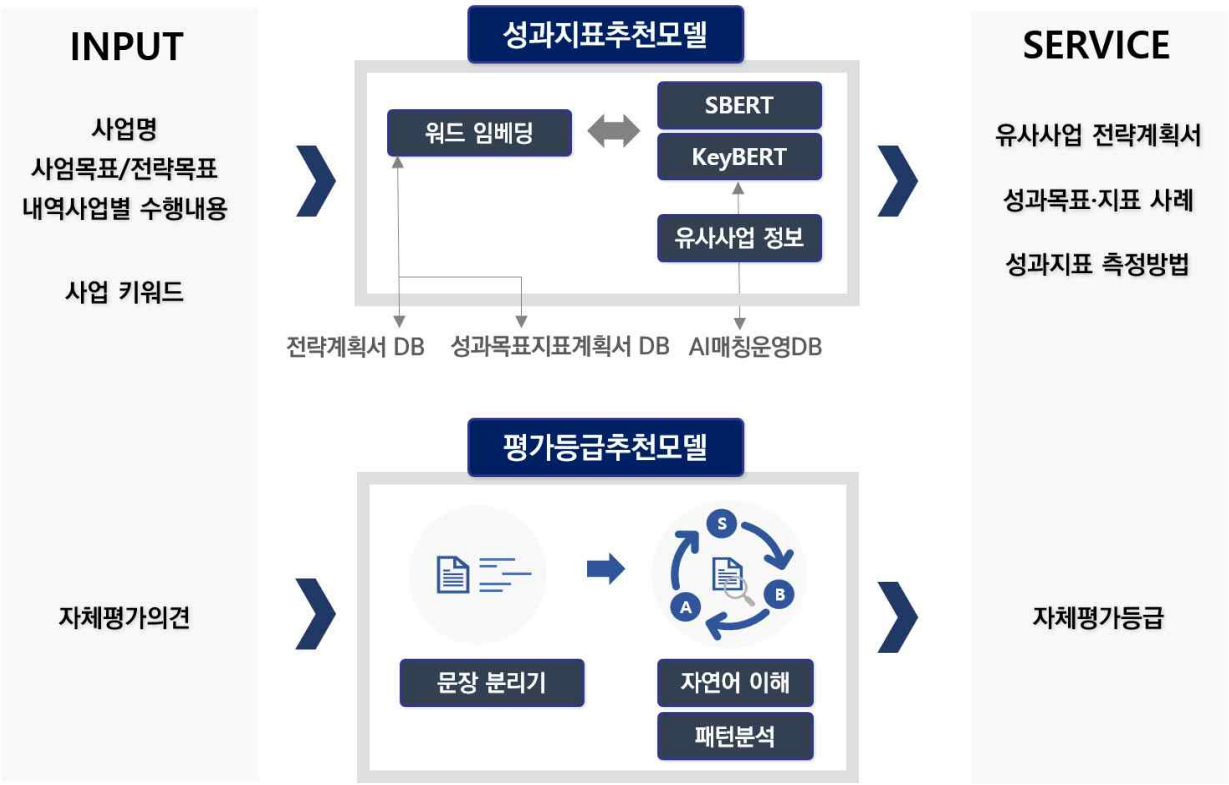
o 국가R&D 프로세스에 먼저 인공지능을 활용한 NTIS나 K2BASE의 사례를 보았을 때, 기존의 데이터로 인한 모델의 편향이나 오류의 가능성을 고려했을 때 부가적인 정보를 제공해주고, 이를 사람이 판단해서 활용할 수 있도록 하는 방식의 보조수단으로서 활용하는 것이 바람직해보이며, 평가와 같은 중요한 의사결정 과정을 대체하기엔 한계가 있음

55) 서호준. "텍스트 네트워크 분석을 활용한 우리나라 과학기술정책 50 년의 주요 의제 분석-[과학기술 50 년사] 를 중심으로." 과학기술정책 2.2 (2019): 171-201.

□ 따라서 국가연구개발사업 평가에서 자연어처리 모델이 가치 있는 도구가 될 가능성이 있지만 현재 개발 수준은 제한된 범위 내에서 다른 평가 방법을 보조하는 수단으로 활용하며 데이터 기반 평가 체계의 활용을 위한 기반을 구축할 필요

- o 국가R&D사업 평가에서 자연어처리 모델이 가치 있는 도구가 될 가능성이 있지만 현재 개발 수준은 제한된 범위 내에서 다른 평가 방법을 보조하는 수단으로 활용되어야 함
- o 자연어처리 모델을 활용한 결과에 전적으로 의존하기보다 평가 시 고려하는 하나의 추가적인 유용한 정보로 활용하며 지속적으로 개선시켜 향후 자연어처리 모델 활용의 효과성을 향상시킬 필요가 있음
- o 우선적으로는 필요한 데이터의 선정, 구축을 통해 데이터의 사용 용이성을 확보하고 데이터 증량 방법 등의 기술 적용을 통하여 데이터의 양적·질적 품질을 확보할 필요가 있음

□ 문헌연구와 선행 사례 연구를 바탕으로 본 연구는 아래와 같이 AI 기반 평가지원체계 마련을 위한 기반연구로서 사업평가 프로세스 효율화가 필요한 세부 업무 프로세스를 선정하고, 인공지능을 활용하여 부가적인 정보를 제공하고, 의사결정을 지원할 수 있는 세부과업을 추진하고자 함



[그림 2-26] 인공지능 기반 국가연구개발사업 평가지원체계 개념(안)

ㅇ AI 기반 국가연구개발사업 평가지원체계의 초기 단계는 반복업무의 비효율성이 높고 인공지능 도입 시 정보제공의 효율성 제고 효과가 가장 높은 전략계획서 성과지표 사례 추천 및 자체평가 의견에 따른 등급 추천으로 세부 과업목표를 선정

ㅇ 사업 유사도 분석 기반 성과지표 사례추천과 자체평가의견에 따른 등급 추천이 가능한 자연어 처리 기반 인공지능 모델을 구축하여 국가R&D 사업평가 프로세스에 적용이 가능할 것으로 기대됨

- 성과지표설정지원 서비스는 사업명, 사업목적, 세부 수행내용, 키워드 등 검색조건과 유사도 기준치를 사용자가 선택하는 등 검색조건별 관련사업 검색이 가능한 모델 구현

- 평가등급설정지원 서비스는 등급별 평가의견의 패턴을 분석해 신규 의견의 등급을 분류할 수 있는 적절한 오피니언 마이닝 모델을 확인하여 적용방안 모색

ㅇ 인공지능 기반 평가지원체계 구축을 통해 궁극적으로는 사업평가 및 상위점검 업무의 효율화를 도모하고, 평가의 전문성과 신뢰성을 제고하는 데 기여하는 것을 목표로 함

- 성과지표 설정 시 사업 특성에 부합하는 관련사업의 우수 사례를 인공지능이 추천해주고, 사업의 추진전략과 성과를 적절히 평가할 수 있는 지표를 선택하도록 지원

- 자체 평가 우수 사례에 대한 학습을 통해 해당 자체평가시 활용 단어 등과의 유사도를 분석하여 AI가 추천등급 표시

Ⅲ. 전략계획서 성과지표 설정 지원 서비스 기획 연구

o 전략계획서 성과지표 설정지원 서비스는 새로운 연구과제의 성과지표 작성시 내용 유사도와 핵심 키워드를 기반으로 기존 사업의 성과지표 사례를 추천해주는 시스템을 개발하여 성과지표 수립 및 점검 업무의 효율성 및 신뢰성 향상 추진을 목적으로 함

제1절 데이터

o 총 데이터의 수는 전략계획서 1,757개로 다른 인공지능 활용을 위한 데이터셋에 비해서는 비교적 규모가 크지 않음

- 전략계획서 대상 문서수와 데이터 수가 다른 이유는 전략계획서 작성 단위가 세부사업을 기본으로 하고 있으나, 방위사업청의 경우 47개 세부사업을 1개의 단위사업으로 묶어 전략계획서를 작성하는 등 일부 상이한 문서형태때문인 것으로 보임

[표 3-1] 인공지능 활용을 위한 평가데이터셋 구축 결과 개요

수집 목적	수집 문서	대상연도	데이터 수	구축 항목
모델구축	전략계획서 (981개)	2021 - 2022	1,005개	사업코드, 부처명, 사업명(단위/세부/내역), 내역사업별 연구활동내용, 사업목적, 추진방식, 사업유형, 다부처여부, 전략목표, 성과목표/지표, 측정방법 등
	자체평가보고서 (277개)	2022 - 2023	646개	부처명, 사업명, 성과목표, (성과유형), 평가부문별 평가의견 요약, 세부 평가의견, 등급 등
모델 검증 · 보완	성과목표지표계획서 (731개)	2016 - 2020	752개	모델구축용 데이터와 동일
	자체평가보고서 (221개)	2019 - 2021	1,240개	

o 데이터 구축은 데이터 부족으로 인한 오류 가능성과 향후 평가·검증 시 활용을 고려하여 평가 데이터 추가 구축 작업을 실시하고 전략계획서 이전의 제도인 성과목표지표계획서를 활용함

- 성과목표지표계획서의 경우 일부 항목이 전략계획서와 상이하나, 관련사업 분석 시에 필수적인 정보항목이 아니라는 점, 사업정보와 지표정보 입력체계가 현행 전략계획서와 유사하다는 점에서 보강 데이터로 활용하기에 적절함

- 다만, 성과목표지표계획서의 경우 2016년 이전 성과지표 구분이 현재와 상이하고, 작성양식 또한 일부 상이하는 등 현재의 데이터와 통합하여 사용하기에 이질성이 높아 2016년부터 2020년까지 5개년을 데이터 구축 대상으로 설정

[표 3-2] PEIS 전략계획서와 성과목표지표계획서의 데이터 입력 항목 비교

구분	전략계획서 (2021년 이후)	성과목표지표계획서 (2018년~2020년)	성과목표지표계획서 (2017년)	성과목표지표계획서 (2016년)
Ⅰ. 사업개요	코드연도	연도	연도	연도
	사업코드(평가코드)	-	-	(사업명)
	부처명	(부처)	(부처명)	-
	단위사업명	단위사업	단위사업	단위사업
	세부사업명	(사업)	(세부사업)	(사업)
	내역사업명	내역사업	내역사업	내역사업
	특이사항	-	-	-
	세부내용	사업별 추진계획	사업별 추진계획	사업별 추진계획
	사업목적	사업목적	사업목적	사업목적
	사업추진경위 (법적근거)	지원근거	지원근거	지원근거
	사업추진경위 (상위계획)	추진경위	추진경위	추진경위
	사업구분	사업구분	사업구분	사업구분
	사업추진방식	사업추진방식	사업추진방식	사업추진방식
	사업유형	(사업유형)	(부처제출유형)	(사업유형(부처제출))
	다부처여부	-	-	-
	참여부처	-	-	-
	사업기간(시작)	사업기간(해당년도)	사업기간	사업기간
	사업기간(종료)	사업기간	사업기간	사업기간
	사업규모	사업규모	사업규모	사업규모
	총사업비	총사업비	총사업비	총사업비
	지원대상	지원대상	지원대상	지원대상
	지원형태	지원형태	지원형태	지원형태
	지원조건	지원조건	지원조건	지원조건
	사업시행주체	사업시행주체	사업시행주체	사업시행주체
	예비타당성통과여부	예비타당성통과여부	예비타당성통과여부	예비타당성통과여부
Ⅱ. 성과목표지표	전략목표	전략목표	전략목표	전략목표
	최종성과목표	성과목표	성과목표	성과목표
	단계	단계	단계	단계
	단계(시작연도)	기간	기간	기간

구분	전략계획서 (2021년 이후)	성과목표지표계획서 (2018년~2020년)	성과목표지표계획서 (2017년)	성과목표지표계획서 (2016년)
	단계(종료연도)	기간	기간	기간
	성과목표명	성과목표	성과목표	성과목표
	성과목표가중치	-	-	-
	성과목표설정근거	설정근거	설정근거	설정근거
	관련내역사업	관련내역사업	관련내역사업	관련내역사업
	성과지표명	성과지표명	성과지표명	성과지표명
	지표단위	단위	단위	단위
	성과유형	성과유형	성과유형	성과유형
	지표유형	지표유형	지표유형	지표유형
	질적지표여부	질적지표	질적지표	질적지표
	성과지표설정사유	성과지표설정사유	성과지표설정사유	성과지표설정사유
	목표치설정방법및근거	목표치설정방법및근거	목표치설정방법및근거	목표치설정방법및근거
	실적치측정산식및 방법과시기	측정산식및방법,시기	측정산식및방법,시기	측정산식및방법,시기
	실적자료출처	자료 출처	자료 출처	자료 출처
	총사업기간(시작)	-	-	-
	총사업기간(종료)	-	-	-
	1단계(시작)	-	-	-
	1단계(종료)	-	-	-
	2단계(시작)	-	-	-
	2단계(종료)	-	-	-
	3단계(시작)	-	-	-
	3단계(종료)	-	-	-
	4단계(시작)	-	-	-
	4단계(종료)	-	-	-
Ⅳ. 사업평가계획	1차평가예상연도	-	-	-
	2차평가예상연도	-	-	-
	3차평가예상연도	-	-	-
	4차평가예상연도	-	-	-
	성과관리활용계획 수행계획	-	-	-
	효과성분석보고서 수행계획	-	-	-

주1: 성과목표지표계획서의 괄호 안 항목은 데이터 추가 구축 필요성 검토 과정에서 추가한 데이터를 의미
주2: 'Ⅲ. 지식재산권 창출 활동'은 해당 시 작성 항목이며, 사업평가 외의 항목으로 분석대상에서 제외

- o 추출된 전략계획서와 성과목표지표계획서의 데이터를 살펴본 결과 양식과 작성방식이 패턴화되어 있어 추후 성과평가정보시스템에 입력된 데이터를 활용하거나 전략계획서 문서로부터 구획화하여 자동추출하는 것이 가능할 것으로 보임
  - 데이터 추출 작업의 효율성을 높이기 위해서는 매년 성과평가정보시스템에 축적되는 전략계획서 데이터셋에 표로 작성되어 포함되지 못하는 사업 세부내용 또는 연차별 연구수행내용, 추진체계 등을 추출하여 포함할 수 있는 방안에 대한 장기적 검토 필요
- o 구축된 데이터는 1차적으로 숫자, 이미지, 표 구분항목 등의 불필요한 텍스트를 모두 제거하고 특수기호, 무관한 영문자 등을 선별적으로 제거함
- o 추천시스템을 위해 전체 연구과제 데이터를 분석 가능한 원형데이터(Raw Data)로 변환하고 필요한 데이터만 추출하되, 기존에 작성된 hwp, pdf 파일의 세부 내용을 활용 가능한 txt 파일로 변환시키고 추천에 필요한 데이터를 추출하여 데이터를 정제함
- o 불필요한 단어 및 기호와 문장을 제거하여 활용 가능한 데이터로 전처리하는 과정 필요하며, 이를 바탕으로 키워드를 추출에 활용함

제2절 모델 탐색 및 보완

- o 자연어 처리 관련 선행 연구에서는 KeyBERT는 키워드 추출, SBERT 문장 임베딩을 통한 특징 추출을 위해 적용하고 있음. 이 때, 특허, 트위터 등 다양한 텍스트 데이터를 활용하였으며, 키워드 추출 및 임베딩을 통해 텍스트 요약, 특허 기반 혁신 지식 그래프 생성, 특허 랜드스케이핑, 불면증 예측, 감성 분석, 개체 일치 등 광범위한 연구 내용을 포함하고 있음
- o 아래 표는 KeyBERT 혹은 SBERT를 활용한 선행 연구를 정리하여 작성함

[표 3-3] KeyBERT 및 SBERT를 활용한 선행연구

연구 내용	모델 (KeyBERT 혹은 SBERT) 활용 내용	출처
사용자 선호도를 활용하여 생성된 요약의 품질을 개선하는 것을 목표로 함	표현 학습 (Representation learning)의 문서 임베딩 과정에서 문서의 중요한 키워드를 포착하기 위해 KeyBERT를 활용하였으며, 이를 통해 제안하고자 하는 학습 모델의 키워드 특징을 벡터화하는 데 활용됨	Nguyen et al., 2022

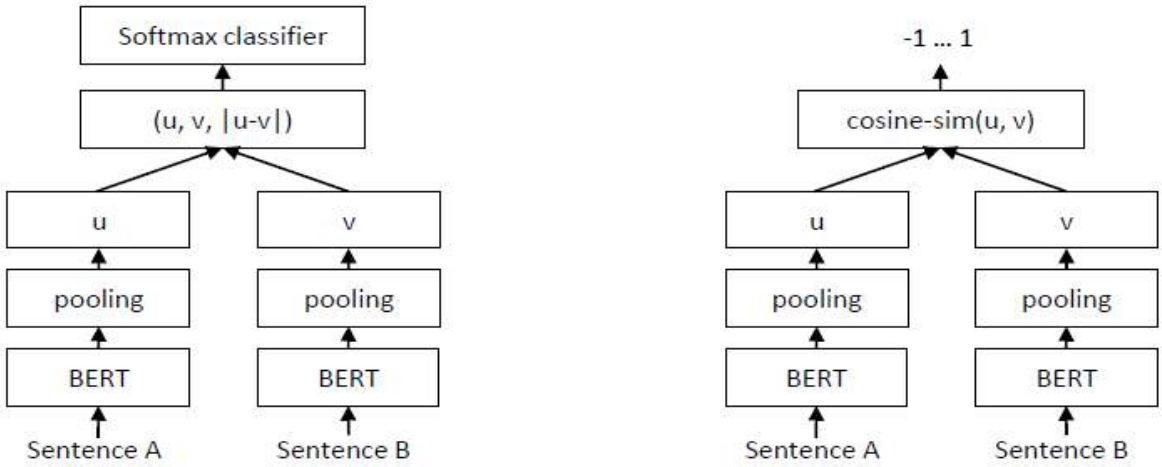
연구 내용	모델 (KeyBERT 혹은 SBERT) 활용 내용	출처
특허 마이닝을 위해 기계 학습 언어 모델을 사용하여 혁신 지식 그래프를 생성함	특허 문서 내 핵심 문장 선택을 위한 문장 분류를 위해 Sentence-BERT를 사용하여 특허 문장의 벡터화를 수행하였으며, 키워드 추출을 위해 KeyBERT를 포함한 기계 학습 언어 모델링 알고리즘을 채택함	Trappey et al., 2022
특정 산업의 기술 개발 동향을 파악하기 위한 특허 랜드스케이핑 방법론을 제시함	KeyBERT 언어 모델을 통해 키워드를 식별하였으며, 각 특허 문서와 가장 유사한 상위 10개의 키워드를 제시하도록 설정함. 이를 통해 온톨로지 구축에 활용됨	Trappey et al., 2023
불면증을 소셜 미디어 (트윗) 게시물을 분석하여 불면증을 예측하는 머신러닝 기반의 모델을 제안함	사전 훈련된 모델인 SBERT를 사용하여 트윗 게시물 텍스트를 벡터화함. 이를 통해 불면증 예측을 위한 가중화된 앙상블 분류기의 가중치로 활용함	Sakib et al., 2021
브라질과 미국의 COVID-19 관련 Twitter 콘텐츠에서 주제 감지 및 감성 분석에 대한 연구를 수행함	트윗 게시글에 대한 감성 분석을 위해 분류기 모델을 활용하였으며, 이에 대한 특징 추출을 위해 SBERT를 포함한 임베딩 모델을 활용함	Garcia & Berton, 2021
소셜 미디어 게시물을 활용하여 COVID-19 대유행 기간 동안 표현된 감정 변화를 조사함	Sentence-BERT를 통해 소셜 미디어 게시물 텍스트를 임베딩하였으며, 고차원의 임베딩 벡터에서 주성분 분석 차원을 기반으로 도출한 상위 100개를 활용해 로지스틱 회귀 분류기를 훈련함	Wang et al., 2022
사전 훈련된 언어 모델을 사용하여 깊은 개체 일치(Deep Entity Matching) 방법을 제안함	Sentence-BERT를 활용하여 벡터 유사성 검색과 함께 사용하여 일치할 가능성이 높은 레코드 쌍을 빠르게 탐색함. 코사인 유사도가 높은 쌍들만 테스트함으로써 훈련된 모델을 모든 후보 쌍에 대한 매칭 시간을 감소시킴	Li et al., 2020
이전에 팩트 체크된 주장을 식별하기 위해 다양한 기계 학습 및 정보 검색 기술을 적용하는 방법을 제안함	sentence-BERT를 사용하여 주장의 텍스트와 관련된 특징을 추출하고, 팩트 체크 데이터셋을 사용하여 학습된 모델을 평가함	Shaar et al., 2020

1. SBERT

ㅇ Sentence-BERT(SBERT)는 Siamese BERT 네트워크를 사용하여 문장 임베딩을 생성하는 방법으로 문장의 의미와 유사성을 포착하고 비교하기 위해 개발됨. SBERT는 문장 검색, 의미론적 유사도 측정, 문장 군집화 등 다양한 자연어 처리 작업에 활용될 수 있으며, BERT 및 RoBERTa보다 훨씬 효율적이고 비교적 빠른 시간 내에 대규모 문장 컬렉션에서 유사한 문장을 찾을 수 있음

ㅇ SBERT는 샴(Siamese)과 트리플릿(Triplet) 네트워크 구조를 사용하여 문장 임베딩을 생성함. 샴 구조는 두 개의 동일한 BERT 네트워크를 공유하여 각각의 입력 문장에 대한 임베딩을 생성하며, 이후 두 개의 임베딩을 비교하여 문장 간의 유사성을 측정할 수 있음. 샴 네트워크 구조를 통해 입력 문장의 고정 크기 벡터를 도출할 수 있음. 해당 연구에서는 코사인 유사도나 및 맨하탄/유클리드 거리와 같은 유사성 측정 방법을 사용하여 의미론적으로 유사한 문장을 도출함. 이러한 유사성 측정 계산은 GPU 등을 활용하여 매우 효율적으로 수행될 수 있으며, 이를 통해 SBERT는 의미론적 유사도 검색 및 군집화 등을 위해 효과적인 방법론으로 사용할 수 있음

ㅇ 아래 그림을 통해 SBERT 모델의 샴 구조 기반 학습 프로세스를 도식화 하여 나타냄. SBERT는 각 문장 쌍에 대하여, 문장 A와 문장 B를 네트워크를 통과시켜 임베딩 u와 v를 얻게 되며, 해당 임베딩의 유사성은 코사인 유사도를 사용하여 계산되고, 그 결과는 실제 유사도 점수와 비교함. 이를 통해 우리의 네트워크는 세밀하게 조정되고 문장의 유사성을 파악함



주: 코사인유사도(Cosine Similarity)는 코사인 각도를 사용해 두 벡터간 유사도를 구하는 방법으로 완전 일치일 경우 1, 완전 반대일 경우 -1의 값을 가짐

[그림 3-1] SBERT 아키텍처 (좌: 분류모델 / 우: 인퍼런스)

- SBERT의 트리플릿 구조는 세 개의 입력 문장을 사용하여 임베딩을 생성함. 첫 번째 문장은 앵커(anchor)로 사용되고, 두 번째 문장은 양(positive) 샘플로, 세 번째 문장은 음(negative) 샘플로 사용됨. 트리플릿 구조는 앵커와 양성 샘플의 유사성을 최대화하고 앵커와 음성 샘플의 유사성을 최소화하도록 임베딩을 조정하는 방식으로 학습됨. 이 때, 해당 연구에서는 유클리드 거리를 거리 측정 방법을 사용함

  - SBERT는 샴 및 트리플릿 네트워크 구조를 사용하여 문장 간의 의미론적 유사성을 계산하며 이를 통해 문장 임베딩이 의미적으로 의미 있는 공간에 매핑된다는 특징이 있음. 또한, SBERT는 효율적인 풀링 전략과 유사도 계산을 사용하여 계산적으로 효율적이며, 대규모 데이터셋에서 빠른 속도로 작동함. SBERT는 100개 이상의 언어에 대한 문장 및 텍스트 임베딩을 제공하며, 다양한 언어에서도 우수한 성능을 보여줌. SBERT는 세밀한 조정을 통해 특정 작업에 맞는 임베딩 모델을 생성할 수 있음. 분석자의 데이터에 맞게 모델을 조정하고 성능을 향상시킬 수 있음

- SBERT는 KeyBERT와 마찬가지로 사전 훈련된 모델이므로 초기에 사전 훈련된 BERT 모델에 의존함. 새로운 데이터에 대해 SBERT를 적용하려면 BERT 모델을 미리 훈련할 필요가 있음. SBERT와 같은 학습 기반의 모델 성능은 사용되는 데이터의 양과 다양성에 따라 달라질 수 있음. 충분한 양과 다양성의 데이터가 필요할 수 있음. BERT 기반 모델은 입력 문장의 길이에 제한이 있을 수 있으며, 긴 문장에 대한 임베딩 생성에 어려움을 겪을 수 있음

- SBERT는 Python에서 사용할 수 있는 sentence-transformers패키지를 설치하여 활용할 수 있음 (https://www.sbert.net/index.html). 아래 그림은 문장 임베딩 예시임

Installation

You can install it using pip:

```
pip install -U sentence-transformers
```

We recommend **Python 3.6** or higher, and at least **PyTorch 1.6.0**. See installation for further installation options, especially if you want to use a GPU.

Usage

The usage is as simple as:

```
from sentence_transformers import SentenceTransformer
model = SentenceTransformer('all-MiniLM-L6-v2')

#Our sentences we like to encode
sentences = ['This framework generates embeddings for each input sentence',
    'Sentences are passed as a list of string.',
    'The quick brown fox jumps over the lazy dog.']

#Sentences are encoded by calling model.encode()
embeddings = model.encode(sentences)

#Print the embeddings
for sentence, embedding in zip(sentences, embeddings):
    print("Sentence:", sentence)
    print("Embedding:", embedding)
    print("")
```

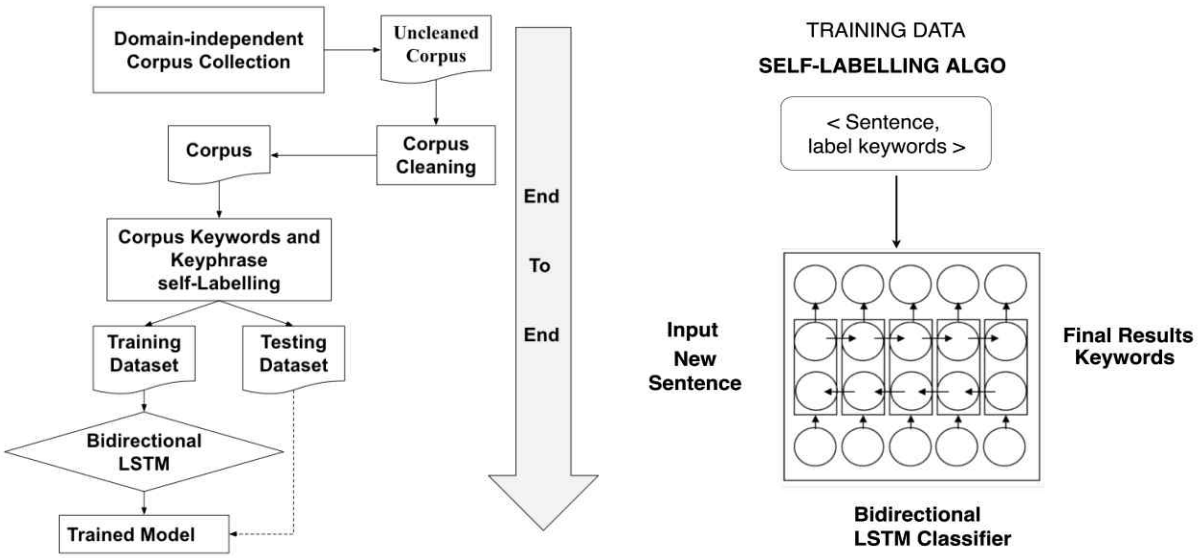
[그림 3-2] 파이썬을 통한 SBERT 활용 예시 : 문장 임베딩

2. KeyBERT

- KeyBERT는 텍스트의 핵심 주제를 추출하기 위해 개발된 Python 라이브러리임. 자연어 처리 작업에 강력한 성능을 발휘하는 사전 훈련 언어 모델인 BERT (Bidirectional Encoder Representations from Transformers)를 기반으로 구축됨. KeyBERT는 특히 텍스트 요약, 주제 추출, 검색 엔진 및 문서 분류 등의 작업에 활용될 수 있음. 이를 통해 텍스트의 핵심 주제를 자동으로 식별하고, 관련 있는 문장을 선택하여 중요한 정보를 추출할 수 있음

- KeyBERT는 텍스트 문서를 입력으로 받아 문서 내의 핵심 주제를 추출함. 이를 위해 텍스트를 BERT 모델에 입력시키고, 문장의 임베딩을 얻음. 이 후 문장 임베딩 간의 유사도를 계산한 후 중요한 문장을 선택하여 핵심 주제를 형성함. 해당 모델은 문서 내에서 중요한 문장을 식별하는 데 도움이 되는 키워드 추출을 수행함

- KeyBERT는 Sharma and Li (2019)의 연구에서 제안된 모델 아키텍처를 기반으로 하며, 해당 모델은 자가 레이블링을 이용한 자기 감독적 맥락 기반 키워드 및 키구문 검색을 위한 엔드투엔드 접근을 제안함. 이는 자가 레이블링을 통해 레이블이 없는 말뭉치를 활용하여 수동 노력을 줄이고 맥락 기반 키워드 추출을 수행함. 제안된 프로세스는 아래 그림(좌)와 같으며, 1) 도메인에 독립적인 말뭉치 수집, 2) 말뭉치 정리, 3) 말뭉치 자가 레이블링, 4) 양방향 LSTM을 사용한 키워드 추출 모델 훈련 등을 포함한 프로세스로 구성됨. 자가 레이블링 단계에서는 아래 그림(우)와 같이 양방향 Transformer 인코더를 활용하여 텍스트에서 맥락적인 특징을 추출하며, 이를 통해 기존의 Rapid Automatic Keyword Extraction (RAKE)나 TextRank 방법보다 더 우수한 키워드 레이블을 얻을 수 있음을 입증함



[그림 3-3] KeyBERT 모델구조(좌) 및 biLSTM 기반 키워드 추출 학습 과정도(우)

- KeyBERT의 장점은 사용하기 쉽고 간단한 인터페이스를 제공하고 있어 pip를 통해 간단하게 설치하여 키워드 추출 작업을 수행할 수 있음. 또한 자연어 처리 작업에서 뛰어난 성능을 발휘하는 BERT 모델을 기반의 모델로 효과적인 키워드 및 키구문 추출을 제공할 수 있음. KeyBERT는 BERT의 문맥 임베딩을 활용하여 텍스트의 문맥적 의미를 포착함. 이를 통해 문서의 핵심 주제와 관련된 키워드를 정확하게 식별할 수 있음

- KeyBERT의 단점 및 이슈사항에는 계산 비용, 높은 메모리, 학습 데이터의 종속성 등이 있음. BERT는 용량이 큰 언어 모델이기 때문에 모델의 크기와 복잡성으로 인해 계산 비용이 많이 소요될 수 있음. 특히, 대규모 텍스트 데이터에 대한 키워드 추출 작업에서는 추가적인 계산 리소스가 필요할 수 있음. 또한, BERT 모델은 큰 양의 메모리를 필요로 하므로 메모리 제약이 있는 환경에서는 사용하기 어려울 수 있음. KeyBERT는 사전훈련된 BERT 모델을 기반으로 하기 때문에 초기 학습 데이터에 의존하는 경향이 있음. 따라서 특정 도메인이나

언어에 대한 키워드 추출 작업에는 도메인 특화된 학습 데이터가 필요할 수 있음. 이는 사전훈련 모델의 공통적인 이슈 사항임

- KeyBERT는 Python에서 사용할 수 있는 KeyBERT 패키지를 설치하여 활용할 수 있음 (https://maartengr.github.io/KeyBERT/). 아래 그림은 키워드 추출 예시임

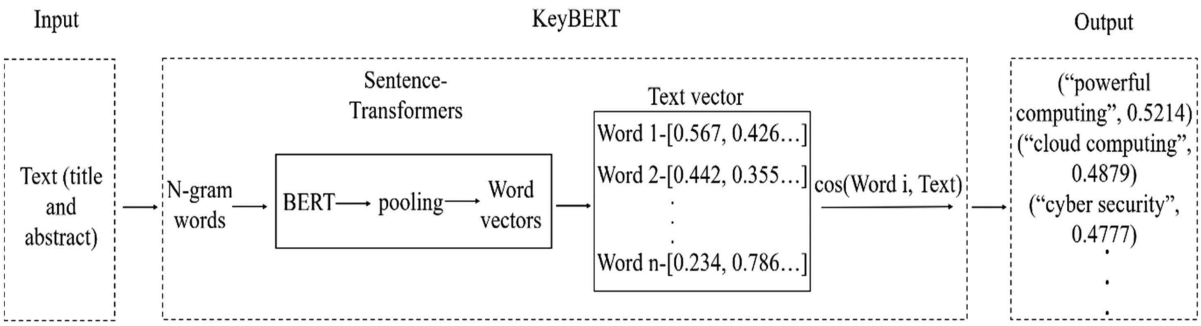
```
from keybert import KeyBERT

doc = """
        Supervised learning is the machine learning task of learning a function that
        maps an input to an output based on example input-output pairs. It infers a
        function from labeled training data consisting of a set of training examples.
        In supervised learning, each example is a pair consisting of an input object
        (typically a vector) and a desired output value (also called the supervisory signal).
        A supervised learning algorithm analyzes the training data and produces an inferred function,
        which can be used for mapping new examples. An optimal scenario will allow for the
        algorithm to correctly determine the class labels for unseen instances. This requires
        the learning algorithm to generalize from the training data to unseen situations in a
        'reasonable' way (see inductive bias).
    """
kw_model = KeyBERT()
keywords = kw_model.extract_keywords(doc)
```

You can set `keyphrase_ngram_range` to set the length of the resulting keywords/keyphrases:

```
>>> kw_model.extract_keywords(doc, keyphrase_ngram_range=(1, 1), stop_words=None)
[('learning', 0.4604),
 ('algorithm', 0.4556),
 ('training', 0.4487),
 ('class', 0.4086),
 ('mapping', 0.3700)]
```

[그림 3-4] 파이썬을 활용한 KeyBERT 모델링 : 키워드 추출



자료: Trappey et al., 2023

[그림 3-5] KeyBERT를 활용한 특허 키워드 추출 프로세스 예시

- 성과지표설정지원 서비스에 적합한 모델 등에 대한 문헌연구를 통해 간결한 문서의 맥락을 이해하고, 유사도 비교에 적합한 SBERT, KeyBERT 모델 활용을 계획함
  - 딥러닝 모델인 SBERT는 Sentence-BERT의 약자로 기존 BERT 모델의 문장 임베딩 성능을 개선하여 더 우수한 모델로 발전시킨 모델이며 학습하는데 시간이 빠르고 성능도 좋다는 장점을 가지고 있음
  - 다만, 기본적으로 문장단위로 임베딩을 하고 있기 때문에 연구과제의 전체 문서를 임베딩하여 비교하는 것은 한계가 있어 문서의 사업 목적 등 대표하는 문장 등 일부 활용도가 높은 부분만 임베딩하여 활용이 필요함
  - KeyBERT 모델은 각 사업의 대표되는 핵심 키워드의 리스트가 구성되고 특정 단어를 검색했을 때 연결된 성과지표를 추천하거나 특정 단어와 의미적으로 유사한 핵심키워드를 가진 사업을 추천하는 방식으로 시스템에 적용이 필요함
- 본 연구에서는 카카오브레인의 KorNLU데이터셋으로 파인튜닝되어 한국어 맥락 해석에 보다 적합한 Ko-sentence-transformer 모델을 활용함
- 유사도 계산 방식은 선행연구와 마찬가지로 전체 문서를 SBERT로 임베딩 후 키워드 추출한 다음, 키워드를 다시 임베딩하여 키워드와 문서 벡터간 코사인 유사도를 구하는 방식을 활용함
- Maximal Marginal Relevance를 통해 전략계획서 내용을 대표할 수 있는 다양한 키워드를 추출하고자 모델을 일부 조정함

[표 3-4] 성과지표 설정지원 모델 활용계획 개요

SBERT
◆ **(개요)** 의미론적으로 유사한 문장 매칭 ◆ **(장점)** 대규모 유사성 비교 및 클러스터링에 탁월/ 연산 속도가 빠른 편 ◆ **(단점)** 많은 양의 데이터로 사전학습 필요 ◆ **(활용방향)** 여러개의 어절로 구성된 사업명, 사업목적, 사업내용 등 기반 유사도 분석시에 활용하되, 맥락해석에 필요한 사전학습모델(예:KorNLU) 등을 활용하여 파인튜닝
KeyBERT
◆ **(개요)** 추출 키워드와 유사한 문서 매칭 ◆ **(장점)** 간단한 문서의 키워드 라벨링 및 문서 그룹핑에 탁월 ◆ **(단점)** 문서용량이 커질수록 리소스가 많이 들어 비효율적 ◆ **(활용방향)** 문서의 키워드를 추출해 유사도를 분석할 경우 활용하되, MMR 등 모델 조정

3. 실험을 통한 모델 보완사항 도출[56)]

(1) 분석개요

- 사업별 전략계획서를 비교하여 유사한 사업을 제안하는 알고리즘을 기본 전제로 하여 문장 간 비교를 통해 관련사업을 제안하는 SBERT 방식과 키워드를 추출한 후 키워드 간 비교를 통해 관련사업을 제안하는 KeyBERT 방식에 대한 각각의 분석을 진행하였음

- SBERT 방식과 KeyBERT 방식을 다양한 샘플사업을 선정하여 관련사업을 분석하도록 하고, 상위 빈도수의 사업들과 키워드 등을 반복적으로 비교 분석하여 반복실험에 따른 유사도 평균, 유효한 diversity 수치, 유효한 유사도 검색기준 등을 선별하고자 하였음

(2) 데이터 구성 및 분석과정

- 데이터는 전략계획서 PDF를 txt 파일로 변환한 후 특수문자, 목차어 제거 등 전처리를 거친 후에 연구에 사용할 데이터 (세부사업명, 사업목적, 전략목표, 세부내용, 통합내용)을 각각 추출하여 구축함

(3) SBERT 샘플사업 3개 유사도 분석 실험

- SBERT 분석은 텍스트 간의 유사성을 탐색하고, 관련사업들 간의 연결고리를 찾기 위해 각 지표(세부사업명, 사업목적, 전략목표, 세부내용, 통합내용)를 근거로 각 지표에 대해 유사한 사업을 제안하는 방법을 사용하였으며, 결과 또한 각각의 분석 옵션별로 도출함

- 각 지표별로 유사한 텍스트 데이터를 추출하기 위해 SBERT(Sentence-BERT) 모델을 활용하여 파이썬 코드를 작성하고, 지표별 가장 유사한 3개의 사업과 유사도를 추출함

- 시범분석을 위해 전략계획서 중 사업의 목적과 추진내용이 사업명에 비교적 잘 드러나있는 3개의 사업을 의도적으로 샘플링하여 실험을 진행하였으며, 3개 사업은 '산업기술국제협력, 수소트럭전기동력부품국산화기술개발, 환경시설재난재해대응기술개발'임

- 실험결과 분석을 통해 유사도 기준항목에 따라 관련사업 결과가 다르게 나타났으며, 각 내용이 상이한 점을 고려하여 실험과 마찬가지로 검색옵션이 여러개 필요하다는 시사점을 도출할 수 있었음

56) 아래의 각 실험은 다양한 세부사업을 대상으로 반복적으로 진행되었으며, 그 과정에서 추가적인 불용어 제거와 결과 비교, 코드 보완/수정 등을 거쳐 최종 조정된 결과를 본 보고서에 포함함

[표 3-5] SBERT 샘플사업 3개 유사도분석 결과 요약

대상사업명	사업명 기준 관련사업 (유사도)	사업목적 기준 관련사업 (유사도)	전략목표 기준 관련사업 (유사도)	사업 세부내용 기준 관련사업 (유사도)	통합내용 기준 관련사업 (유사도)
산업기술 국제협력	기후기술국제 협력촉진 (79.8%)	산학연Collabo R&D (81.6%)	산학연플랫폼 협력기술개발 (77.3%)	해외원천기술 상용화기술개발 R&D (74.2%)	해외우수기관 협력허브구축 (84.2%)
	산학연플랫폼 협력기술개발 (72.4%)	산학연협력 활성화지원사업 (80.5%)	해외우수연구 기관유치 (74.8%)	소재부품글로벌 투자연계 기술개발R&D (72.3%)	소재부품글로벌 투자연계 기술개발R&D (83.8%)
	산업혁신 기반구축 (70.6%)	산업혁신 기반구축 (77.6%)	월드클래스플러스 프로젝트지원 (73.3%)	정보통신방송기술 국제공동연구 (72.1%)	월드클래스플러스 프로젝트지원 (82.4%)
수소트럭 전기동력부품 국산화기술개발	수소트럭개조 기술개발및실증 (92.1%)	수소차용차세대 연료전지시스템 기술개발 (69.8%)	수소버스안전성 평가기술및장비 개발사업 (64.8%)	친환경동력원적용 농업기계기술개발 (68.0%)	수소트럭개조기술 개발및실증 (76.8%)
	수소에너지혁신 기술개발 (84.6%)	수소트럭개조 기술개발및실증 (65.6%)	수소트럭개조 기술개발및실증 (63.9%)	전기식건설기계용 충전인프라및기반 기술개발 (66.0%)	해외수소기반 대중교통인프라 기술개발 (71.0%)
	수소차용차세대 연료전지시스템 기술개발 (83.6%)	헬리콥터전기식 다중테일로터 기술개발 (65.6%)	수소선박안전 기술개발사업 (63.8%)	수소연료전지기반 민군겸용탑재중량 200kg급카고드론 기술개발 (64.1%)	온실가스감축을 위한SUV용 하이브리드시스템 고도화기술개발 (70.2%)
환경시설 재난재해대응 기술개발	자연재해대응영향 예보생산기술개발 (66.1%)	기후변화대응AI 기반풍수해위험도 예측기술개발 R&D (69.1%)	사회복합재난예방 대응기술개발 (70.2%)	사회복합재난예방 대응기술개발 (70.7%)	사회복합재난예방 대응기술개발 (76.1%)
	대기환경관리기술 사업화연계 기술개발(65.6%)	기후변화영향 최소화기술개발 (65.8%)	재난피해복구역량 강화기술개발 (67.0%)	기후변화대응AI 기반풍수해위험도 예측기술개발 R&D (69.6%)	자연재해대응영향 예보생산기술개발 (73.7%)
		자연재난정책 연계형기술개발 (63.3^%)	기후변화대응AI 기반풍수해위험도 예측기술개발 R&D (65.2%)	화학사고예측예방 고도화기술개발 사업 (69.0%)	자연재난정책 연계형기술개발 (73.6%)

[표 3-6] SBERT 샘플사업 3개 유사도분석 세부 결과(세부사업명, 사업목적 기준)

대상 사업	관련사업명	유사도	대상사업 목적	관련사업	관련사업목적	유사도
산업기술 국제협력	기후기술국제 협력촉진	0.798	국내 산학연과 해외 기관과의 공동연구 기술인력정보 교류 전략적 기술협력을 지원하여 우리 산업기업의 글로벌 혁신역량 강화 및 해외시장진출 지원	산학연Collabo R&D	산학연 협력RD 활성화를 통한 중소기업 혁신성장과 일자리 창출 산학협력기술개발 대학의 보유자원인력기술장비 등을 활용하여 연구인력 확보가 어려운 중소기업의 협력RD를 지원 산연협력기술개발 연구기관의 전문기술분야에 기반하여 중소기업의 혁신과 성장에 필요한 사업화 중심의 협력RD를 지원	0.816
	산학연플랫폼 협력기술개발	0.724		산학연협력활성화 지원사업	대학연구소의 기술사업화 인프라 및 혁신 역량을 활용하여 기업과의 협력을 통해 기술사업화 및 기술창업을 촉진하고 지역의 혁신성장과 일자리 창출 등 지역경제 활성화 유도	0.805
	산업혁신기반 구축	0.706		산업혁신기반구축	개별기업이 구축하기 힘들지만 산업기술개발에 필수적인 공동활용 인프라 구축 지원을 통해 중소기업의 산업기술혁신 역량을 제고	0.776
수소트럭 전기동력 부품 국산화 기술개발	수소트럭개조 기술개발및 실증	0.921	해외 수입 의존도가 높은 대형40Ton 초과 수소트럭의 핵심부품 400kW급 구동모터 인버터 변속기 등을 국산화하여 기술경쟁력 확보	수소차용차세대 연료전지시스템 기술개발	수소차 핵심기술의 기술력 확보 및 보급 확대를 위한 차세대 연료전지시스템 스택의 무게당 출력밀도kWkg 50 향상 및 풀스택 30 경량화 기술 개발 기존 넥쏘2세대 대비 고효율경량화 된 3세대 연료전지시스템 개발	0.698
	수소에너지 혁신기술개발	0.846		수소트럭개조 기술개발및실증	수소트럭 보급 활성화를 위하여 공공부문에서 사용되는 특장차량을 대상으로 수소동력 기반 특장차용 요소부품 및	0.656

대상 사업	관련사업명	유사도	대상사업 목적	관련사업	관련사업목적	유사도
					차량 시스템을 개발하고 차량 장착을 통한 검증 및 효율적 연계 운용방안 도출을 위한 실도로 실증기술을 개발	
	수소차용 차세대연료 전지시스템 기술개발	0.836		헬리콥터전기식 다중테일로터 기술개발	테일로터 전기동력시스템의 제조공정 기술 국산화를 통한 국제 기술경쟁력 강화 및 헬리콥터 기술 자립화 달성	0.656
환경시설 재난재해 대응기술 개발	자연재해대응 영향예보생산 기술개발	0.661	자연재난지진 태풍 등으로 인한 환경시설의 파괴 기능정지 등에 신속하게 대응하여이로 인한 2차 환경피해를 최소화하기 위한 재난관리 기술개발	기후변화대응AI 기반풍수해위험도 예측기술개발R&D	기후사회환경 변화 인한 풍수해 재난환경 변화에 대응 재난상황 및 피해예측 기술 개선 등 풍수해관리시스템 고도화 기술개발	0.691
	자연재해대응 영향예보생산 기술개발	0.661		기후변화영향 최소화기술개발	기후변화로 인한 사회경제적 영향 최소화를 위해 대기 수자원 등 환경관리 기술 개발	0.658
	대기환경관리 기술사업화연계기술개발	0.656		자연재난정책 연계형기술개발	재난안전관리 정책과 연계된 자연재난 2대 분야 관련 기술의 현장 적용성 강화 풍수해 폭염 관리정책	0.633

[표 3-7] SBERT 샘플사업 3개 유사도분석 세부 결과(세부사업명, 사업목적 기준)

대상사업	전략목표 관련사업	관련사업전략목표	유사도	세부내용 관련사업	유사도	통합내용 관련사업	유사도
산업기술국제협력	산학연플랫폼 협력기술개발	산학연 플랫폼 구축 및 협력 RD지원을 통한 중소기업의 연구개발 투자 촉진 및 기술경쟁력 확보	0.773	해외원천기술상용화 기술개발RD	0.742	해외우수 기관협력 허브구축	0.842
	해외우수연구 기관유치	해외우수연구기관과의 국제교류를 통한 선진 우수 기술 확보 및 과학기술 국제협력 네트워크 구축	0.748	소재부품글로벌투자 연계기술개발RD	0.723	소재부품 글로벌투자 연계기술 개발RD	0.838
	월드클래스 플러스프로젝트 지원	성장의지와 기술잠재력을 갖춘 중견중견후보기업에 대한 RD 및 사업화 집중지원을 통한 중견기업의 기술혁신역량 제고 및 신시장 진출 촉진	0.733	정보통신방송기술 국제공동연구	0.721	월드클래스 플러스 프로젝트 지원	0.824

대상사업	전략목표 관련사업	관련사업전략목표	유사도	세부내용 관련사업	유사도	통합내용 관련사업	유사도
수소트럭전기동력부품국산화기술개발	수소버스안전성평가기술및장비개발사업	수소버스 차량 및 수소 부품 단위의 안전성 평가 검사기술 및 장비 개발과 국제 안전기준 마련으로 수소버스 보급기반 활성화	0.648	친환경동력원적용 농업기계기술개발	0.680	수소트럭 개조기술 개발및실증	0.768
	수소트럭개조기술개발및실증	대형 수소트럭GVW 18톤급용 특장키트 및 특장키트차량의 장착성 기술개발실증 기술개발 대형 수소트럭GVW 18톤급용 ePTO연료전지 파워팩 및 특장 키트 4종 수소 특장트럭 5대 개조기술 개발 실증 대형 수소 특장트럭의 누적 10만km 운행실증을 통한 수소트럭 신뢰성 검증	0.639	전기식건설기계용 충전인프라및 기반기술개발	0.660	해외수소 기반 대중교통 인프라 기술개발	0.710
	수소선박안전 기술개발사업	수소 추진선박 및 운송선박 엑셀러레이터 안전기준 개발을 통한 수소선박 안전기반 조성	0.638	수소연료전지기반민군겸용탑재중량200kg급카고드론기술개발	0.641	온실가스 감축을위한 SUV용 하이브리드 시스템 고도화 기술개발	0.702
환경시설재난재해대응기술개발	사회복합재난 예방대응 기술개발	사회복합재난 대응역량 확보를 위한 전주기 통합재난관리체계 구축	0.702	사회복합재난예방대응기술개발	0.707	사회복합 재난예방 대응 기술개발	0.761
	재난피해복구역량강화기술개발	현장 활용성을 확보한 재난관리자원 운영관리 기술 및 재난 피해 시 국가기반시설 기능 정상화 기술 개발을 통한 실용적 재난피해 복구 역량 제고	0.670	기후변화대응AI기반 풍수해위험도예측기술개발RD	0.696	자연재해 대응영향 예보생산 기술개발	0.737
	기후변화대응AI기반풍수해위험도예측기술개발RD	기후변화 이상기후로 인한 풍수해 재난환경 변화에 대응을 위한 풍수해위험도 예측기술 및 풍수해관리시스템 고도화	0.652	화학사고예측예방고도화기술개발사업	0.690	자연재난 정책연계형 기술개발	0.736

(4) KeyBERT 키워드 추출 실험

- KeyBERT 분석은 각 사업의 핵심 키워드들을 추출한 후 각 키워드 간 유사도를 계산하여 가장 유사한 사업을 제안해주는 방식으로 설계하였고 다양한 변수 (Diversity, 키워드 수, ngram, 검색어 수)에 따른 결과를 분석하여 최적의 결과를 찾음

- 사업별 통합내용을 바탕으로 후보 키워드를 추출하고 SBERT를 활용하여 후보 키워드와 통합내용을 각각 임베딩 한 후 유사도 계산을 통해 가장 유사한 키워드를 추출하는 파이썬 코드를 작성함

- Diversity는 0.1, 0.3, 0.5, 0.7, 1.0 다섯 단계로 설정하고 ngram은 1과 2 두 단계로 설정하고 키워드 수는 5개, 10개로 설정하여 각 실험을 진행함

- Diversity 0.5, 0,7, ngram 2, 키워드 10개로 변수를 설정하고 검색어를 1개부터 5개까지 넣었을 때 추출된 관련사업 결과를 비교하는 실험을 진행함

- 분석결과 diversity 조정, ngram 수 조정 등에 따라 추출되는 키워드가 사업별로 일관된 특성을 보이지 않아 서비스 제공 시 사용자가 조정을 통해 적절한 결과를 찾도록 하는 것이 필요하나, 개발소요 및 정보입력 비용을 고려하여 반복 실험을 통해 diversity 0.7 또는 0.5 두 개의 옵션을 제공대상으로 선별함

[표 3-8] KeyBERT diversity 조정에 따른 키워드 추출 결과 예시 (bi-gram/ 5개 키워드 기준)

대상사업명	Diversity (파라미터)	1	2	3	4	5
국가위성통합 운영시스템 개발사업	1.0	효율 위성	데이터 관리	사이버 공격	촬영	네트워크
	0.7	효율 위성	데이터 관리	사이버 공격	인공지능 기반	네트워크
	0.5	효율 위성	데이터 관리	인공지능 기반	장비 구축	계획 자동화
	0.3	효율 위성	통합 자동화	시스템 국가	장비 구축	데이터 시스템
	0.1	효율 위성	운영 시스템	자동화 영상	계획 자동화	데이터 시스템
지진·지진해일 ·화산감시응용 기술개발	1.0	지진계 활용	품질	자연	운동	정보 생산
	0.7	지진계 활용	자연	추구 미래	수요자 맞춤	단축 현장
	0.5	지진계 활용	구축 자연	분석 신기술	화산 분화	대응 기반
	0.3	지진계 활용	자동 분석	지진 분석	대응 기반	기반 지진
	0.1	지진계 활용	지진 분석	지진 정밀	기반 지진	지진동 예측

(5) KeyBERT 검색어 입력 후 관련사업 추천 실험

- KeyBERT 키워드 추출 실험을 바탕으로 모델의 작동과 분석원리를 검토한 후 키워드 그룹 입력에 따라 관련사업을 잘 추출해내는지에 대하여 실험을 통해 검증하고자 함

- 각 키워드 그룹은 사업목적과 지원대상이 사업내용에 영향을 미칠 수 있는 중소기업 키워드군, 인재양성군, 사업내용의 특수성이 높은 치매군 총 3가지 그룹으로 나누어 diversity를 0.7과 0.5로 두 버전의 입력실험을 진행하고 결과를 비교함

[표 3-9] 중소기업 키워드별 상위 10개 관련사업 추출 결과(diversity 0.7)

검색어수	관련사업명	유사도
중소기업	ICTRD혁신바우처지원사업예타	0.818
	중소기업지원선도연구기관협력기술개발	0.8178
	지역중소기업공동수요기술개발RD	0.8156
	산단대개조지역기업RD	0.766
	연구기반활용플러스	0.7292
	산학연플랫폼협력기술개발	0.7186
	중소기업기술혁신개발소부장회계	0.7147
	제조중소기업글로벌역량강화신규	0.707
	강소벤처형중견기업육성	0.7053
	중소기업상용화기술개발	0.7046
중소기업, 지원	지역중소기업공동수요기술개발RD	0.751
	월드클래스플러스프로젝트지원	0.7039
	중소기업상용화기술개발	0.6927
	산학연플랫폼협력기술개발	0.6827
	ICTRD혁신바우처지원사업예타	0.6749
	중소기업지원선도연구기관협력기술개발	0.6616
	강소벤처형중견기업육성	0.658
	연구기반활용플러스	0.6549
	건강기능식품개발지원사업	0.6459
	ICT미래시장최적화협업기술개발	0.6383
중소기업, 지원, 산업단지	중소기업상용화기술개발	0.6544
	지역특화산업육성	0.6381
	지역중소기업공동수요기술개발RD	0.6344
	월드클래스플러스프로젝트지원	0.6311
	연구기반활용플러스	0.6135
	산학융합지구조성사업	0.6097
	강소벤처형중견기업육성	0.5933
	ICTRD혁신바우처지원사업예타	0.588

검색어수	관련사업명	유사도
	월드클래스300프로젝트기술개발	0.584
	지역대표중견기업육성	0.5762
중소기업, 지원, 산업단지, 장비활용	연구기반활용플러스	0.7056
	중소기업상용화기술개발	0.6354
	지역특화산업육성	0.6112
	지역중소기업공동수요기술개발RD	0.5994
	나노제품성능안전평가기술개발및기업지원사업	0.5879
	중소기업상용화기술개발소부장회계	0.5813
	강소벤처형중견기업육성	0.5799
	월드클래스플러스프로젝트지원	0.5737
	제조데이터공동활용플랫폼기술개발	0.5693
	ICTRD혁신바우처지원사업예타	0.565
중소기업, 지원, 산업단지, 장비활용, 일자리	연구기반활용플러스	0.6382
	중소기업상용화기술개발	0.585
	지역특화산업육성	0.5812
	지역중소기업공동수요기술개발RD	0.5467
	연구개발특구육성	0.5446
	중소기업상용화기술개발소부장회계	0.5439
	강소벤처형중견기업육성	0.5407
	월드클래스플러스프로젝트지원	0.5389
	제조데이터공동활용플랫폼기술개발	0.5368
	나노제품성능안전평가기술개발및기업지원사업	0.5352

[표 3-10] 중소기업 키워드별 상위 10개 관련사업 추출 결과(diversity 0.5)

검색어수	관련사업명	유사도
중소기업	ICTRD혁신바우처지원사업예타	0.7933
	지역중소기업공동수요기술개발RD	0.7847
	중소기업상용화기술개발소부장회계	0.7779
	성과공유형공통기술RD	0.7751
	제조중소기업글로벌역량강화신규	0.7736
	산학연플랫폼협력기술개발	0.7632
	중소기업지원선도연구기관협력기술개발	0.7529
	중소기업기술사업화역량강화	0.734
	연구기반활용플러스	0.7082
	중소기업상용화기술개발	0.7033

검색어수	관련사업명	유사도
중소기업, 지원	산학연플랫폼협력기술개발	0.8096
	지역중소기업공동수요기술개발RD	0.7783
	월드클래스플러스프로젝트지원	0.7526
	중소기업기술사업화역량강화	0.7418
	제조중소기업글로벌역량강화신규	0.7301
	중소기업상용화기술개발소부장회계	0.722
	창업성장기술개발RD기후대응기금	0.7154
	ICTRD혁신바우처지원사업예타	0.7154
	중소기업지원선도연구기관협력기술개발	0.6994
	중견기업상생혁신사업RD	0.6925
중소기업, 지원, 산업단지	지역대표중견기업육성	0.744
	지역특화산업육성	0.7143
	강소벤처형중견기업육성	0.6841
	산학연플랫폼협력기술개발	0.684
	창업성장기술개발RD기후대응기금	0.6691
	월드클래스플러스프로젝트지원	0.667
	ICTRD혁신바우처지원사업예타	0.6667
	연구장비산업육성	0.663
	중소기업상용화기술개발소부장회계	0.6543
	ESG형산단공동혁신지원사업균특	0.645
중소기업, 지원, 산업단지, 장비활용	연구기반활용플러스	0.7093
	연구장비산업육성	0.6556
	중소기업지원선도연구기관협력기술개발	0.626
	제조기술융합센터테스트베드	0.6207
	시스템산업거점기관지원	0.5825
	나노융합현장수요기반실증지원사업	0.5749
	지역중소기업공동수요기술개발RD	0.5743
	기업부설연구소RD역량강화지원	0.5695
	조선해양산업기술개발사업	0.5633
	기계장비산업기술개발	0.5576
중소기업, 지원, 산업단지, 장비활용, 일자리	연구개발특구육성	0.5761
	연구기반활용플러스	0.5646
	이공학학술연구기반구축	0.5643
	중소기업지원선도연구기관협력기술개발	0.5515
	지역대표중견기업육성	0.543
	실험실창업지원	0.5424
	중소기업상용화기술개발	0.5357
	지역중소기업공동수요기술개발RD	0.5331
	중견기업상생혁신사업RD	0.5327
	해양수산기술창업기업Scaleup지원사업	0.5309

[표 3-11] 인력양성 키워드별 상위 10개 관련사업 추출 결과(diversity 0.7)

검색어수	관련사업명	유사도
인재	융합형과학기술인재양성기반구축	0.7728
	농식품기술융합창의인재양성	0.6853
	보건의료인재양성지원사업국민건강증진기금	0.6489
	마이스터대지원RD	0.6384
	재난안전산업기술사업화지원RD	0.6147
	산단대개조지역기업RD	0.6016
	산림융복합전문인력양성	0.6008
	중견기업핵심연구인력성장지원사업RD	0.5978
	소재부품글로벌투자연계기술개발RD	0.5972
	정보통신방송혁신인재양성	0.5962
인재 교육	농식품기술융합창의인재양성	0.6741
	융합형과학기술인재양성기반구축	0.6733
	BK21플러스사업	0.624
	보건의료인재양성지원사업국민건강증진기금	0.5921
	마이스터대지원RD	0.5786
	문화콘텐츠RD전문인력양성	0.5744
	인공지능반도체응용기술개발	0.5689
	에듀테크RD지원사업	0.5608
	정보통신방송혁신인재양성	0.5587
	산림융복합전문인력양성	0.5523
인재 교육 연수	농식품기술융합창의인재양성	0.5701
	보건의료인재양성지원사업국민건강증진기금	0.5697
	융합형과학기술인재양성기반구축	0.5595
	정보통신방송혁신인재양성	0.5416
	규제과학인재양성사업	0.4967
	산업혁신인재성장지원	0.4967
	중소기업연구인력지원	0.496
	에듀테크RD지원사업	0.4917
	인공지능반도체응용기술개발	0.4904
	중견기업핵심연구인력성장지원사업RD	0.4745
인재 교육 연수 대학	전문대학미래기반조성RD	0.652
	BK21플러스사업	0.6227
	대학혁신지원	0.6156
	인공지능융합혁신인재양성	0.6145
	정보통신방송혁신인재양성	0.5354
	산학융합지구조성사업	0.5344

검색어수	관련사업명	유사도
	산업혁신인재성장지원	0.5302
	중소기업연구인력지원	0.5298
	집단연구지원	0.5238
	융합형과학기술인재양성기반구축	0.5113
인재 교육 연수 대학 참여	전문대학미래기반조성RD	0.624
	대학혁신지원	0.5946
	인공지능융합혁신인재양성	0.5943
	BK21플러스사업	0.5839
	정보통신방송혁신인재양성	0.5549
	산학융합지구조성사업	0.524
	산업혁신인재성장지원	0.5235
	중소기업연구인력지원	0.5198
	융합형과학기술인재양성기반구축	0.5037
	규제과학인재양성사업	0.499

[표 3-12] 인력양성 키워드별 상위 10개 관련사업 추출 결과(diversity 0.5)

검색어수	관련사업명	유사도
인재	ICTRD혁신바우처지원사업예타	0.7933
	지역중소기업공동수요기술개발RD	0.7847
	중소기업상용화기술개발소부장회계	0.7779
	성과공유형공통기술RD	0.7751
	제조중소기업글로벌역량강화신규	0.7736
	산학연플랫폼협력기술개발	0.7632
	중소기업지원선도연구기관협력기술개발	0.7529
	중소기업기술사업화역량강화	0.734
	연구기반활용플러스	0.7082
	중소기업상용화기술개발	0.7033
인재 교육	산학연플랫폼협력기술개발	0.8096
	지역중소기업공동수요기술개발RD	0.7783
	월드클래스플러스프로젝트지원	0.7526
	중소기업기술사업화역량강화	0.7418
	제조중소기업글로벌역량강화신규	0.7301
	중소기업상용화기술개발소부장회계	0.722
	창업성장기술개발RD기후대응기금	0.7154
	ICTRD혁신바우처지원사업예타	0.7154
	중소기업지원선도연구기관협력기술개발	0.6994
	중견기업상생혁신사업RD	0.6925

검색어수	관련사업명	유사도
인재 교육 연수	지역대표중견기업육성	0.744
	지역특화산업육성	0.7143
	강소벤처형중견기업육성	0.6841
	산학연플랫폼협력기술개발	0.684
	창업성장기술개발RD기후대응기금	0.6691
	월드클래스플러스프로젝트지원	0.667
	ICTRD혁신바우처지원사업예타	0.6667
	연구장비산업육성	0.663
	중소기업상용화기술개발소부장회계	0.6543
	ESG형산단공동혁신지원사업균특	0.645
인재 교육 연수 대학	연구기반활용플러스	0.7093
	연구장비산업육성	0.6556
	중소기업지원선도연구기관협력기술개발	0.626
	제조기술융합센터테스트베드	0.6207
	시스템산업거점기관지원	0.5825
	나노융합현장수요기반실증지원사업	0.5749
	지역중소기업공동수요기술개발RD	0.5743
	기업부설연구소RD역량강화지원	0.5695
	조선해양산업기술개발사업	0.5633
	기계장비산업기술개발	0.5576
인재 교육 연수 대학 참여	연구개발특구육성	0.5761
	연구기반활용플러스	0.5646
	이공학학술연구기반구축	0.5643
	중소기업지원선도연구기관협력기술개발	0.5515
	지역대표중견기업육성	0.543
	실험실창업지원	0.5424
	중소기업상용화기술개발	0.5357
	지역중소기업공동수요기술개발RD	0.5331
	중견기업상생혁신사업RD	0.5327
	해양수산기술창업기업Scaleup지원사업	0.5309

[표 3-13] 치매 키워드별 상위 10개 관련사업 추출 결과(diversity 0.7)

검색어수	관련사업명	유사도
치매	치매극복연구개발사업과기부	0.6907
	치매극복연구개발사업복지부	0.6907
	뇌질환극복연구사업	0.5795
	5G기반이동형유연의료플랫폼기술개발	0.4708
	미래뇌융합기술개발	0.4706
	국민생활안전긴급대응연구사업RAPID	0.4614
	감염병관리기술개발연구	0.4558
	보건의료인재양성지원사업국민건강증진기금	0.4524
	철도기술연구사업	0.4521
	만성병관리기술개발연구	0.4479
치매 뇌과학	뇌질환극복연구사업	0.6921
	치매극복연구개발사업과기부	0.6914
	치매극복연구개발사업복지부	0.6914
	미래뇌융합기술개발	0.6625
	혁신도전자폐혼합형디지털치료제개발	0.5393
	첨단의료기술개발	0.4975
	차세대지능형반도체기술개발소자	0.4945
	KMedi융합인재양성지원사업	0.4842
	인공지능활용혁신신약발굴	0.4823
	보건의료인재양성지원사업국민건강증진기금	0.4815
치매 뇌과학 뇌파	뇌질환극복연구사업	0.712
	치매극복연구개발사업과기부	0.6707
	치매극복연구개발사업복지부	0.6707
	미래뇌융합기술개발	0.6319
	차세대지능형반도체기술개발소자	0.5228
	혁신도전자폐혼합형디지털치료제개발	0.5206
	지진화산업무지원및활용기술개발	0.4682
	양자컴퓨팅기술개발사업	0.4624
	첨단방사선융합치료기술개발	0.4604
	국립재활원재활연구개발용역사업	0.4556
치매 뇌과학 뇌파 알츠하이머	뇌질환극복연구사업	0.6512
	치매극복연구개발사업과기부	0.6155
	치매극복연구개발사업복지부	0.6155
	미래뇌융합기술개발	0.5712
	혁신도전자폐혼합형디지털치료제개발	0.4727
	차세대지능형반도체기술개발소자	0.4523
	사람중심AI강국실현을위한차세대인공지능핵심원천기술개발	0.4056
	첨단방사선융합치료기술개발	0.4001
	인공지능활용혁신신약발굴	0.3905
	양자컴퓨팅기술개발사업	0.3896

검색어수	관련사업명	유사도
치매 뇌과학 뇌파 알츠하이머 인공지능	미래뇌융합기술개발	0.6931
	혁신도전자폐혼합형디지털치료제개발	0.6465
	뇌질환극복연구사업	0.6156
	영상진단의료기기탑재용AI기반영상분석솔루션개발	0.6073
	인공지능융합선도프로젝트	0.6002
	인공지능활용혁신신약발굴	0.5963
	한국어대형인공지능언어모델기술개발	0.5876
	첨단방사선융합치료기술개발	0.5848
	인공지능기반의건축설계자동화기술개발	0.5803
	비대면비즈니스디지털혁신기술개발	0.5745

[표 3-14] 치매 키워드별 상위 10개 관련사업 추출 결과(diversity 0.5)

검색어수	관련사업명	유사도
치매	치매극복연구개발사업복지부	0.8194
	치매극복연구개발사업과기부	0.8194
	뇌질환극복연구사업	0.6547
	뇌기능규명조절기술개발사업	0.5536
	국립정신건강센터연구개발사업	0.4778
	미래뇌융합기술개발	0.4702
	국민생활안전긴급대응연구사업RAPID	0.4606
	감염병관리기술개발연구	0.4558
	지역사회기반재활운동서비스기술개발	0.453
	소상공인자영업자를위한생활혁신형기술개발	0.4491
치매 뇌과학	치매극복연구개발사업복지부	0.8002
	치매극복연구개발사업과기부	0.8002
	뇌기능규명조절기술개발사업	0.7505
	뇌질환극복연구사업	0.7321
	미래뇌융합기술개발	0.67
	지역사회기반재활운동서비스기술개발	0.5123
	국립정신건강센터연구개발사업	0.5076
	문화유산스마트보존활용기술개발	0.4922
	빅데이터기반인공지능도시계획기술개발	0.4828
	인공지능활용혁신신약발굴	0.4823
치매 뇌과학 뇌파	뇌질환극복연구사업	0.7507
	뇌기능규명조절기술개발사업	0.7327
	치매극복연구개발사업과기부	0.726
	치매극복연구개발사업복지부	0.726
	미래뇌융합기술개발	0.6256
	지역사회기반재활운동서비스기술개발	0.5093
	극한지개발및탐사용협동이동체시스템기술개발사업RD	0.4893
	극한지개발및탐사용협동이동체시스템기술개발다부처	0.4893
	문화유산스마트보존활용기술개발	0.4892
	빅데이터기반인공지능도시계획기술개발	0.474

검색어수	관련사업명	유사도
치매 뇌과학 뇌파 알츠하이머	치매극복연구개발사업과기부	0.7296
	치매극복연구개발사업복지부	0.7296
	뇌질환극복연구사업	0.6815
	뇌기능규명조절기술개발사업	0.6795
	미래뇌융합기술개발	0.5718
	ICT기반사회문제해결기술개발사업	0.4345
	문화유산스마트보존활용기술개발	0.4285
	지역사회기반재활운동서비스기술개발	0.4214
	빅데이터기반생활전자파예측기술개발사업	0.409
	사람중심AI강국실현을위한차세대인공지능핵심원천기술개발	0.4056
치매 뇌과학 뇌파 알츠하이머 인공지능	미래뇌융합기술개발	0.6874
	뇌기능규명조절기술개발사업	0.6871
	치매극복연구개발사업복지부	0.6797
	치매극복연구개발사업과기부	0.6797
	ICT기반사회문제해결기술개발사업	0.635
	뇌질환극복연구사업	0.6264
	한국어대형인공지능언어모델기술개발	0.6085
	영상진단의료기기탑재용AI기반영상분석솔루션개발	0.6073
	빅데이터기반인공지능도시계획기술개발	0.6069
	지역사회기반재활운동서비스기술개발	0.6044

- 세 차례의 키워드 검색실험 결과, 키워드 그룹의 특성에 따라 결과가 달라지는 것을 확인할 수 있었으며, diversity 값을 0.7에서 0.5로 조정하는 것으로도 유사사업 추천결과와 유사도가 달라지는 것을 확인할 수 있었음

- 다만, 유사사업 추천 결과에 대한 활용도는 사업정보나 기술분야의 도메인 지식이 바탕이 되어야 한다는 점에서 diversity 기준치는 두 개의 옵션을 모두 제공해야 할 것으로 보임

(6) 사업별 SBERT, KeyBERT 기반 관련사업 추출 실험

- 앞서 진행된 실험들을 바탕으로 각 부처별 샘플사업들의 SBERT 기준 관련사업 3개와 KeyBERT 기반 관련사업 3개를 추출하고 단어 빈도수 높은 키워드와 워드클라우드 결과를 도출하여 비교해봄으로써 두 모델의 추천 결과가 유의미하게 다른지 살펴봄

- 샘플 사업은 '과학기술정보통신부의 ICT기반개방형혁신서비스개발사업', '농림축산식품부의 고부가가치식품기술개발', '보건복지부의 감염병방역기술개발', '산업통상자원부의 가스터빈 부품제조기업기술역량강화및품질/신뢰성지원인프라구축기술개발사업(R&D)', '해양수산부의 스마트항만-자율운항선박연계기술개발' 5개를 정하여 분석에 활용함

[표 3-15] 샘플사업에 대한 SBERT/KeyBERT 추천결과 비교

부처 / 사업	구분	관련사업 부처	관련사업명	핵심키워드 (KeyBERT)	유사도
과기정통부 ICT기반개방형 혁신서비스개발사업 (R&D)	SBERT1	중기부	리빙랩활용기술개발사업	-	0.835
	SBERT2	과기부	공공연구성과 가치창출기술키움	-	0.83
	SBERT3	과기	연구개발특구육성	-	0.826
	KEYBERT1	행안부	국민수요맞춤형생활안전 연구개발사업	혁명 신기술,국민 생활,안전 분야,요소 발굴,아이디어 기반,현장 실증,기반 재난,생활 적용,국민 투표,수요 맞춤	0.792
	KEYBERT2	중기부	리빙랩활용기술개발사업	실증 지원,활용 사용자,반영 시장,단계 참여,주민 제품,지원 방식,사업 성과,최대 정부,발비 비중,지원 연구개	0.785
	KEYBERT3	국토부	빅데이터기반인공지능도시계획기술개발	인공지능 도시,진단 사업,계획 데이터,이용 기반,도시 관리,사회 변화,수요자 중심,구축 토지,지원 서비스,실증 데이터	0.759

[표 3-16] SBERT/KeyBERT 관련사업의 키워드 순위 비교 (빈도수 기준)

구분	SBERT1		SBERT2		SBERT3		KEYBERT1		KEYBERT2		KEYBERT3	
키워드 빈도 순위	리빙랩활용기술 개발사업		공공연구성과 가치창출기술키움		연구개발특구육성		국민수요맞춤형 생활안전 연구개발사업		리빙랩활용기술 개발사업		빅데이터기반 인공지능 도시계획기술개발	
1	지원	6	키우다	6	지원	22	국민	9	지원	6	도시	16
2	활용	4	실용	6	사업	18	생활	8	활용	4	계획	15
3	사업	3	사업	5	특구	13	안전	8	사업	3	기반	8
4	소비자	3	시장	5	창업	10	발굴	4	소비자	3	데이터	7
5	수요	2	지원	4	실증	9	수요	3	수요	2	수립	7
6	반영	2	검증	4	지역	8	맞춤	3	반영	2	지원	5
7	실증	2	후속	4	육성	7	아이디어	3	실증	2	진단	4
8	참여	2	기업	4	신기술	7	재난	3	참여	2	사업	3
9	제품	2	가능	3	통하다	6	요소	3	제품	2	서비스	3
10	지정	2	상용	3	위하다	6	사업	2	지정	2	인공지능	2

- 결과적으로 사업의 특징에 따라 두 모델을 활용했을 때 관련사업 추천 결과가 달라질 수도, 동일할 수도 있음을 확인함

- 본 실험을 통해 두 개의 모델과 검색항목, 파라미터값과 같은 검색옵션을 다양하게 제공하는 것이 필요하며, 필요에 따라 두 개의 결과를 모두 참조할 수 있도록 기획하는 것이 필요함을 알 수 있음

리빙랩활용기술개발사업

공공연구성과가치창출기술키움

연구개발특구육성

국민수요맞춤형생활안전연구개발사업

빅데이터기반인공지능도시계획기술개발

리빙랩활용기술개발사업

[그림 3-6] 샘플사업에 대한 키워드 클라우드 (uni-gram)

[표 3-17] 샘플사업에 대한 SBERT/KeyBERT 추천결과 비교

부처 / 사업	구분	관련사업 부처	관련사업명	핵심키워드 (KeyBERT)	유사도
농림부 고부가가치식품 기술개발	SBERT1	중기부	건강기능식품 개발지원사업	-	0.806
	SBERT2	농진청	작물시험연구	-	0.794
	SBERT3	농림부	농식품기술 융합창의인재양성	-	0.786
	KEYBERT1	농진청	농업과학기반기술연구	농업 미생물,자원 다양성,지원 기반,생태 데이터,과학 기반,안전 건강,생산 과정,활용 부가가치,농업 기반,기계화 자동화	0.79
	KEYBERT2	농진청	바이오그린연계 농생명혁신기술개발	육종 기반,생명 첨단,표지 형질,확산 국가,기반 작물,매몰 방지,성장 잠재력,생산 시스템,세포 장형,가능 분자	0.771
	KEYBERT3	복지부	재생의료임상연구 기반조성	구축 세포,의약품 안전,생산 공정,활용 지원,조성 첨단,은행 특성,선택 기반,지원 임상시험,소재 바이오,벡터 생산	0.747

[표 3-18] SBERT / KeyBERT 관련사업의 키워드 순위 비교 (빈도수 기준)

구분	SBERT1		SBERT2		SBERT3		KEYBERT1		KEYBERT2		KEYBERT3	
키워드 빈도 순위	건강기능식품개발 지원사업		작물시험연구		농식품기술융합창의인재양성		농업과학기반기술 연구		바이오그린연계농생명혁신기술개발		재생의료임상연구 기반조성	
1	지원	9	목표	144	인력	17	농업	18	생명	11	의료	17
2	기능	5	품종	135	식품	16	기반	8	육종	10	재생	15
3	건강	3	작물	97	현장	12	안전	6	소재	10	임상	14
4	식품	3	보급	97	기관	11	구축	5	분자	7	지원	14
5	중소기업	3	식량	88	지원	10	관리	5	기반	7	세포	13
6	주기	3	성과	73	사업	9	식품	5	공학	6	첨단	8
7	원료	3	실적	64	양성	8	자원	5	활용	6	생산	7
8	위하다	3	활용	61	위하다	8	생물	5	바이오	4	바이러스	7
9	임상	3	지표	60	분야	8	과학	3	통하다	4	소재	6
10	사업	2	지수	58	인재	7	생산	3	농업	4	벡터	6

건강기능식품개발지원사업

작물시험연구

농식품기술융합창의인재양성

농업과학기반기술연구

바이오그린연계농생명혁신기술개발

재생의료임상연구기반조성

[그림 3-7] 샘플사업에 대한 키워드 클라우드 (uni-gram)

[표 3-19] 샘플사업에 대한 SBERT/KeyBERT 추천결과 비교

부처 / 사업	구분	관련사업 부처	관련사업명	핵심키워드 (KeyBERT)	유사도
복지부 감염병방역기술개발	SBERT1	질병청	신기술기반백신플랫폼 개발지원사업	-	0.794
	SBERT2	복지부	감염병위기대응기술개발	-	0.793
	SBERT3	산업부	백신원부자재생산 고도화기술개발	-	0.778
	KEYBERT1	중기부	현장수요맞춤형 방역물품기술개발	중소기업 방역,사항 반영,기업 창의성,사업 성과,지원 기술력,현장 애로,물품 기기,수요 기업,창출 현장,투입 의료진	0.756
	KEYBERT2	복지부	감염병위기대응기술개발	감염병 대응,국산 개량,국가 차원,백신 신약,발생 국민,데이터 활용,접종 국가,보건의료기술,구축 확산,대비 고도	0.732
	KEYBERT3	복지부	범부처전주기의료기기 연구개발(복지부)	기기 의료,화형 글로벌,복지 구현,기반 확보,플랫폼 내역,지역 격차,단계 임상시험,복지부 부처,제품 국산,부처 과기	0.73

[표 3-20] SBERT / KeyBERT 관련사업의 키워드 순위 비교 (빈도수 기준)

구분	SBERT1		SBERT2		SBERT3		KEYBERT1		KEYBERT2		KEYBERT3	
키워드 빈도 순위	신기술기반백신 플랫폼개발 지원사업		감염병위기대응 기술개발		백신원부자재생산 고도화기술개발		현장수요맞춤형 방역물품기술개발		감염병위기대응 기술개발		범부처전주기 의료기기연구개발 (복지부)	
1	위하다	8	감염병	15	생산	17	방역	8	감염병	15	의료	23
2	백신	7	백신	9	백신	15	물품	6	백신	9	기기	17
3	지원	6	대응	8	위하다	5	현장	4	대응	8	위하다	7
4	플랫폼	4	국가	7	원부	4	수요	3	국가	7	시장	6
5	대응	4	위기	4	자재	4	중소기업	3	위기	4	지원	5
6	효능	4	국민	4	제조	4	기기	3	국민	4	구현	5
7	감염병	3	강화	4	경쟁력	4	지원	3	강화	4	제품	4
8	코로나	2	위하다	3	소재	4	맞춤	2	위하다	3	미래	4
9	신속	2	국산	3	공정	4	감염병	2	국산	3	내역	4
10	가능	2	지원	3	고도	3	대응	2	지원	3	중점	4

신기술기반백신플랫폼개발지원사업

감염병위기대응기술개발

백신원부자재생산고도화기술개발

현장수요맞춤형방역물품기술개발

감염병위기대응기술개발

범부처전주기의료기기연구개발 (복지부)

[그림 3-8] 샘플사업에 대한 키워드 클라우드 (uni-gram)

[표 3-21] 샘플사업에 대한 SBERT/KeyBERT 추천결과 비교

부처 / 사업	구분	관련사업 부처	관련사업명	핵심키워드 (KeyBERT)	유사도
산업부 가스터빈부품제조기업 기술역량강화및품질/신뢰성지원인프라구축 기술개발사업(R&D)	SBERT1	산업부	소재부품산업기술 개발기반구축(R&D)	-	0.828
	SBERT2	산업부	LNG발전용가스터빈 고온부품성능검증 기술개발	-	0.806
	SBERT3	산업부	원전안전부품경쟁력강화 기술개발	-	0.801
	KEYBERT1	산업부	LNG발전용가스터빈 고온부품성능검증 기술개발	가스 터빈,기반 구축,부품 제조,블레이드 엔지니어링,운전 특화,적용 효율,중소기업 경쟁력,고온,특화 가스,국내 발전	0.846
	KEYBERT2	산업부	저열화성노후전력기자재 재제조기술개발	전력 기자재,인증 기반,화성 노후,제조 최적,상용 사업,기업 지원,잔존 수명,부합 스펙,성능 안전,확산 기반	0.739
	KEYBERT3	산업부	표준가스복합화력시스템 및TestBed구축 기술개발사업	표준가스복합화력시스템및testbed구축기술개발사업 국내,복합 효율,표준화 모델,발전 플랜트,성장 동력,표준 가스,설계 제작,초초임계압,국내 최적화,친환경 설비	0.734

[표 3-22] SBERT / KeyBERT 관련사업의 키워드 순위 비교 (빈도수 기준)

구분	SBERT1		SBERT2		SBERT3		KEYBERT1		KEYBERT2		KEYBERT3	
키워드 빈도 순위	소재부품산업기술 개발기반구축 (R&D)		LNG발전용가스 터빈고온부품 성능검증기술개발		원전안전부품 경쟁력강화 기술개발		LNG발전용가스 터빈고온부품성능 검증기술개발		저열화성노후전력 기자재재제조기술 개발(신규추가)		표준가스복합화력 시스템및 TestBed구축 기술개발사업	
1	소재	16	터빈	8	원전	9	터빈	8	제조	9	복합	7
2	부품	16	가스	7	안전	7	가스	7	전력	7	표준	6
3	지원	14	부품	6	부품	5	부품	6	기자재	7	가스	5
4	장비	12	고온	5	강화	4	고온	5	노후	5	발전	4
5	구축	9	발전	4	기자재	4	발전	4	기반	5	설계	3
6	기업	9	성능	3	경쟁력	3	성능	3	성능	4	플랜트	3
7	기반	7	검증	2	관련	3	검증	2	사업	3	최적화	2
8	활용	6	통하다	2	지원	3	통하다	2	고도	2	성능	2
9	분야	5	기반	2	중소	3	기반	2	설비	2	기준	2
10	인프라	5	구축	2	중견기업	3	구축	2	부합	2	기기	2

소재부품산업기술개발기반구축(R&D)

LNG발전용가스터빈고온부품성능검증기술개발

원전안전부품경쟁력강화기술개발

LNG발전용가스터빈고온부품성능검증기술개발

저열화성노후전력기자재제조기술개발신규추가

표준가스복합화력시스템 및
TestBed구축기술개발사업

[그림 3-9] 샘플사업에 대한 키워드 클라우드 (uni-gram)

[표 3-23] 샘플사업에 대한 SBERT/KeyBERT 추천결과 비교

부처 / 사업	구분	관련사업 부처	관련사업명	핵심키워드 (KeyBERT)	유사도
해수부 스마트항만-자율운항 선박연계기술개발	SBERT1	해수부	자율운항선박기술개발	-	0.806
	SBERT2	해수부	수출입자율주행차량자동 하역지원시스템기술개발	-	0.786
	SBERT3	해수부	스마트자동화항만상용화 기술개발	-	0.772
	KEYBERT1	해수부	초연결디지털해상물류 통합성능검증테스트베드 기술개발	디지털 항만,성능 검증,해상 최적,구축 필요,신기술 장비,지능,시뮬레이션 가상,서비스 통합,시뮬레이션 항만,구축 해양	0.837
	KEYBERT2	해수부	수출입자율주행차량자동 하역지원시스템기술개발	지능 하역,주행 차량,기반 마련,관리 플랫폼,지능 플랫폼,효율 증대,항만 인프라,수출입 자율,스마트 해상,물류 기반	0.828
	KEYBERT3	해수부	항만내환적화물자동운송 시스템(무인트램)개발	스마트 항만,효율 생산성,스테이션 방식,제작 적재,셔틀 제어,기반 시설,시스템 추진,차량 복합,터미널,트윈로더 방식	0.792

[표 3-24] SBERT / KeyBERT 관련사업의 키워드 순위 비교 (빈도수 기준)

구분	SBERT1		SBERT2		SBERT3		KEYBERT1		KEYBERT2		KEYBERT3	
키워드 빈도 순위	자율운항선박 기술개발		수출입자율주행 차량자동하역지원 시스템기술개발		스마트자동화항만 상용화기술개발		초연결디지털해상 물류통합성능검증 테스트베드 기술개발		수출입자율주행 차량자동하역지원 시스템기술개발		항만내환적화물 자동운송시스템 (무인트램)개발	
1	자율	12	하역	6	항만	10	디지털	7	하역	6	항만	10
2	운항	12	수출입	5	스마트	9	통합	6	수출입	5	화물	8
3	선박	6	자율	5	자동화	9	성능	6	자율	5	설계	8
4	시스템	5	주행	5	상용	4	검증	6	주행	5	셔틀	7
5	실증	4	차량	5	시스템	4	해상	5	차량	5	운송	6
6	운용	3	자동	3	생산성	3	물류	3	자동	3	제작	6
7	국제	3	지원	3	테스트	3	센터	3	지원	3	자동	4
8	지능	2	시스템	3	베드	3	구축	3	시스템	3	시스템	4
9	자동화	2	관리	3	컨테이너	2	시뮬레이션	3	관리	3	확보	4
10	통하다	2	자동차	2	하역	2	항만	3	자동차	2	적재	4

자율운항선박기술개발

수출입자율주행차량자동하역지원시스템기술개발

스마트자동화항만상용화기술개발

초연결디지털해상물류통합성능검증테스트베드기술개발

수출입자율주행차량자동하역지원시스템기술개발

항만내환적화물자동운송시스템(무인트램)개발

[그림 3-10] 샘플사업에 대한 키워드 클라우드 (uni-gram)

제3절 평가 및 활용성 검증

1. 검토방법 및 기준

- SBERT, KeyBERT 모델이 추천해준 관련사업 결과에 대한 유사도에 대한 전문가의 평가·검증을 수행하였으며, 평가자로 참여한 전문가는 KISTEP 전략계획서 사업별 간사를 담당했던 6인과 기술분야별 전문기관 사업평가 및 사업관리 담당자 9인을 대상으로 함

- 평가방법은 8개 분야 160개 샘플사업의 관련사업으로 추천된 474개에 대하여 유사성, 유용성을 검토하는 방식으로 양식과 기준을 제공하여 판단하도록 함

- 예비타당성조사 지침에 따르면 사업 유사도는 사업목적, 내용, 수행주체, 추진체계 등의 항목을 기준으로 유사도를 검토할 수 있으며(KISTEP, 2023), NTIS 유사중복과제 검토기준에 따르면 연구목표, 내용, 기대효과, 키워드 등을 기준으로 과제의 유사도를 검토할 수 있음(KISTEP, 2012)을 제시하고 있음

- 이를 참고하여 아래의 유사도 판단기준을 제시하고, AI 추천결과에 대하여 전문가들이 각 항목을 기준으로 했을 때 유사하다고 보는지(O), 아닌지(X) 여부를 체크하도록 하고, 종합적으로 추천 결과가 어느 정도 유사하다고 판단하는지(유사성), 유사도를 고려하지 않더라도 추천결과를 어느 정도 활용할 수 있다고 판단하는지(유용성) 등 두 개의 기준으로 5점 척도로 주관적인 평가를 병행하도록 함

[표 3-25] AI 관련사업 추천결과에 대한 유사도 판단기준

판단요소		내용	판단기준	전략계획서 항목
주요 정보	사업 목표	사업을 통해 달성하고자 하는 목적·목표	내용의 동일성	사업목적, 전략목표 (사업유형 참고)
	사업내용	해당 사업의 세부내용		내역사업별 연구활동내용
보조 정보	지원 대상	사업의 주요 지원 대상(수행주체) - 대학, 출연연, 기업 등		지원대상(수행주체)
	추진 체계	해당 사업의 추진 형태 - 상향식, 하향식, 혼합식		사업추진방식
종합 판단기준		※ (원칙) 개별 판단요소 중 목표, 내용에서 유사하게 분류될 경우, '유사'로 판단(지원대상 및 추진체계는 부수적 판단요소로 활용) ※ (예외) 성과지표 설정 등 측면에서 참고사례로 판단 가능한 경우에는 사업수행내용에 대한 검토 후 '유사'로 판단 가능		

자료 : 예타수행 세부지침(2023, p.189); KISTEP(2012, p.62); NTIS 연구개발과제 차별성 검토기준을 참고하여 연구진 작성

2. 전문가 평가 결과

[표 3-26] 전문가 유사도 평가 결과의 예시

<table>
<tr><th colspan="2">대상사업</th><th colspan="2">비교대상사업</th><th colspan="4">사업 속성</th></tr>
<tr><th>부처</th><th>사업명</th><th>부처</th><th>사업명</th><th>사업목표</th><th>사업내용</th><th>지원대상</th><th>추진체계</th></tr>
<tr><td rowspan="11">산업부</td><td rowspan="11">산업기술
국제협력</td><td>과기
정통부</td><td>해외우수기관협
력허브구축</td><td>o</td><td>x</td><td>x</td><td>o</td></tr>
<tr><td colspan="2">종합의견</td><td colspan="4">사업목표가 국제R&D협력 및 네트워킹이라는 측면에서 유사하나, 사업내용 상 지원분야가 화학/유전학/생명과학과 정보통신/ 에너지/기계 등으로 상이한 점, 지원대상이 대학/출연연과 기업으로 상이한 점 등을 근거로 관련사업으로 판단되지 않음. 국제협력이라는 목표를 제외하고 인력 및 기술협력 측면에서 ‘산학연CollaboR&D’가 더 유사한 사업에 가깝다고 판단됨.</td></tr>
<tr><td colspan="6"></td></tr>
<tr><td>부처</td><td>사업명</td><td>사업목표</td><td>사업내용</td><td>지원대상</td><td>추진체계</td></tr>
<tr><td>산업부</td><td>소재부품글로벌
투자연계기술개
발R&D</td><td>o</td><td>o</td><td>o</td><td>x</td></tr>
<tr><td colspan="2">종합의견</td><td colspan="4">사업목표가 국내 기업의 글로벌 역량 강화라는 측면에서 유사하며, 사업의 세부내용 상 정보통신분야 연구기관(기업)의 국제협력 또는 글로벌 시장진출을 지원하며, 지원대상도 중소, 중견기업으로 동일한 점에서 유사하다고 판단됨. 산업기술국제협력고 달리 소재부품글로벌투자연계기술개발 사업은 상향식 추진이나, 해당사항은 부수적 판단기준으로 동 사업을 관련사업으로 판단함</td></tr>
<tr><td colspan="6"></td></tr>
<tr><td>부처</td><td>사업명</td><td>사업목표</td><td>사업내용</td><td>지원대상</td><td>추진체계</td></tr>
<tr><td>산업부</td><td>월드클래스
플러스
프로젝트지원</td><td>o</td><td>o</td><td>o</td><td>o</td></tr>
<tr><td colspan="2">종합의견</td><td colspan="4">사업목표가 국내 기업의 신시장 진출 및 글로벌 역량 강화라는 측면에서 유사하며, 중소, 중견기업의 글로벌 시장진출을 지원한다는 점, 상향식 추진인 점에서 관련사업으로 판단</td></tr>
</table>

- SBERT 추천결과에 대한 전문가 평가 결과, 추천사업 474개 중 전문가가 유사성을 ‘보통’ 이상으로 평가한 사업은 43.0% (204개), 유용성을 ‘보통’ 이상으로 평가한 사업은 67.9% (322개)이었으며, 유사성과 유용성 평가에 차이가 있으며, 목적, 내용, 지원대상이 유사할 경우 전문가(사용자)의 평가점수가 높은 경향을 보임

- 따라서 전문가 평가기준을 참고하여 사업목적 및 내용 유사도에 가중치를 부여하는 방식으로 모델 보완을 검토해야 함을 시사점으로 도출함

- 키워드 검색실험 결과와 검색옵션(파라미터) 조정에 따른 추천결과에 대한 전문가 검증 결과에 대해서는 가장 유사한 사업을 잘 찾아내는 것은 0.7 옵션이, 상위 3개 사업을 잘 찾아내는 것은 0.5 옵션이 적합한 것으로 나타남

- 평가 검증을 통해서 실험분석과 마찬가지로 사용자가 직접 파라미터값을 조정할 수 있도록 옵션 제공이 필요함을 재확인하였음

[표 3-27] (좌) 판단요소별 평가결과와 점수의 관계 / (우) AI 모델과 전문가 평가 일치도

	유사성		유용성	
사업목적	1.26	***	0.72	***
사업내용	0.79	***	0.59	***
지원대상	0.22	***	0.37	***
추진체계	0.05		-0.04	

기준	파라미터	중소기업	대학	치매극복
top5	0.7	80%	80%	80%
	0.5	60%	60%	100%
top3	0.7	33%	33%	100%
	0.5	67%	33%	67%
top1	0.7	O	X	O
	0.5	X	X	O

주1 : 좌측 표의 수는 상관계수, ***은 0.01의 p-value 수준에서 유의함을 나타냄

주2: 우측 표의 top 5는 관련사업 유사도 기준 상위 5개 사업의 일치도, top3는 상위 3개, top1은 상위 1개사업의 일치도 또는 일치여부를 나타냄

Ⅳ. 중간평가 등급 설정 지원 서비스 기획 연구

o 중간평가 자체평가 등급설정 지원 서비스는 중간평가 자체평가보고서 작성 시 평가의견과 평가등급간 일치하지 않아 발생하는 행정적·기술적 오류로 인한 비효율을 최소화하고 향후 시스템 고도화 이후에는 평가등급 설정이라는 의사결정을 지원할 수 있는 서비스로 발전이 기대되는 탐색적 연구를 추진함

제1절 데이터

1. 데이터 수집과 데이터셋 구축

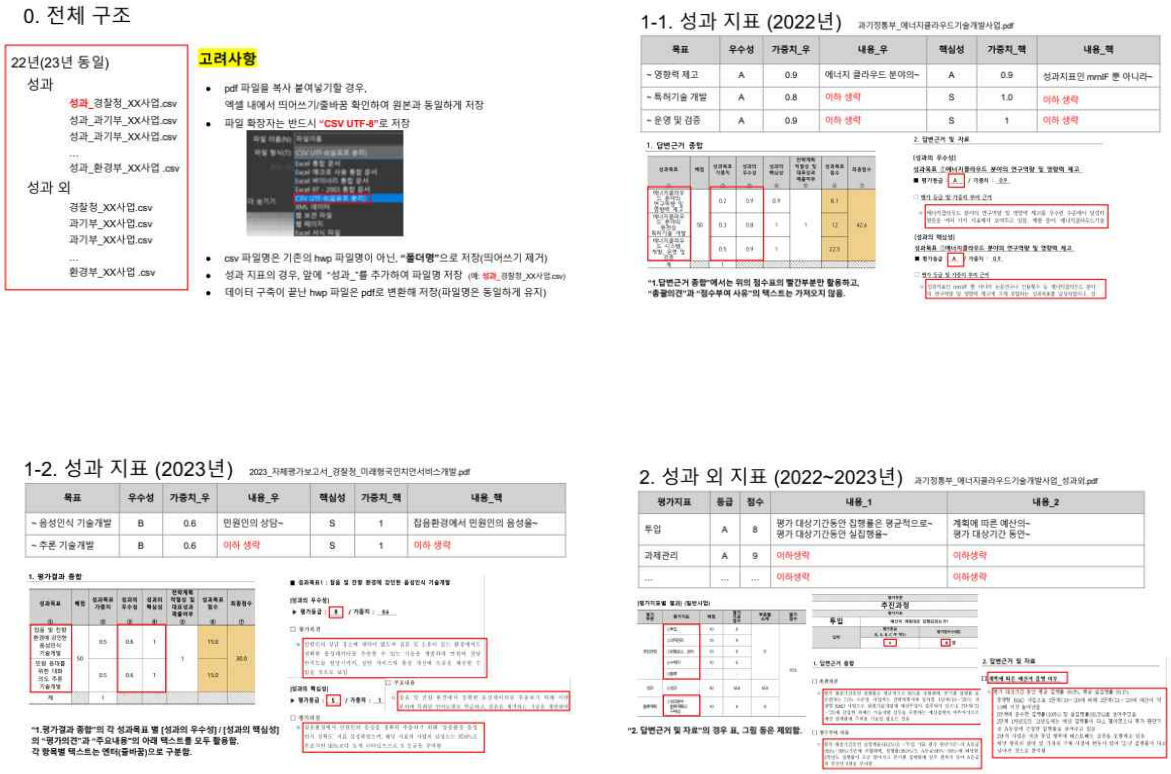
[표 4-1] 인공지능 활용을 위한 평가데이터셋 구축 결과 개요

수집 목적	수집 문서	대상연도	데이터 수	구축 항목
모델구축	전략계획서 (981개)	2021 - 2022	1,005개	사업코드, 부처명, 사업명(단위/세부/내역), 내역사업별 연구활동내용, 사업목적, 추진방식, 사업유형, 다부처여부, 전략목표, 성과목표/지표, 측정방법 등
	자체평가보고서 (277개)	2022 - 2023	646개	부처명, 사업명, 성과목표, (성과유형), 평가부문별 평가의견 요약, 세부 평가의견, 등급 등
모델 검증 · 보완	성과목표지표계획서 (731개)	2016 - 2020	752개	모델구축용 데이터와 동일
	자체평가보고서 (221개)	2019 - 2021	1,240개	

o 자체평가보고서 데이터의 경우 구축 결과 데이터의 수가 적어 분석에 한계가 있을 것으로 예상했으나, 평가의견 텍스트의 길이가 길고, 여러 의견이 종합적으로 작성되어 있는 특징을 고려하여 평가지표별 의견으로 세분화하여 구축하는 것으로 데이터 단위를 조정함

- 그럼에도 불구하고, 추가적인 텍스트 데이터 보완이 필요할 것으로 예상되어 2차 구축 작업을 통해 2019-2021 자체평가보고서 데이터를 보완용 데이터로 구축함[57)]

57) Random Deletion과 Random Swap 등 통상적 한국어 텍스트 증량을 시도하고 그 결과를 살펴보았으나, 평가의견에 반복적으로 작성되는 문구가 다양한 점, 그 결과 최다빈도 단어 기준 분석결과가 유의미하게 달라지지 않는 점, 등급별 불균형으로 인한 bias가 더욱 커질 수 있다는 우려 등을 고려하여 증량을 통한 학습데이터 증강보다는 가장 유사한 이전 제도의 자체평가의견 데이터를 추가 구축하는 것을 선택함



[그림 4-1] 자체평가보고서 데이터셋 구축 방법 (1차, 2022-2023 자체평가보고서 대상)

- 2019-2021 자체평가보고서 데이터는 평가제도 개선 이전의 지침에 따라 추진과정 지표의 데이터가 없는 점, 평가의견을 성과목표 단위가 아닌 성과 유형별 배점에 따라 작성하도록 되어 있는 등 그 구조가 크게 상이하다는 점에서 우려사항이 있음

- 그러나 현재와 같이 등급별 지표의 가중치를 배점한다는 점, 성과유형별로 작성된 평가의견을 통합하여 현재의 자체평가 의견 데이터셋과 유사한 구조로 데이터셋을 구축할 수 있다는 점에서 2019-2021 데이터를 추가 구축 대상으로 설정함

o 추출된 자체평가보고서 데이터를 살펴본 결과 양식은 정해져 있으나, 작성범위, 내용, 기준 등의 작성 행태가 다양하고, 자유도가 매우 높아 코딩을 통한 추출에 한계가 있었으며, 일부 부처의 경우 개정된 작성지침에 따르지 않아 양식명이 달라 구획화하여 추출하는 데에도 한계가 있었음

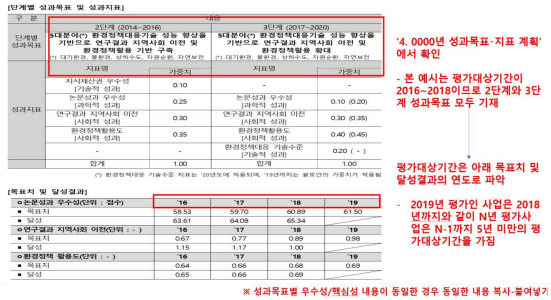
- 전략계획서는 정형화된 양식·항목으로 자동수집이 가능한 반면, 자체평가보고서는 별도 입력시스템을 마련하지 않는 한 활용을 위한 데이터를 전체 수작업을 통해 구축해야한다는 점, 매년 업데이트하는 작업과 불용어 처리 등의 기반 작업이 반복적으로 이루어져야한다는 점에서 데이터 구축으로 인한 비용 효율성이 매우 낮다고 할 수 있음

- 딥러닝 모델을 활용하는 경우에는 데이터 추출과 최소한의 stopword 제거 작업만을 추진할 수 있어 전처리의 시간과 비용을 줄일 수 있으나, 보고서 항목이나 불필요한 문장과 단어들이 평가의견에 많이 포함되어 있는 점, 평가의견과는 다소 상관이 없는 사업별, 기술분야별 전문용어가 반복적으로 작성되어 있어 이를 반복학습할 경우 편향이 발생할 수 있는 점 등을 고려해 모델 성능 제고를 위한 불용어 작업 또한 반복해서 이루어져야할 것으로 보임

양식 및 예시 _ 2019 환경부_환경정책기반공공기술개발사업 : 별첨 엑셀과 한글파일 참고

목표	성과유형	우수성	가중치_우	내용_우	핵심성	가중치_핵	내용_핵
5대분야(*) 환경정책 대응기술 성능 향상을 기반으로 연구결과 지역사회 이전 및 환경정책활용 기반 구축	과학적 성과	작성x	0.8	환경정책기반사업의 특성을 고려할 때 논문의 63%가 SCI급 논문에 해당하고, 세계 최상위 5% 수준의 저널에도 16건이 게재되는 등 우수한 질적 수준을 확보함으로써 해당 성과가 세계수준으로 0.8 부여 고해적으로 새로운 사실을 발견하는 등의 기초연구가 사업의 -	작성x	0.7	해당 성과를 논리모형(Logic Model)에 기반 하면, 논문 성과는 지식재산권, 기술이전 등으로 연계되어 사업의 최종 목적 달성을 위해 선행적으로 발생되는 필수적인 성과로 본 사업의 핵심성이 높음.
5대분야(*) 환경정책 대응기술 성능 향상을 기반으로 연구결과 지역사회 이전 및 환경정책활용 확대 1.	과학적 성과	작성x	0.8	환경정책기반사업의 특성을 고려할 때 논문의 63%가 SCI급 논문에 해당하고, 세계 최상위 5% 수준의 저널에도 16건이 게재되는 등 우수한 질적 수준을 확보함으로써 해당 성과가 세계수준으로 0.8 부여 고해적으로 새로운 사실을 발견하는 등의 기초연구가 사업의 -	작성x	0.7	해당 성과를 논리모형(Logic Model)에 기반 하면, 논문 성과는 지식재산권, 기술이전 등으로 연계되어 사업의 최종 목적 달성을 위해 선행적으로 발생되는 필수적인 성과로 본 사업의 핵심성이 높음.

(2019) 환경부_환경정책기반공공기술개발사업 P.7 - **목표**



'4. 0000년 성과목표·지표 계획'에서 확인

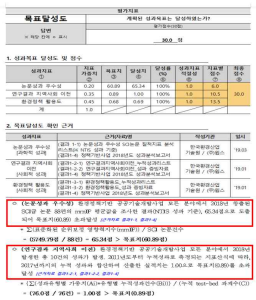
- 본 예시는 평가대상기간이 2016~2018이므로 2단계와 3단계 성과목표 모두 기재

평가대상기간은 아래 목표치 및 달성결과의 연도로 파악

- 2019년 평가인 사업은 2018년까지와 같이 N년 평가사업은 N-1까지 5년 미만의 평가대상기간을 가짐

※ 성과목표별 우수성/핵심성 내용이 동일한 경우 동일한 내용 복사-붙여넣기

(2019) 환경부_환경정책기반공공기술개발사업 P.10 - **핵심성**

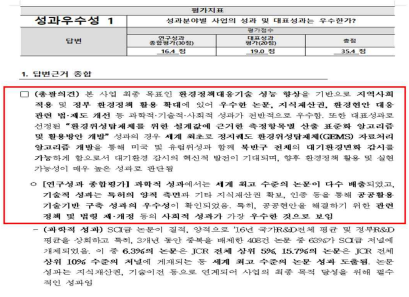


목표달성도 – 2. 목표달성도 확인 근거에서 텍스트 추출

1. 표 제외, 텍스트 중 산식 제외, 숫자로 계산된 실적치 제외

2. 근거자료 결과1-1, 결과1-4와 같이 근거자료 지칭어 제외

(2019) 환경부_환경정책기반공공기술개발사업 P.12~14 - 우수성



성과우수성1 – 1. 답변근거 종합에서 텍스트 추출

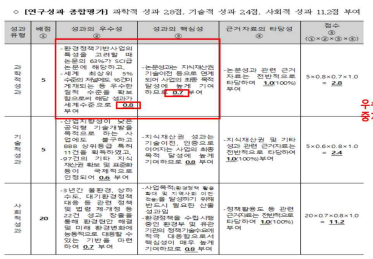
1. 네모, 등그라미, 괄호안 목차어(예: 총괄의견, 과학적 성과, 기술적 성과 등등) 삭제

2. 대표성과 관련 의견은 포함X

3. 총괄의견, 종합평가 등 성과유형 구분없이 공통된 의견은 모든 성과유형에 복-붙하고, 성과유형별 구분된 의견은 구분하여 입력

4. PDF인 경우 복사붙여넣기 시 문장이 끊어지는 경우가 있어 원본과 같이 문장 단위로 이어지도록 엑셀에서 수정 필요

(2019) 환경부_환경정책기반공공기술개발사업 P.14 – 우수성/핵심성

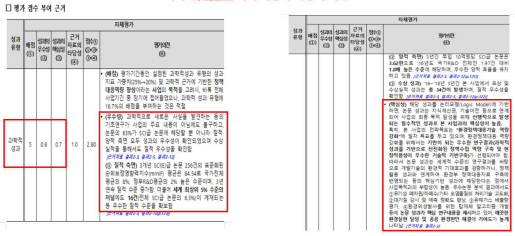


성과유형 참고

우수성/핵심성 각각 내용과 가중치 작성

(2019) 환경부_환경정책기반공공기술개발사업 P.17~22 – 우수성/핵심성

우수성/핵심성 각각 내용과 가중치 작성



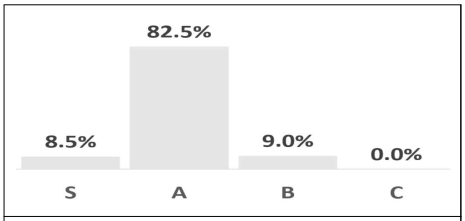
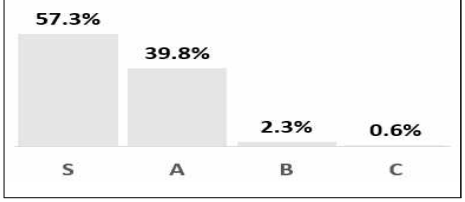
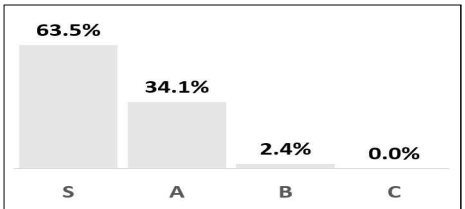
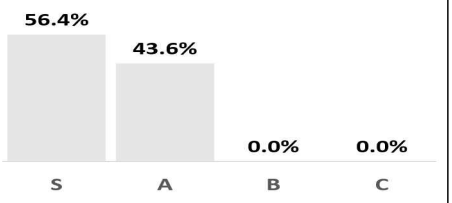
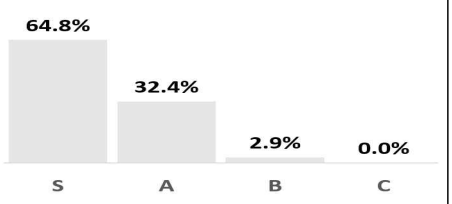
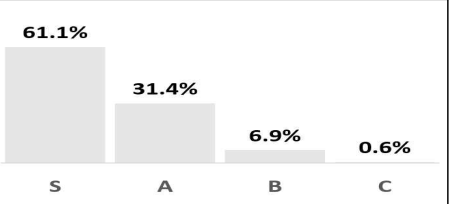
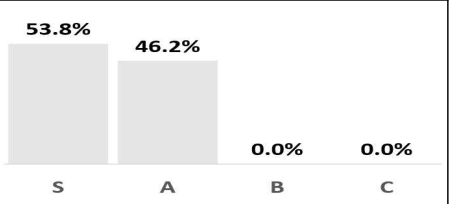
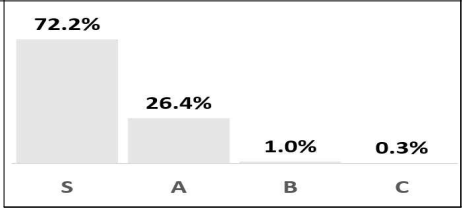
[그림 4-2] 자체평가보고서 데이터셋 구축 방법 (2차, 2019-2021 자체평가보고서 대상)

o 자체평가보고서 데이터의 양적 부족은 평가의견 단위의 문장 분석을 통해 해소 가능하지만, 가장 큰 문제점은 등급별 의견의 분포가 매우 불균형적이라는 점임

- 성과외 추진과정 지표에 대해서는 2019-2021 평가대상이 아니었으므로 그 수가 매우 적을 뿐만 아니라, B와 C등급으로 평가된 사례가 거의 없음

- 성과의 경우도 성과의 우수성에 C등급이 없어 신규 평가의견 데이터가 B 또는 C 등급에 해당하는 경우 등급 예측의 정확도가 현저히 떨어질 것을 우려함

[표 4-2] 자체평가 등급별 의견 분포

구분	평가지표	총 갯수	등급	데이터 수	등급별 분포
성과	성과의 우수성	646	S	55	
			A	533	
			B	58	
			C	-	
	성과의 핵심성	646	S	370	
			A	257	
			B	15	
			C	4	
성과외	투입	211	S	134	
			A	72	
			B	5	
			C	-	
	과제관리	282	S	159	
			A	123	
			B	-	
			C	-	
	위험요소관리	210	S	136	
			A	68	
			B	6	
			C	-	
	수혜자	175	S	107	
			A	55	
			B	12	
			C	1	
	환류	91	S	49	
			A	42	
			B	-	
			C	-	
	성과분석과 환류계획의 구체성	295	S	213	
			A	78	
			B	3	
			C	1	

2. 전처리와 형태소 분석기

ㅇ 데이터셋 구축 이후 데이터의 활용을 위해 형태소분석기의 성능을 검토하고, 불용어 제거 등 전처리를 진행함

ㅇ 머신러닝 모델의 경우, 분류 성능을 높이기 위해 각 지표별 Vocab을 만들어 등급(S, A, B, C)마다 어떤 Vocab이 유의미한지 확인하고, 평가 의견을 분류하고자 함

- 각 등급별 Vocab의 차이는 유의미하게 나타나지 않았기 때문에 단순 Count 행렬을 만들어 모델 학습에 활용함. 이 때, Vocab은 문맥 유추를 위해 세 단어 조합인 tri-gram으로 구성하였음

ㅇ 딥러닝 모델의 경우, 원문 텍스트를 학습하여도 분류에 강건한 성능을 보이는 것으로 알려져 있으므로 최소한의 전처리를 통해 최대한 평가 의견 그대로 모델 학습에 활용함

- 이에 따라 특수기호 제거와 띄어쓰기를 모두 한 칸으로 변경하는 작업만 진행함. 다만 특수기호 중, %의 경우에는 100% 달성 등 등급을 구분하는 내용으로 자주 언급되었기에 제거하지 않음

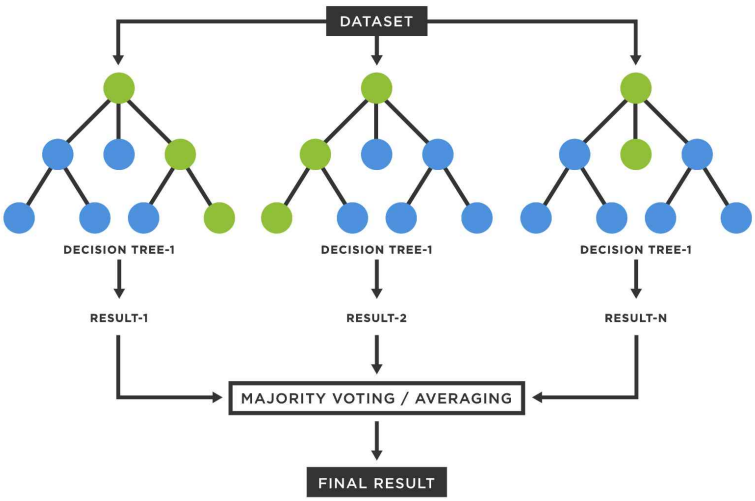
ㅇ 머신러닝 모델의 입력 값으로 텍스트를 활용하기 위해 형태소 단위로 텍스트를 분절하는 전처리 과정을 수행함. 형태소란 의미를 가진 가장 작은 단위로, 일반적으로 형태소 분리를 위해 아래의 형태소 분석기를 검토하였음

- ckonlpy라이브러리(김현중, 2018)를 통해 보다 쉽게 사용자사전 추가가 가능하고, 선택적으로 사전이 추가되는 양상도 보이지 않는 등 장점을 고려하여 성과 데이터 전처리에 twitter 형태소 분석기를 사용하였으며, 조사(Josa), 어미(Eomi), 구두점(Punctuation), 외국어, 한자 및 기타기호(Foreign)외 모든 품사를 선정하여 분석하였음

제2절 모델 탐색 및 보완

1. 랜덤포레스트(Random forest)

○ 랜덤포레스트(Random Forest) 는 레오 브레이먼의 논문에서 제시된 개념의 머신러닝에서 사용되는 지도학습 알고리즘[58]으로, 다수의 의사결정나무 (Decision Tree)를 조합한 모델이며, 분류모델을 구축하는데 사용되는 앙상블 학습방법론임 (L Breiman, 2001)



자료 : 김준기 외(2022), 빅데이터 기반의 도로안전성 분석에 관한 기초 연구, 기본 22-10, 국토연구원

[그림 4-3] 랜덤포레스트 개념도

- 의사결정트리들을 생성해 각 트리의 예측을 결합하여 최종 결정을 내리는 방식으로, 각 트리는 데이터세트의 무작위 샘플에서 독립적으로 학습되며, 이는 과적합의 방지에 도움이 됨

- 각 트리의 예측이 집계되어 최종 예측이 이루어지는 방식으로 작동이 되며, 분류의 경우 다수결, 회귀의 경우 평균 예측치가 출력됨

○ 랜덤포레스트모델의 장점은 분류와 회귀 모두에 사용할 수 있으며, 불균형 데이터의 분류에서 우수한 성능을 가지고 있어 자연어 데이터에 적용하기 용이하나, 랜덤포레스트 모델은 트리가 많을수록 학습과 예측에 더 많은 시간이 필요하여 실행 시간이 길어진다는 단점이 있음

58) 개념도 출처: https://www.spotfire.com/glossary/what-is-a-random-forest

[표 4-3] 랜덤포레스트의 장단점[59)]

구분		주요내용
장점	범용성	분류와 회귀 모두에 사용할 수 있으며, 다양한 유형의 데이터에 적용 가능함
	정확성	여러 결정 트리의 조합을 사용하기 때문에 정확도가 높음
	용이성	결측치를 다루기 쉽고, 대용량 데이터 처리에 효과적임. 하이퍼 파라미터 튜닝을 많이 하지 않아도 잘 작동하고, 데이터 전처리에서 스케일링 단계가 필요없음
	과적합 방지	개별 결정 트리에 비해 과적합의 위험이 적음
	피쳐 중요도 평가	각 특징(Feature)이 예측에 얼마나 중요한지 평가할 수 있어 feature selection을 중요도를 이용해 선택 가능함 (변수의 선정 및 랭킹 구하기 가능)
	병렬 처리	각 트리는 독립적이기 때문에 병렬 처리가 가능해 계산 효율이 높음
단점	실행 시간	트리가 많을수록 학습과 예측에 더 많은 시간이 필요함. 알고리즘에서 수백 또는 수천개의 트리를 만들기 때문이며, 메모리를 많이 사용함
	모델 크기	많은 수의 트리로 인해 모델이 크고 저장 공간을 많이 사용할 수 있음
	해석의 어려움	앙상블이라 단일 결정 트리보다 해석하기 어려움. 블랙박스와 같이 시각화해서 설명하지 못함
	노이즈에 민감	노이즈가 많은 데이터에서는 성능이 저하될 수 있음
	데이터별 성능차이	텍스트 데이터와 같이 차원이 높고 희소한 경우 잘 작동하지 않음. 그러한 경우 random forest보다는 선형 모델 또는 신경망이 적합함

- 사용자의 선호도나 행동을 바탕으로 제품이나 서비스를 추천하거나, 질병 조기진단, 발병가능성의 예측에 대표적으로 사용

[표 4-4] 랜덤포레스트의 활용분야

활용분야	주요내용
의료 분야	질병 조기진단, 발병가능성의 예측에 대표적으로 사용되며, 환자 데이터를 바탕으로 질병을 진단하거나 위험 요소 예측할 수 있음
금융 분야	대출 신청자의 신용 위험을 예측하고 신용 스코어를 계산하여 대출의 승인 또는 거부 결정을 내릴 때 도움을 줌. 보험 청구 데이터를 분석해 사기성 청구를 탐지하는 비정상적 패턴 식별 탐지에도 활용됨
생태학	생물 다양성 및 환경 변화의 영향을 예측
추천 시스템	사용자의 선호도나 행동을 바탕으로 제품이나 서비스를 추천
이미지 분류	복잡한 이미지 데이터에서 특정 객체를 식별하고 분류

- Albadi 등 (2019) 은 아랍어 트위터에 종교적 증오를 퍼뜨리는 데 있어 아랍어 봇을 정확하게 식별할 수 있는 새로운 회귀 모델로 랜덤 포레스트를 소개[60)]

59) https://heytech.tistory.com/149

60) Albadi, N., Kurdi, M., & Mishra, S. (2019). Hateful people or hateful bots? Detection and characterization of bots spreading religious hatred in Arabic social media. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-25.

- 영어 봇을 매우 정확하게 탐지하는 기존 도구가 아랍어 봇에서는 제대로 작동하지 않는 것으로 나타나 언어, 콘텐츠, 행동 및 네트워크 기능에 대한 분석을 수행하고, 아랍어와 영어 봇 간의 차이점에 대해 보고

- 로지스틱 회귀 및 그래디언트 부스팅 회귀 트리와 같은 다른 회귀 알고리즘과 비교하였을 때, 랜덤 포레스트 모델에 비해 해당 알고리즘들이 성능이 떨어지는 것으로 나타남

ㅇ 국민대학교 연구진은 1,295개 국내 상장기업을 대상으로 하는 기업신용등급 평가모형 구축에 랜덤 포레스트를 적용하여 경영분야 문제, 특히 기업신용위험 관리에 있어서의 RF 알고리즘의 적용 가능성을 확인[61)]

- 랜덤 포레스트 알고리즘의 성과를 비교평가하기 위하여 다중판별분석, 인공신경망, 다분류 SVM 모형을 사용하였으며, 실증분석 결과, 랜덤 포레스트 기법이 비교한 기법들에 비해 보다 정확한 예측결과를 산출함을 확인

- 단일 등급 분류(one class classification), 다양성 및 정확성에 기반한 분류모형 선택(diversity and accuracy-based classifier selection), 데이터 불균형(data imbalance) 문제 해소 등의 이슈를 종합적으로 다룰 필요가 있다고 제언

ㅇ 김판준(2019)[62)]은 랜덤포레스트 모델을 국내 학술지 논문의 자동분류에 적용한 사례에서 트리수 구간, 자질선정, 학습데이터의 크기에 대한 비교분석을 실시

- 학술지 논문은 하나의 논문에 다수의 분류가 할당되는 1:n 분류 범주이며, 불균등 데이터라는 특성이 있어 분류 알고리즘의 성능이 저하되는 문제가 있었고, 이를 해결하기 위해 랜덤포레스트모델을 선택함

2. SVM (support vector machine)

ㅇ 서포트 벡터 머신(Support Vector Machine, SVM)은 머신러닝의 지도학습 모델 중 하나로, 주로 분류와 회귀 분석에 사용됨[63)]

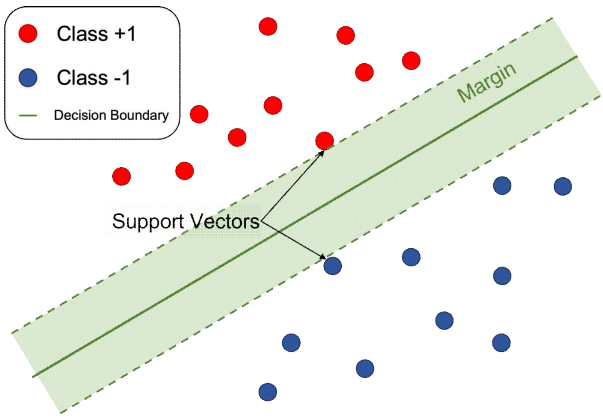
- 러시아의 통계학자인 Vapnik에 의해 처음 소개된 학습기법으로, 입력공간과 관련된 비선형문제를 선형문제로 대응시켜 나타내 수학적으로 분석하는 것이 수월하다는 장점을 가진 모델 (Hearst et al., 1998)

61) 김성진, & 안현철. (2016). 기업신용등급 예측을 위한 랜덤 포레스트의 응용. 산업혁신연구, 32(1), 187-211.
62) 김판준. “랜덤포레스트를 이용한 국내 학술지 논문의 자동분류에 관한 연구.” 정보관리학회지 36.2 (2019): 57-77.
63) 개념도 출처: https://bit.ly/42ofdQg

- 이 알고리즘은 데이터를 분류하기 위해 결정 경계(Decision Boundary) 또는 분리 초평면(Separating Hyperplane)을 찾는 것을 목표로 하며, 딥러닝이 나타나기 전 가장 유명하고 성능좋은 머신러닝 모델로 활약 (Son et al., 2004; Ahn et al., 2005)

- SVM은 데이터 포인트들을 서로 구분하는 최적의 초평면을 찾아내어 두 카테고리 간에 명확한 간격을 두는 방식으로 함. 이 간격을 마진(Margin)이라고 하며, SVM은 이 마진을 최대화하는 초평면을 찾음

- 비선형 데이터에 대해서는, 커널 트릭을 사용하여 데이터를 높은 차원으로 변환한 후 선형 분리가 가능하게 함



자료 : https://velog.io/@shlee0125/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-%EC%A0%95%EB%A6%AC-Support-Vector-Machine-05.-Why-does-SVM-maximize-margin (최종접속 : 24.3.8.)

[그림 4-4] SVM 개념도

ㅇ SVM의 가장 대표적인 장점은 효율성으로, 비교적 적은 수의 서포트 벡터에 의존하여 메모리가 절약된다는 강점이 있으며, 자연어 처리에도 SVM은 강점을 보이고 있는데, 국내에서는 기술정보 문서, 특허문서 분류 등에 SVM을 활용하여 문서의 내용을 바탕으로 미리 정의된 범주를 문서에 부여하는 문서 범주화 등에 대한 연구를 수행한 바 있음

- 강윤희 & 박용범 (2004)은 정보통신 웹 디렉터리 내의 문서로부터 추출된 단어 집합을 기반으로 SVM을 학습시킨 후 신규 문서에 대해 문서분류를 수행하였으며, 분류 실험을 통해 학습 벡터 구성과정에서 잡음에 의한 다른 클래스 문서분류에 미치는 영향을 고려한 SVM 기반 문서 분류 기법의 강건성을 확인[64]

- 경기대학교 연구진은 SVM을 이용한 특허문서 분류기의 설계 및 구현 관련 연구를 진행하였으며, 실험 결과 기술 간의 관련성이 적을수록 문서분류 정확도가 향상되는 것으로 나타났고, 이와 반대로 관련성이 가까워질수록 정확도는 이에 반비례하여 하락하는 것으로 나타남[65]

64) 강윤희, & 박용범. (2004). SVM 을 이용한 디렉토리 기반 기술정보 문서 자동 분류시스템 설계. 전기전자학회논문지, 8(2), 186-194.
65) 박찬정, 성동수 & 이건배. (2010) SVM을 이용한 특허문서 분류기의 설계 및 구현. 산업기술종합연구소논문집., 38, 115-128.

o 평가문서의 분류와 일맥상통할 수도 있는 단문메세지의 감성분류에 대해 SVM을 적용한 사례가 다수 존재

- 김현우 & 이승룡(2013)은 140자 이내의 단문 메시지에 대한 긍정 및 부정 감성분류를 SVM을 활용해 음운과 음절단위로 커널을 비교해 정답률을 비교한 연구를 실시했으며, F1 score 관점에서 음절보다 음운관점의 모델이 비교적 우수하다고 보고함[66]

o 반면, 모형 구축 시간이 비교적 오래 걸리고, 결과에 대해 직접적으로 확률적 해석을 할 수 없어 설명력이 떨어진다는 단점이 있음

[표 4-5] SVM의 장단점[67][68]

구분		주요내용
장점	효율성	고차원 데이터에서도 잘 작동하고 비교적 적은 수의 서포트 벡터에 의존하기 때문에 메모리가 절약됨
	유연성	다양한 커널 함수(선형, 다항식, RBF, 시그모이드 등)를 사용하여 다양한 유형의 데이터에 적용할 수 있음
	과적합 방지	마진을 최대화하는 원리로 인해 과적합의 위험이 상대적으로 낮음
	결정 경계의 명확성	최적의 분리 초평면을 찾음으로써 명확한 결정 경계를 제공함
단점	직관적인 해석의 어려움	직접적으로 확률적 해석을 할 수 없고, 결정 경계가 복잡하거나 고차원일 경우, 모델의 해석이 어려울 수 있음.
	대용량 데이터 처리	모형 구축 시간이 비교적 오래 걸림. 매우 큰 데이터 세트에 대해 시간소요가 크고 효율이 떨어질 수 있음
	스케일링 필요	변수의 스케일링에 민감하여 사전에 데이터 정규화가 필요함.
	모델 매개변수 선택	적절한 커널 선택과 매개변수 설정이 중요하지만 복잡할 수 있음

o 신호 처리, 의료 응용 분야, 자연어 처리, 음성 및 영상 인식을 비롯한 여러 분류 및 회귀 문제에 사용

[표 4-6] 랜덤포레스트의 활용분야

활용분야	주요내용
이미지 인식	이미지 데이터에서 특정 객체나 패턴을 식별하고 분류
생체 인식	지문, 안면, 홍채 인식등에 강력하게 사용됨
텍스트 분류	스팸 메일 분류, 문서 분류 등에 활용
생물정보학	단백질 분류, 유전자 데이터 분석 등에 사용
금융 분석	신용 위험 평가, 주식 시장 분석 등에 적용

66) 김현우, and 이승룡. "모바일 텍스트의 감성분류를 위한 SVM 기반 음운 커널 기법." 정보과학회논문지: 소프트웨어 및 응용 40.6 (2013): 350-355.

67) https://velog.io/@khsfun0312/SVMSupport-Vector-Machine

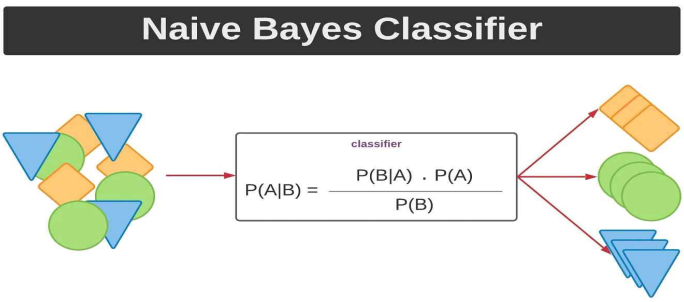
68) 박찬정, et al. "특허정보 문서를 이용한 자질선택 방법 및 분류 알고리즘의 성능비교 1." 대한전자공학회 학술대회 (2011): 1034-1036.

ㅇ George & Vinod (2018)는 SVM 기법을 활용하여 정상적인 메일과 스팸 메일을 분류하는 방안을 제시했으며, 특히 Enron 데이터세트에서 복합 기능을 추출하기 위해 기계 학습과 NLP도 적용[69]

- 머신러닝 알고리즘을 적용하기 전 추출된 특징(문자 기반, 단어 기반, 태그 기반, 구조 기반 특징 포함)의 순위를 매기기 위해 차원 알고리즘을 사용했으며, 테스트된 5가지 알고리즘 중 SVM은 가장 성능이 좋은 알고리즘으로 나타남

3. 나이브베이즈(Naive Bayes)

ㅇ 나이브 베이즈(Naive Bayes)는 베이지안 이론을 기반으로 하는 머신러닝 알고리즘[70]으로 주로 분류 작업에 사용되며, 특히 텍스트 데이터 분석에 많이 활용됨 (I Rish, 2001).



자료: https://zdnet.co.kr/view/?no=20220725093548 (최종접속: 24.3.8.)

[그림 4-5] 나이브베이즈 개념도

- 나이브(Naive) 란, 예측한 특징이 상호 독립적이라는 가정하에 확률 계산을 단순화하는 것이며, 베이즈(Bayes) 정리를 기반으로 전체 확률 대비 특정 클래스에 속할 확률을 계산하는 것으로, 사건 B가 주어졌을 때 사건 A가 일어날 확률인 조건부 확률과 베이즈 정리를 이용한 분류기라고 여길 수 있음

- 나이브 베이즈 classifier는 주어진 특성들이 서로 독립임을 가정하는 베이지안 분류 방법으로, 한 특성의 존재가 다른 특성의 존재에 영향을 미치지 않는다고 가정하며, 베이즈 정리를 토대로 각 클래스의 조건부 확률을 계산하고, 가장 높은 확률을 가진 클래스를 예측 결과로 선택함

ㅇ 매우 단순하고 적은 학습데이터로도 잘 수행되며, 결측 데이터가 있어도 우수한 성능을 보이나 모든 데이터의 특징을 독립적인 사건이라고 가정하기 때문에 문서 분류에는 적합하나 다른 분류 모델에는 제약이 있을 수 있음

69) George, P., & Vinod, P. (2018). Composite email features for spam identification. In Cyber Security: Proceedings of CSI 2015 (pp. 281-289). Springer Singapore.

70) 개념도 출처: https://zdnet.co.kr/view/?no=20220725093548

o 나이브 베이즈 분류의 가장 큰 장점은 적은 적은 양의 데이터에도 준수한 성능을 보여준다는 것으로, 500개 미만의 데이터만으로 이루어진 적은 데이터에서 KoBERT 모델과 견줄 수 있는 결과를 보임

- 조희련 등 (2021)의 연구에서 A, B, C, D 등급으로 채점된 외국인의 주제별 한국어쓰기 글짓기 답안지 300건을 대상으로 나이브베이즈, 로지스틱 회귀, KoBERT 모델을 사용해 성능평가를 한 결과 accuracy 기준 KoBERT 54.5%, 나이브베이즈 50.7%, 로지스틱회귀 44.7%로 로지스틱 회귀보다는 우수하며, KoBERT 모델과 근소한 차이로 떨어지는 성능을 보였음 (통합 테스트 데이터 성능 기준)[71]

[표 4-7] 나이브 베이즈의 장단점[72][73]

구분		주요내용
장점	간단하고 빠름	구현이 간단하고, 학습 및 예측 속도가 빠름
	처리력	노이즈와 결측치 처리가 용이해 상대적으로 적은 양의 학습데이터로도 좋은 성능을 낼 수 있음
	고차원 데이터에 적합	텍스트 데이터 같은 고차원 데이터에서 잘 작동함
	예측 정확도	예측을 위한 추정 확률을 쉽게 얻을 수 있음.
단점	특성 독립성 가정	모든 특징이 동등하게 중요하고 독립이라는 가정이 잘못될 경우가 자주 있는데, 일기예보에서 습도와 같은 중요한 변수를 사소한 변수들과 동등하다고 판단해 버리는 사례가 있음.
	과적합의 가능성	데이터에 특정 카테고리가 지나치게 많은 경우 과적합이 발생할 수 있음
	연속적인 데이터 처리의 어려움	수치 특징이 많은 데이터셋에는 이상적이지 않음

o 나이브 베이즈의 활용에 가장 많이 알려진 것은 스팸 메일 필터링으로, 텍스트 분류, 감정 분석, 추천 시스템 등에 광범위하게 활용

[표 4-8] 나이브 베이즈의 활용분야

활용분야	주요내용
스팸 메일 필터링	이메일의 텍스트를 분석하여 스팸 여부를 판별
문서 분류	뉴스 기사, 학술 논문 등 다양한 문서의 카테고리 분류에 사용
감정 분석	소셜 미디어, 리뷰 등에서 긍정적 또는 부정적 감정 분석에 활용
의료 진단	환자 데이터를 분석하여 특정 질병의 발병 가능성을 예측
추천 시스템	사용자의 선호도나 이전 행동을 바탕으로 제품이나 서비스를 추천

71) 조희련, et al. "KoBERT, 나이브 베이즈, 로지스틱 회귀의한국어 쓰기 답안지 점수 구간 예측 성능 비교." 한국정보처리학회 학술대회논문집 28.1 (2021): 501-504.

72) 조한철, and 조근식. "나이브 베이지안 분류자와 메세지 규칙을 이용한 스팸메일 필터링 시스템." 한국정보과학회 학술발표논문집 29.1B (2002): 223-225.

73) https://bit.ly/3SoJolV

ㅇ Shirodkar 등 (2020)은 트위터 데이터에 대하여 나이브베이즈 알고리즘을 활용한 감정 분석(sentiment analysis)을 통해 선거 결과를 예측하는 기계 학습 모델을 제안[74)]

- N-gram과 POS-tag를 특징으로 사용하는 Naive Bayes 분류기를 기반으로 정당을 언급하는 메시지를 양성, 음성 및 중립으로 분류하고 감정 분석을 통해 선거 결과를 예측

- 트위터 데이터로부터 훈련 데이터를 생성하고 선거 결과를 예측하기 위한 확장 가능한 기계 학습 모델을 제안하기 위해 2단계 프레임워크를 구성하였으며, 기존 연구에서는 단어에 대한 감성 분석을 실시해 왔지만 전체 문장에 대한 감성 분석을 적용하면 더 나은 결과를 얻을 수 있을 것으로 기대

ㅇ 이상기 외(2010)는 대규모 정보센터나 도서관에서 학술논문을 효율적이고 지능적으로 추천하기 위해 협업필터링과 나이브베이즈 모델을 결합한 하이브리드 방식의 추천시스템을 제시

- 본인이 저술한 논문이나 열람한 논문을 학술논문 DB와 로그파일을 통해 주기적으로 수집하여 초기 학습문서를 구축하고, 같은 논문을 열람한 관련분야 동료연구자나 공저자 관계에 있는 연구자들이 저술한 논문 또는 열람한 논문을 수집하여 학습문서에 추가한 후 같은 문서를열람한 횟수가 많은 동료 연구자나 공저자 빈도에 따라 가중치를 달리 부여하고 이를 토대로 추천문서 우선순위를 산정

- 기존 콘텐츠 기반 추천시스템의 과도한 특성화(Over-specialization) 문제와 이용로그나 평가정보가 축적되기 전까지 신규논문을 추천할 수 없었던 협업필터링 방식의 문제점을 동시에 해소하였으나, 다양한 분야에서 활동하는 연구자들을 고려한 추천모델 성능 개선의 추가 연구 필요성 제시

4. XG Boost

ㅇ XGBoost (eXtreme Gradient Boosting)는 부스팅 방법 중 하나로, 약한 학습기(Weak Learner)들을 순차적으로 학습시켜, 그 예측들을 결합해 강력한 최종모델을 만드는 과정[75)]

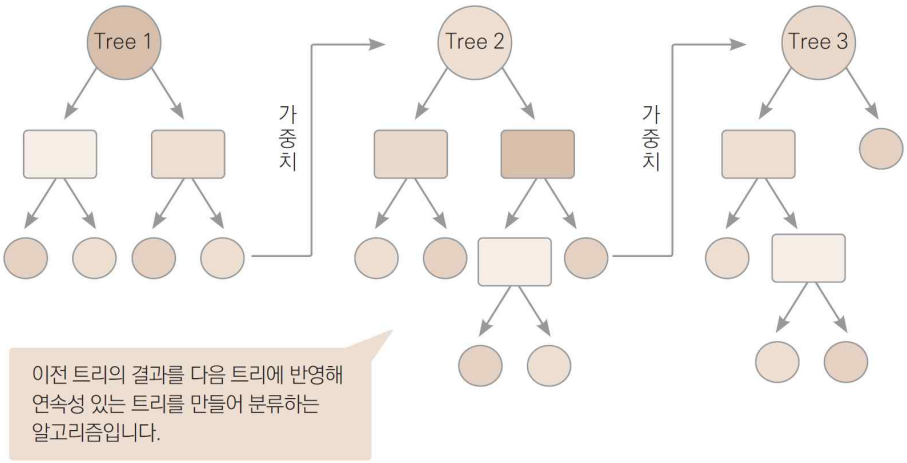
- 머신러닝에서 널리 사용되는 앙상블 학습기반의 알고리즘으로, tianqi chen에 의해 제안됨(T Chen et al., 2016)

- XGBoost는 여러 개의 결정 트리를 결합하여 작동하며, 각 트리는 이전 트리의 오류를 바탕으로 학습되며, 이러한 접근 방식은 모델의 정확도를 점진적으로 향상시킴

74) Shirodkar, M., Nimbalkar, S., Ingole, A., Mishra, P., & Sahu, S. (2020). Election result prediction using Sentiment Analysis. International Research Journal of Engineering and Technology, 7(2), 2855-2857.

75) 개념도 출처: 권시현. “데싸노트의 실전에서 통하는 머신러닝”, 골든래빗 (2022)

- 각 반복에서, XGBoost는 모델의 성능을 최적화하는 방향으로 트리를 추가하며, 이 과정에서 그래디언트 부스팅 기법을 사용하고 과적합을 방지하기 위한 Regularization이 포함됨



자료 : https://anodos.tistory.com/entry/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98-XG%EB%B6%80%EC%8A%A4%ED%8A%B8XGBoost

[그림 4-6] XG Boost 개념도

o 여러 트리를 조합하는 앙상블 기법을 사용하여 높은 정확도와 성능을 제공하고, 다양한 유형의 데이터에 대해 강력한 일반화 능력을 가지고 있으나, 트리 기반 모델의 특성상 과적합이 발생할 수 있어 이를 방지하기 위해 모델의 복잡도를 조절하는 하이퍼파라미터 튜닝이 필요

[표 4-9] XG boost 의 장단점[76)]

구분		주요내용
장점	높은 성능	다른 알고리즘보다 낮은 오차율을 보이며, 데이터의 패턴을 잘 학습할 수 있음
	병렬 처리	효율적인 병렬 처리로 빠른 학습이 가능함
	규제화	과적합의 규제를 위한 기능이 있음 (L1, L2)
	유연성	사용자 정의 최적화 옵션을 제공하며, 분류 및 회귀 문제에 모두 사용할 수 있음
	결측치 처리	자체적으로 결측치를 처리할 수 있음
	평가지표 다양성	예측모델의 성능을 평가하기 위한 다양한 평가지표를 제공 ※ 분류의 경우 정확도/정밀도/재현율/F1 score 등을 제공하고, 회귀의 경우 MSE, R-squared 등을 사용가능). 이를 활용해 예측능력을 정량적으로 평가하고 개선할 수 있음
단점	설정 복잡성	많은 하이퍼파라미터가 있어 튜닝이 복잡할 수 있음. 많은 매개변수들을 조정해야하는데, 이에 많은 시간과 노력이 들어감
	해석의 어려움	결정 트리의 앙상블이기 때문에, 모델의 해석이 복잡하고 어려움
	자원소모	많은 메모리와 프로세싱 파워를 요구하는 특성이 있어 주의를 요하며, 병렬처리를 위해 CPU 코어가 많이 필요할 수 있음
	소규모 데이터셋에서의 과적합	비교적 작은 데이터셋에서는 과적합의 위험이 있을 수 있음

76) https://velog.io/@jjw9599/MLboostingconcept1

- XGBoost는 대규모 데이터셋에서 높은 성능을 보이며, 효과적으로 확장이 가능하여 특히, 피처의 수가 많거나 차원이 높은 NLP 작업에 적합하고, 특성 중요도를 계산하여 모델이 학습 중에 중요하지 않은 피처를 제거할 수 있어 노이즈가 많은 데이터셋에서 모델의 성능을 향상시키는 데 도움이 됨

- 작은 규모의 데이터셋에서는 과적합의 위험이 있으며 대규모의 데이터셋이나 복잡한 모델에서는 많은 계산 리소스가 필요할 수 있어 대용량의 텍스트 데이터를 다루는 데는 부적합할 수 있음

ㅇ 이진 분류 및 다중 클래스 분류 문제, 온라인 광고, 웹사이트 방문자 동작 예측 등 다양한 분야에서 사용

[표 4-10] XG Boost의 활용분야

활용분야	주요내용
금융 분야	신용 점수 평가, 사기 탐지
생물정보학	유전자 데이터 분석 및 질병 예측에
이커머스	고객 구매 예측, 제품 추천 시스템
기상학	기후 변화 예측, 날씨 패턴 분석
자연어 처리	텍스트 분류, 감정 분석

ㅇ XGBoost를 활용하여 서울특별시 응답소 민원 데이터에 대한 카테고리 확인 및 담당부서 분류를 자동화하기 위한 연구가 수행[77)]

- 7년치 민원 사례 17,700건의 데이터를 수집하여, XGBoost 모델과 Random Forest 모델과 비교하여 한국어 텍스트 분류의 적합성을 확인하였으며, 그 결과 RandomForest에 대비 XGBoost의 정확도가 전반적으로 높게 나타남

ㅇ 이경란 & 강창묵(2023)은 XG Boost를 활용하여 기계학습에 기반한 자연어처리 기법을 활용하여 공시자료의 분류를 자동화하는 방법을 제안[78)]

- 비밀처리(confidential treatment, CT)를 가지는 미국 수시공시 회계문서 8-K 양식의 자동판별을 위한 자연어처리(natural language processing, NLP) 기계학습 모델을 제안

- XGBoost 모형과 인공신경망 기반의 EmbedMixed, BERT 모형을 비교하였을 때, 재현율과 정밀도가 80%~90% 사이에서 서로 상쇄하는 수준을 보인 XGBoost 모형이 가장 우수한 성능을 보임

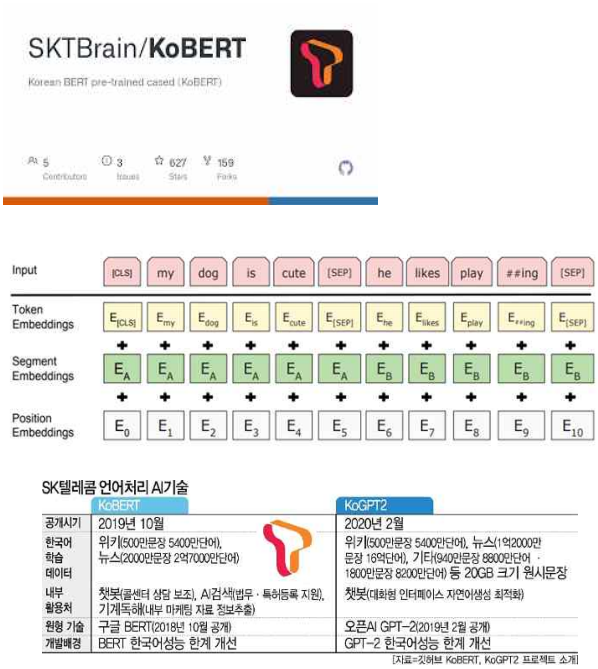
77) 하지은, 신현철, & 이준기. (2017). RandomForest 와 XGBoost 를 활용한 한국어 텍스트 분류: 서울특별시 응답소 민원 데이터를 중심으로. 한국빅데이터학회지, 2(2), 95-104.

78) 이경란, & 강창묵. (2023). 자연어처리 기계학습 기법을 이용한 공시문서의 자동분류: Confidential treatment 를 가진 8-K 문서를 중심으로. The Journal of Society for e-Business Studies, 28(2), 21-36.

5. KoBERT

ㅇ KoBERT는 SKTBrain이 자연어처리를 위해 설계하여 공개한 BERT(Bidirectional Encoder Representations from Transformers) 기반의 언어 모델[79]

- KoBERT는 트랜스포머(Transformer) 아키텍처를 기반으로 하며, 한국어의 특성을 고려한 사전 훈련된 언어 모델이며, BERT의 양방향 특성을 활용하여 문장 내 단어의 맥락을 보다 정확하게 파악하고 한국어의 복잡한 어휘와 문법 구조를 효과적으로 학습함



자료 : 아주경제, SK텔레콤은 '언어신동 AI'에 어떻게 한국어를 가르쳤나' (2020.10.12.일자 기사), https://www.ajunews.com/view/20201011091342159, (최종접속: 24.3.8.)

[그림 4-7] KoBERT 개념도

ㅇ KoBERT는 한국어에 특화된 사전학습을 수행하여, 한국어 문장의 표현을 효과적으로 학습할 수 있으나, 다른 언어의 데이터에 대해 일반화하기 어려울 수 있으며 사전학습 및 미세조정에 필요한 학습 시간과 계산 비용이 많이 소요될 수 있음

- 이전에 학습된 언어 모델을 사용하여 적은 데이터로도 효과적으로 모델을 학습시킬 수 있고, 사전 훈련된 언어 모델로 얻은 임베딩을 활용하여 다양한 자연어 처리 작업에 활용할 수 있음

- 모델이 학습한 데이터에 따라 한국어의 다양성을 반영하지 못할 수 있으며, 많은 데이터를 기반으로 사전 훈련되었기 때문에 풍부한 데이터가 없는 상황에서는 제한된 성능을 보일 수 있음

79) 개념도 출처1. https://www.ajunews.com/view/20201011091342159
개념도 출처2. https://github.com/SKTBrain/KoBert

[표 4-11] KoBERT의 장단점[80)81)]

구분		주요내용
장점	양방향 맥락 이해	문장 내 앞뒤 단어의 맥락을 모두 고려하여 언어를 이해함
	한국어 최적화	한국어의 어휘적, 문법적 특성을 반영한 사전 훈련 모델을 사용
	높은 성능	다양한 한국어 자연어 처리 작업에서 우수한 성능을 보여줌
	다양한 활용 가능성	분류, 질의응답, 감정 분석 등 여러 분야에 적용 가능함
단점	유연성 부족	위키피디아, 뉴스와 같은 정제된 문장형 텍스트로 사전훈련을 하여 신조어, 축약어, 구어체등에 대처가 어려울 수 있음.
	리소스 요구량	높은 계산 자원과 메모리 필요
	튜닝의 복잡성	사전 훈련된 모델로, 특정 작업에 맞게 미세 조정하는 것이 복잡함
	한정된 언어 범위	주로 한국어에 특화되어 있어, 다국어 처리에는 제한적임

o KoBERT는 한국어 텍스트의 감정을 효과적으로 이해하고 분석할 수 있어, 제품 리뷰, 소셜 미디어 콘텐츠 등에서 감정 분석에 활용될 수 있으며, 문서, 뉴스 기사, 리뷰 등의 텍스트를 다양한 카테고리로 분류하는 작업에서 뛰어난 성능을 보임

[표 4-12] KoBERT의 활용분야

활용분야	주요내용
감정 분석	소셜 미디어, 리뷰, 고객 피드백 등에서 감정분석
문서 분류	뉴스 기사, 학술 논문 등을 카테고리별로 분류
기계 번역	한국어와 다른 언어 간의 번역 작업에 활용
챗봇	자연스러운 대화형 인터페이스를 위한 챗봇 개발

o 이종하 등(2022)은 이공계 구직자에게 적합한 직무를 추천하는 모델 연구로 KoBERT를 이용하여 설명가능한 AI를 구현함[82)]

- 실제 약 1,000여명의 연구 이력서를 자연어처리하고 자동차 산업의 3가지 연구개발 직무인 전자제어, PE시스템, 연료전지/배터리로 분류하도록 파인 튜닝(Fine tuning)하여 약 평균 80% 정확도를 갖는 분류 모델과 더불어 희망하는 분야에 얼마나 적합한지, 이력서 중 가장 높은 적합성을 가지는 문장은 어떤 것인지 출력하는 설명가능한 AI를 구현

80) Lee, Sangah, et al. “Kr-bert: A small-scale korean-specific language model.” arXiv preprint arXiv:2008.03979 (2020).

81) 황상흠, and 김도현. “한국어 기술문서 분석을 위한 BERT 기반의 분류모델.” 한국전자거래학회지 25.1 (2020): 203-214.

82) 이종하, 구명완, 이경표. (개최날짜). KoBERT 기반의 연구개발 직무 추천 모델 연구. 대한산업공학회 춘계공동학술대회 논문집, 개최지.

ㅇ 노이즈나 이상값이 없는 예제를 가정하여 설계된 기계 학습 모델의 한계점을 해소하기 위한 KoBERT 모델 기반 연구가 진행 [83]

- 다중 레이블 데이터의 불균형 특성을 해결하기 위해 데이터 오버샘플링 기술을 사용하고 레이블 예측을 위한 전역 임계값을 설정하면서 파인 튜닝(Fine tuning)한 결과, 불균형하고 시끄러운 환경 뉴스 데이터에 대해 분류 성능을 80% 이상 향상시키는 EnvBERT 모델을 제시

ㅇ 문서 분류의 정확도를 높이기 위해 문맥 정보와 키워드 정보를 모두 사용하는 이중 접근(Dual Approach) 방법론이 제안됨[84]

- 한국어 말뭉치를 사전학습한 KoBERT를 사용하여 문맥 정보를 CLS 토큰 형태로 추출하고, 키워드 정보는 문서별 키워드 집합을 Autoencoder의 잠재 벡터를 통해 하나의 벡터값으로 생성하여 사용

- 제안 방법을 국가과학기술정보서비스(NTIS)의 국가 R&D 과제 문서 중 보건 의료에 해당하는 40,130건의 문서에 적용하여 실험을 수행한 결과, 제안 방법이 문서 정보 또는 단어 정보만을 활용하여 문서 분류를 진행하는 기존 방법들에 비해 정확도 측면에서 우수한 성능을 나타냄을 확인

6. RoBERTa

ㅇ RoBERTa (Robustly optimized BERT approach)는 워싱턴대학과 Facebook AI가 개발한 BERT(Bidirectional Encoder Representations from Transformers)의 개선된 버전으로, BERT의 핵심 아키텍처를 유지하면서 학습 방법과 데이터 처리 방식을 최적화하여 성능을 향상시킨 것으로 알려져 있음[85]

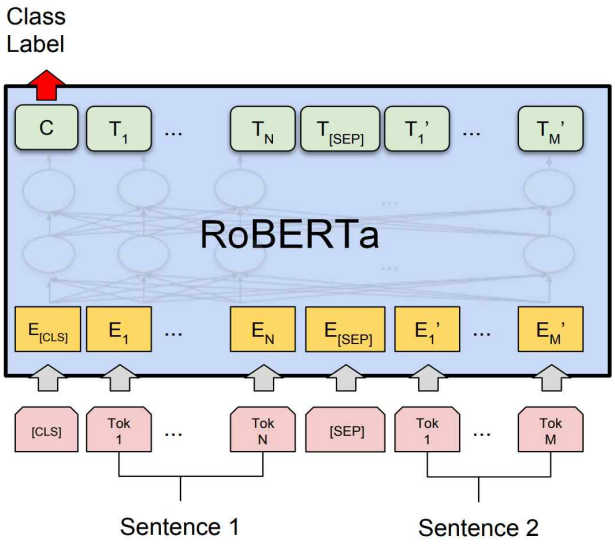
- RoBERTa는 BERT*의 변형으로, 더 큰 데이터 세트와 더 긴 학습 시간, 더 큰 배치 크기 등을 통해 BERT의 성능을 향상시킴 (마스킹 된 부분을 바꿔주며 학습, NSP제거, 롱 시퀀스 학습, 빅 데이터, 큰 배치 학습)

- RoBERTa는 여전히 트랜스포머 기반의 구조를 사용하며, 문장의 양방향 맥락을 파악하는 데 초점을 맞추었으며, BERT에서 사용된 몇 가지 학습 기법을 변경하거나 제거하여 성능을 개선시킴

83) Kim, D., Koo, J., & Kim, U. M. (2021, January). EnvBERT: multi-label text classification for imbalanced, noisy environmental news data. In 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1-8). IEEE.

84) Yoon, S., & Kim, N. (2023). Document Classification Methodology Using Autoencoder-based Keywords Embedding. Journal of the Korea Society of Computer and Information, 28(9), 35-46. https://doi.org/10.9708/JKSCI.2023.28.09.035

85) 개념도 출처: https://sooftware.io/roberta/



자료 : https://sooftware.io/roberta/, (치종접속: 24.3.8.)

[그림 4-8] RoBERTa 개념도

○ BERT보다 더 많은 데이터를 사용하여 사전학습되어 더 다양하고 일반화된 언어 이해 능력을 갖을 수 있으나 모델을 사용하는데 필요한 계산 자원이나 메모리가 많이 필요

- RoBERTa는 BERT와는 달리 문장의 순서 정보를 무시하고 문장을 쪼개어 학습하기 때문에 더 많은 학습 데이터를 활용할 수 있으며, 미세 조정을 통해 다양한 NLP 작업에 적용할 수 있는 유연성을 제공

- RoBERTa는 복잡한 모델 구조를 가지고 있어 추론 시간이 길어질 수 있으며, RoBERTa는 초기에 주로 영어에 중점을 두고 개발되어, 다국어 지원이 BERT보다는 제한적일 수 있음

[표 4-13] RoBERTa의 장단점[86)]

구분		주요내용
장점	높은 성능	최적화된 학습 전략으로 인해 BERT보다 성능이 향상됨
	양방향 맥락 이해	문장 내 각 단어의 양방향 맥락을 잘 파악하여 언어이해가 정확해짐
	범용성	다양한 자연어 처리 작업에 효과적으로 적용 가능
	대규모 데이터 학습 가능	큰 데이터 세트와 긴 학습 시간을 들였기 때문에 복잡한 언어 패턴을 학습할 수 있음
단점	대형모델	큰 데이터와 학습시간이 길어 컴퓨팅 리소스가 많이 필요하고 배포가 힘듦
	복잡성	최적의 성능을 달성하기 위해 세밀한 튜닝이 필요할 수 있지만 어려움
	사전학습 데이터 양의 의존	대량의 다양한 데이터에 의존하기 때문에 적은 양의 데이터만 존재할 경우 성능 향상이 제한될 수 있음

86) https://velog.io/@mmodestaa/RoBERTa-A-Robustly-Optimized-BERT-Pretraining-Approach

ㅇ RoBERTa는 문장이나 문서를 다양한 카테고리로 분류하는 텍스트 분류 작업이나 주어진 문맥과 질문에 대한 답변을 생성하는 질문 응답 작업에서 문장의 문맥을 이해하여 답을 도출하는 등 다양한 분야에 사용될 수 있음

[표 4-14] RoBERTa의 활용분야

활용분야	주요내용
텍스트 분류	뉴스 분류, 감정 분석, 스팸 탐지
질의응답 시스템	사용자의 질문에 대한 정확한 답변 찾기 (챗봇)
기계 번역	다양한 언어 간의 텍스트 번역
요약	자연스러운 대긴 문서 또는 기사의 요약화형 인터페이스를 위한 챗봇 개발
개체명 인식	텍스트에서 사람, 장소, 조직 등의 명사 식별

ㅇ Angin 등(2023)은 SDGs 관련성에 따라 보고서를 성공적으로 분류하기 위해 NLP 기반 텍스트 분류 작업에 최적화된 다양한 기계 학습 접근 방식을 조사하였으며, RoBERTa가 높은 성능을 보임[87)]

- NLP 연구 및 SDG의 본질에 대한 통찰력을 도출하는 연구를 지원하는 것을 목표로 하는 공개 데이터 세트인 OSDG 커뮤니티 데이터세트를 사용하여 광범위한 실험을 수행
- 미세 조정(fine tuning)된 RoBERTa 기반 분류 모델이 SDG와의 관련성을 탐지하기 위한 대규모 문서 컬렉션의 자동화된 처리에 높은 성능을 달성

ㅇ 정부입법 제·개정령안과 의원입법안에 대하여 평가인력의 침해평가 업무를 지원하도록 AI 기반 개인정보 침해평가 시스템이 개발[88)]

- 신규 양식의 제· 개정내용은 4종의 개인정보와 정보주체, 개인정보처리자, 개인정보처리 목적으로 나누어지며 모두 학습된 개체명 인식기를 통해서 추출
- 개체명 인식기는 xlm-roberta-large 사전언어 모델로 부터 Fine-tuning 되었으며 8종 개체명에 대하여 micro F1-score는 0.707이었으며, 4종 개인정보의 micro F1-score는 0.737로 나타남
- AI 시스템의 도입을 통해 새로 세분화되어 양식이 변경되었음에도 불구하고 업무 효율성을 향상시킬 수 있었으며 신속하게 침해평가를 진행할 수 있도록 평가업무를 지원

87) Angin, M., Taşdemir, B., Yılmaz, C. A., Demiralp, G., Atay, M., Angin, P., & Dikmener, G. (2022). A RoBERTa Approach for Automated Processing of Sustainability Reports. Sustainability, 14(23), 16139.

88) 채정민, & 김효정. (2023). AI 기반 개인정보 침해평가 시스템: 정부입법 및 의원입법안에 대한 NLP 활용. 정보과학회 컴퓨팅의 실제 논문지, 29(12), 545-554.

ㅇ 자체평가 등급설정 지원에 적합한 모델 등에 대한 문헌연구를 통해 등급 다중분류가 가능한 머신러닝, 딥러닝 모델을 모두 활용해보고 성능과 전문가 평가결과를 고려하여 최종 활용 모델을 결정하고자 함

- 랜덤포레스트는 다중분류 가능 모델로, 그룹 내 의견의 등급을 동일하게 예측하는 의사결정 나무 알고리즘 구성 후 투표 등으로 분류

- SVM은 주어진 데이터집합을 토대로 카테고리로 분류하는 이진 선형분류모델로, 가장 큰 폭의 경계를 찾아 분류의 정확도 제고

- XG Boost는 랜덤포레스트 모델을 기반으로 하되, 등급별 데이터 불균형을 고려하여 S/A와 B/C 그룹을 구분 후 S와 A 등급을 구분하는 형태의 2-layer 모델로 구성

ㅇ 딥러닝 모델로는 KoBERT를 기본적으로 활용하되, 결과의 성능 비교를 위해 RoBERTa를 활용한 결과도 검토

- KoBERT는 트랜스포머 인코더를 여러개 쌓아 구성된 모델로, 2천만개 이상 한국어 뉴스와 500만개 이상 문장으로 사전학습하여 임베딩 성능이 향상됨

- Self-attention을 통해 단어간 관계를 학습 후 양방향 문맥을 학습할 수 있으며, 가장 높은 확률을 가지는 등급을 출력하는 방식으로 정확도를 제고할 수 있어 활용함

[표 4-15] 자체평가 등급 설정지원 모델 활용계획 개요

(딥러닝) KoBERT, RoBERTa
◆ **(개요)** 문맥과 단어의 의미 기반 텍스트 이해 ◆ **(장점)** 문장, 문서의 맥락을 이해하는 데 효과적/ 낮은 구축비용 ◆ **(단점)** 결과를 도출한 원리에 대한 설명 불가 / 모델 보완에 한계 ◆ **(활용방향)** 최소한의 전처리를 토대로 KoBERT 모델을 기본으로 하되, 성능비교를 위해 RoBERTa 모델도 병행하여 활용
(머신러닝) 랜덤포레스트, SVM, XG Boost
◆ **(개요)** 단어(n-gram) 빈도수 기반 텍스트 이해 ◆ **(장점)** 해당 결과를 도출한 원리(특징단어, PMI) 해석 가능 / 모델 보완 용이 ◆ **(단점)** 데이터가 변동마다 단어뭉치 재구성, 불용어 제거, 모델 구축 등 매번 반복해야함 / 의미적 유사성과 맥락 이해에 한계 ◆ **(활용방향)** tri-gram 기반 불용어 제거와 의견분석을 통해 가능한 머신러닝 모델을 활용하여 그 성능을 비교하고, 데이터 구조 등 정보를 토대로 등급을 예측하는 로직에 대해서도 분석

7. 실험을 통한 모델 보완사항 도출[89)]

(1) 딥러닝 모델 분석

□ 분석 개요

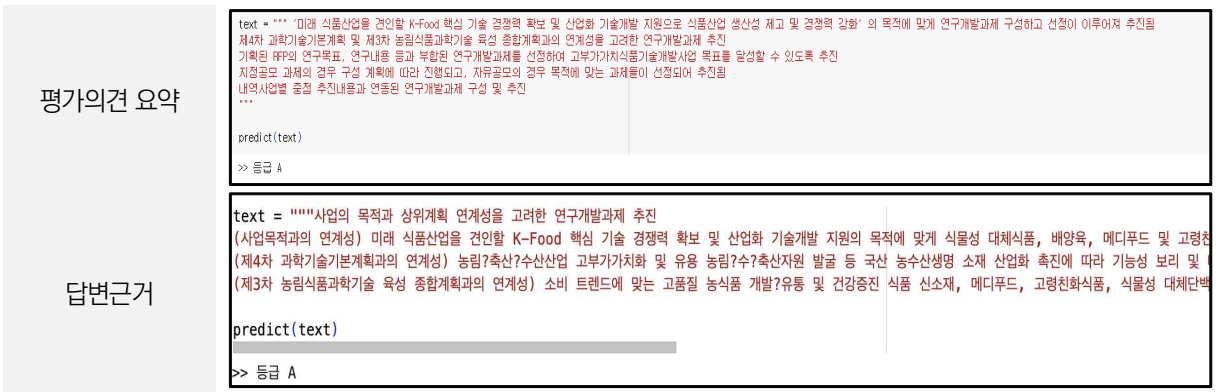
- KoBERT 모델과 RoBERTa-base 모델은 학습과정에서 학습데이터와 평가데이터의 비율을 8:2로 분리하고, 최대한 많은 텍스트를 학습하기 위하여 실험환경에서 분석 가능한 최대 문장길이였던 512자로 설정하여 KoBERT는 20번(epoch), RoBERTa는 30번 반복 학습하였음

- 딥러닝 실험에서는 문장 원문을 최대한 보존하여 문서의 뉘앙스와 규칙을 학습하는 것이 필요하기 때문에 특수기호만 제외 후 학습을 진행하였으며, 성과외 추진과정 지표에 대하여 세부지표별이 아닌 통합 모델로 한번에 학습을 진행함

- 딥러닝 모델로 성과외 추진과정 지표에 대해서 학습과 모델링을 할 때에는 머신러닝에 비해 상대적으로 성능이 좋은 점, 지표별로 나누었을 때 평가의견의 수가 현저히 적어지게 되는데 성과외 지표의 특성 상 평가의견의 문장 또한 길이가 충분치 않은 점을 고려하여 정확도와 효율성 제고를 위해 통합 모델을 구축하여 학습을 진행함

□ 성과지표 평가의견에 대한 분석

- 성과 지표에 대해서는 문장의 길이가 충분히 확보되어 종합평가결과와 의견이 드러나 있는 평가결과 요약(1)과 평가근거를 종합적으로 포함하고 있는 답변근거(2)로 나누고 각각에 대한 학습과 평가 시 성능을 비교하는 실험을 추가적으로 수행함



[그림 4-9] 추진과정 지표 평가의견 요약과 답변근거 데이터의 비교와 예측결과 예시 (KoBERT)

89) 아래의 각 실험은 다양한 세부사업을 대상으로 반복적으로 진행되었으며, 그 과정에서 추가적인 불용어 제거와 결과 비교, 코드 보완/수정 등을 거쳐 최종 조정된 결과를 본 보고서에 포함함

- 성과지표의 KoBERT모델 실험 결과로 비교해보면, 요약의 학습정확도는 99.3%, 평가 정확도는 90%, 답변근거의 학습 정확도는 99.3%, 평가 정확도는 63.8%로 나타났으며, 이는 평가의견과 지침에 따른 평가기준이 담겨있는 평가결과 요약이 평가등급 설정과 관련이 있는 비교적 더 유효한 텍스트 데이터라고 해석하고 평가결과 요약만 채택함

KoBERT 우수성

학습 정확도:82.5%, 평가 정확도:83%

```
# B(1)
predict("""해당 목표는 전주기 기술지원 이후 수출국가 및 품목(수출기반지수), 수출액(수출액) 등 경제적 성과가 성과지표로 설정되어 있음 (사업화 양적성과)
사업화 성과는 작년 대비 409.3% 증가하였으며(42.8억→218.2억), 특히 수출액이 904.6% 증가함(21.6억→217.4억)
(사업화 10억 원 당 성과) 2019년 0.19에서 2021년도 3.12로 비약적으로 증가하였으며, 수출이 0.39(2020년)에서 3.11(2021년)로 증가한 것에 기인함
수출 사업화 성과를 살펴보면 ' ICT기반 수출용 깻잎과 아스파라거스의 병해충 방제기술 개발 '(19년 시작), '유자제품 수출확대 원료생산 안정화 및 제품 고급?다
첫 사업화 성과가 나타나기 시작할 경우, 연속적인 성과 창출이 기대되므로, 아직 성과가 나타나지 않은 과제들도 전주기적 관리가 지속적으로 이뤄져야 할 필요가 있음
사업화 대상 과제들은 대부분이 기술료 성과도 함께 나타나고 있는 과제들로서 본 사업의 목적인 시장 분석→기술이전 실시→사업화 수출의 전주기를 따르고 있어 향후 사
또한, 동 사업이 지원하는 기업의 경우 대부분 중소·영세기업으로, 중소 영세기업이 수출을 위한 기반을 마련하여, 해외 수출 성과가 나타는 것은 매우 어려우나, 본
(사업화 효율성 분석) 2020~2021년도 농식품수출비즈니스전략모델 구축사업에서 창출된 10억원 당 사업화 매출액(수출액)의 평균은 1.127으로 나타남
사업화 매출액 비교분석DB는 2020년, 2021년 상위평가보고서에서 사업화 매출액을 성과지표로 활용하고 있는 사업을 전수조사하여 예산과 매출액을 작성하였으며, NT
사업화 매출액을 성과지표로 활용하는 사업은 15개로 집계되었으며, 예산 평균은 302억원, 사업화 매출액 평균은 31.9억 원으로 산출되어 사업화매출액 지수는 0.10
농식품수출비즈니스전략모델 구축사업은 1.127(2020~21)로 산출되어 산업부 에너지자원순환기술개발 2.981 다음으로 높은 수준으로 나타났으며, 사업의 효율성이 상

>> 등급 A
```

KoBERT 핵심성

학습 정확도:99.8%, 평가 정확도:78.6%

```
# S(3)
predict("""직접적인 사업 및 성과목표와 비교하였을 때, 성과목표 1의 경우 3개년간 목표치를 100% 달성해왔으며, 성과목표 2의 경우에도 목표치를 100% 달성하
'천리안위성 2A호 기반 고품질 위상정보 서비스 기반 구축'을 위해 설정한 1개 성과지표를 목표 대비 3년간 100% 달성. 동 사업의 성과는 기상위성 활용서비스 기술
동 사업의 성과로 다분야 활용을 위한 기상위성과 미래기술을 융합한 맞춤형 기상 위성정보 산출 및 온실가스 감시체계 구축을 위한 기상·기후 변화 감시체계 마련, 고

>> 등급 S
```

RoBERTa 우수성

학습 정확도:99.8%, 평가 정확도:78.6%

```
predict("""해당 목표는 전주기 기술지원 이후 수출국가 및 품목(수출기반지수), 수출액(수출액) 등 경제적 성과
사업화 성과는 작년 대비 409.3% 증가하였으며(42.8억→218.2억), 특히 수출액이 904.6% 증가함(21.6억→217.4억)
(사업화 10억 원 당 성과) 2019년 0.19에서 2021년도 3.12로 비약적으로 증가하였으며, 수출이 0.39(2020년)에서
수출 사업화 성과를 살펴보면 ' ICT기반 수출용 깻잎과 아스파라거스의 병해충 방제기술 개발 ' (19년 시작), '
첫 사업화 성과가 나타나기 시작할 경우, 연속적인 성과 창출이 기대되므로, 아직 성과가 나타나지 않은 과제들도
사업화 대상 과제들은 대부분이 기술료 성과도 함께 나타나고 있는 과제들로서 본 사업의 목적인 시장 분석→기술
또한, 동 사업이 지원하는 기업의 경우 대부분 중소·영세기업으로, 중소 영세기업이 수출을 위한 기반을 마련하
(사업화 효율성 분석) 2020~2021년도 농식품수출비즈니스전략모델 구축사업에서 창출된 10억원 당 사업화 매출액
사업화 매출액 비교분석DB는 2020년, 2021년 상위평가보고서에서 사업화 매출액을 성과지표로 활용하고 있는 사업
사업화 매출액을 성과지표로 활용하는 사업은 15개로 집계되었으며, 예산 평균은 302억원, 사업화 매출액 평균은
농식품수출비즈니스전략모델 구축사업은 1.127(2020~21)로 산출되어 산업부 에너지자원순환기술개발 2.981 다음으

>> 등급 A

predict("""2019년~2020년 특허 성과는 총 38건(출원특허 20건, 등록특허 18건)이며, SMART값은 19년 3.64, 20년
원천기술 창출의 기반이 되는 특허는 사업·성과목표와의 부합성이 높은 편으로 판단되나 성과목표를 66.2%(19년)

>> 등급 A
```

RoBERTa 핵심성

학습 정확도:97.17%, 평가 정확도:62.96%

```
predict("""성과목표인 외래생물 관리의 과학기술적 핵심인 국내 유입종의 생태특성 연구를 생태계교란종 전종(포유

Truncation was not explicitly activated but 'max_length' is provided a specific value, please use 'truncati
'A'

predict("""인공지능이 적용된 뇌신경생물 실험기기 시작품을 1단계에서부터 제작하는 것으로 연구개시 2차년도부터

'A'

predict("""직접적인 사업 및 성과목표와 비교하였을 때, 성과목표 1의 경우 3개년간 목표치를 100% 달성해왔으며,
'천리안위성 2A호 기반 고품질 위상정보 서비스 기반 구축' 을 위해 설정한 1개 성과지표를 목표 대비 3년간 100%
동 사업의 성과로 다분야 활용을 위한 기상위성과 미래기술을 융합한 맞춤형 기상 위성정보 산출 및 온실가스 감시

'A'
```

[그림 4-10] KoBERT와 RoBERTa의 평가등급 예측 결과

- KoBERT 성과지표 등급예측 결과를 보면, 우수성의 학습정확도와 평가 정확도가 오차범위 내로 준수한 성능을 보인 반면, 핵심성의 경우에는 학습 성능이 우수성 비해 더 높게 나타났으나, 학습 정확도와 평가 정확도의 차이가 다소 크게 나타남

- 또한 위의 [표]와 같이 등급 예측 결과를 일일이 검토한 결과 핵심성과 우수성 모델 모두 B 등급을 예측해내지 못해 S와 A 등급을 예측함으로써 정확도가 높아진 수치임을 알 수 있음

- 비교를 위해 RoBERTa의 경우 우수성과 핵심성 모두 97.2%로 성과외 지표보다 더 좋은 성능을 보여줌. 다만 평가 정확도가 60-70%대로 학습 정확도만큼을 못 따라오는 것을 확인할 수 있었음

- 구체적으로 RoBERTa 모델일 때에는 학습 정확도가 KoBERT에 비해 다소 떨어졌으며, 그에 따라 평가 정확도도 낮게 나타남

  - KoBERT 모델은 한국어 학습과 예측에 비교적 효과적인 반면, 한국어 맥락 이해에 한계가 있는 RoBERTa 모델의 특성이 학습과 평가 시 성능에도 작용한 것을 알 수 있음

- 따라서 성과부문의 평가의견 데이터를 분석할 때 딥러닝 모델을 활용할 경우, 한국어 분류 모델에 강건한 KoBERT 모델을 활용하는 것이 적절할 것이라는 결론을 내림

(2) 머신러닝 모델 분석

□ 분석 개요

- 랜덤포레스트, SVM, 나이브베이즈 세 종류의 모델을 활용하였으며, 랜덤포레스트 모델은 위험단어 제거 전(RF1), 위험단어 제거 후(RF2), 위험단어 제거와 하이퍼파라미터튜닝 후(RF3) 3가지 버전으로 나누어 수행함

  - 위험단어란 데이터누수 위험단어를 뜻하며, 등급정보(정답데이터)를 그대로 담고 있는 단어들을 의미함 (예를 들어, 10점 부여, S등급 부여, A등급 부여, 1.0점 부여, 가중치 0.8 부여)

- SVM의 경우 ovoclassifier, rbfkenerl 두 개의 옵션으로 나누어 진행했고, 하이퍼파라미터튜닝은 ovoclassfier에 대해서만 진행하였음

□ 성과 지표 평가의견에 대한 분석

- 일반적으로 랜덤포레스트 위험단어 제거 전이 성능수치는 높게 나왔으며, 이는 지침에서 제시하는 등급별 기준에 따라 성과의 우수성과 핵심성 평가의견을 작성하기 때문에 지침어의 영향이 크게 작용한 것으로 보임

[표 4-16] 성과지표 평가의견에 대한 머신러닝 정확도

구분	랜덤포레스트 (RF1)	랜덤포레스트 (RF2)	랜덤포레스트 (RF3)	나이브베이즈	SVM ovoclassifier	SVM rbfkernel
성과의 우수성	0.8462	0.8308	0.8231	0.7923	0.8231	0.8462
성과의 핵심성	0.7462	0.6769	0.6462	0.6307	0.6461	0.6

□ 성과외 지표 평가의견에 대한 분석

o 성과외 지표에 대해서는 공통적인 불용어를 제거한 데이터로 학습했을 때(모델1)와 세부지표별 불용어와 어휘사전을 별도로 구축해서 학습했을 때(모델2)로 나누어 실험을 진행함

[표 4-17] 성과지표 평가의견에 대한 머신러닝 정확도

		랜덤포레스트 (위험단어 제거 전)	랜덤포레스트 (위험단어 제거 후)	랜덤포레스트 (하이퍼파라미터 튜닝 후)	나이브베이즈	SVM (ovoclassifier)	SVM (하이퍼파라미터 튜닝 후)	SVM (rbfkenerl)
투입	1	0.7442	0.6977	0.6512	0.6512	0.6512	0.6047	0.6512
	2	0.7441	0.7441	0.6744	0.5116	0.6279	0.7674	0.7906
과제관리	1	0.7544	0.6491	0.6316	0.6140	0.6316	0.6140	0.6316
	2	0.7719	0.6666	0.6666	0.6140	0.6491	0.6491	0.6315
위험요소관리	1	0.7857	0.6429	0.6429	0.6429	0.6429	0.6429	0.5952
	2	0.7857	0.6666	0.6428	0.5	0.6428	0.6666	0.6428
수혜자	1	0.7143	0.6	0.6286	0.5426	0.6286	0.5143	0.5715
	2	0.6285	0.6285	0.6285	0.5428	0.6285	0.6285	0.5714
환류	1	0.7368	0.4737	0.4737	0.4211	0.5263	0.5263	0.5263
	2	0.5263	0.6315	0.5263	0.5789	0.4736	0.4210	0.6842
성과분석과 환류계획의 구체성	1	0.7780	0.7288	0.7288	0.6441	0.7288	0.7288	0.5932
	2	0.7288	0.7288	0.7288	0.5932	0.7288	0.7118	0.6779

o 모델1의 경우 일반적으로 대부분의 지표에서 랜덤포레스트 위험단어 제거 후, 하이퍼파라미터 튜닝 전(RF2)의 모델이 성능이 좋은 것을 확인하였으며, 성능제고를 기대하고 지표별 어휘사전을 구축해 진행한 실험에서는 실제 대부분의 지표에서 이전 실험보다 다소 향상된 성능을 보임

o 결론적으로 머신러닝 모델 기반의 서비스를 제공할 경우 랜덤포레스트 위험단어 제거 후, 하이퍼파라미터 튜닝 전 모델과 SVM 모델을 지표별로 성능에 따라 적용하는 것이 가장 바람직하다고 할 수 있음

(3) 부스팅 계열 모델 분석

□ 분석 개요

ㅇ 부스팅 계열 모델은 머신러닝 모델을 2개의 layer로 쌓아서 보다 예측 정확도를 높이기 위한 것으로 본 연구에서는 XGBoost와 LightGBM 두 개의 모델을 다양한 옵션과 함께 분석에 활용함

- 부스팅 계열 모델의 경우 0과 1의 이진분류에서만 활용가능하기 때문에 S/A와 B/C등급을 구분할 수 있는 단어들을 찾아 1단계 모델로 활용하고, 이후 S와 A등급을 분류하는 2단계 모델로 구축함

ㅇ XGBoost의 경우 랜덤포레스트 모델의 복합모델이므로 랜덤포레스트와 동일하게 위험단어 제거 전후와 하이퍼파라미터 튜닝 전후로 나누어 3가지 옵션의 모델을 활용함

ㅇ LightGBM에 대해서도 하이퍼파라미터튜닝 전후를 비교함

□ 성과 지표 평가의견에 대한 분석

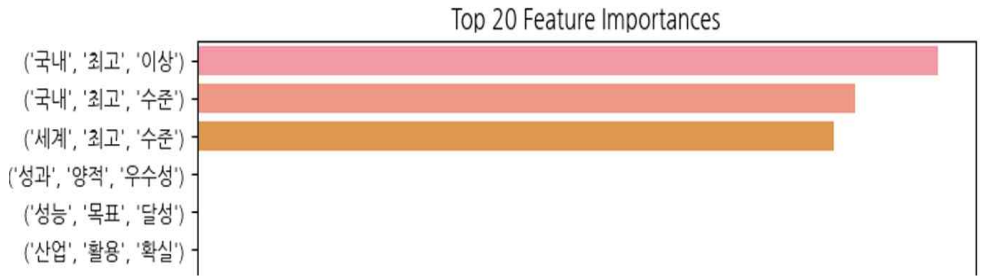
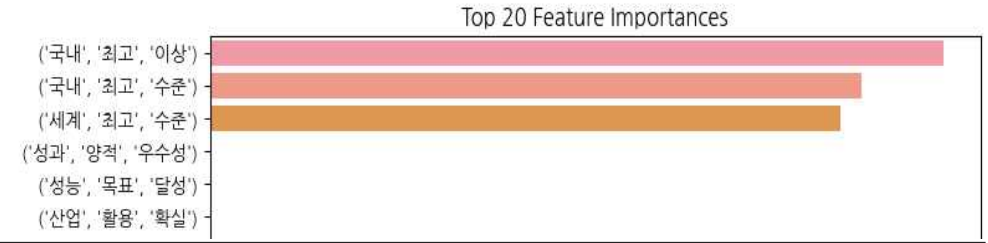
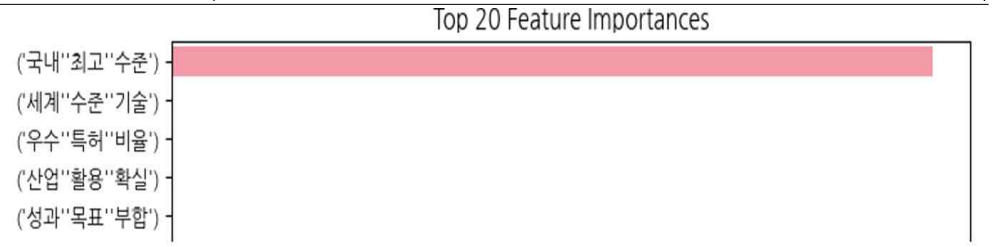
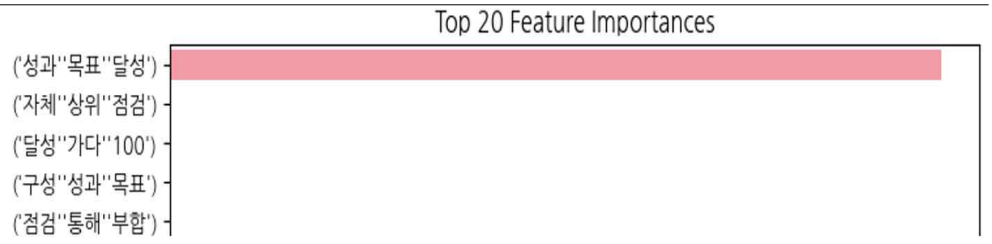
[표 4-18] 부스팅 계열 모델 성능 비교 (성과)

구분	XGBoost (위험단어 제거 전)	XGBoost (위험단어 제거 후)	XGBoost (하이퍼파라미터 튜닝 후)	lightGBM (기본모델)	lightGBM (하이퍼파라미터 튜닝 후)
성과의 우수성	1.0	1.0	1.0	1.0	1.0
성과의 핵심성	1.0	1.0	1.0	1.0	1.0

ㅇ 데이터 수가 적은 B, C 등급의 데이터를 제외하고, S, A등급의 데이터를 부스팅계열 모델에 학습한 결과, 랜덤포레스트, SVM모델보다 성능향상이 있는 것으로 판단됨

ㅇ 이런 성능향상에 근거해서 B, C등급과 S, A등급 그룹을 첫 번째 이진 분류모델에서 정확히 구분해낼 수 있는 trigram을 추출해낼 수 있다면 부스팅 계열 모델을 활용하는 것이 단순 머신러닝 모델을 활용하는 것보다는 나은 결과를 낼 수 있음을 확인함

ㅇ 두 종류의 부스팅 계열 모델 중 어떤 것이 적합한지 살펴보기 위해 각 모델별 주요 20개 유효한 단어 뭉치(trigram)을 살펴봄

XGBoost 성과의 우수성	
XGBoost 성과의 핵심성	
lightGBM 성과의 우수성	
lightGBM 성과의 핵심성	

[그림 4-11] 부스팅 계열 top 20 feature importances (주요단어) 비교

- 같은 정확도 1.0이 나왔지만 모델 성능에 영향을 주는 주요 feature들을 구분했을 때, XGBoost가 보다 지침과 의도에 맞는 등급분류결과를 보여주는 것을 알 수 있음

- 다만, 여전히 부스팅 계열 모델의 정확도가 완벽히 100%가 나온 결과에 대해서는 그만큼 데이터의 수가 적고 데이터가 단순하다는 것을 반영하는 것이므로 추후 B, C등급의 예측 수요가 높아질 경우 모델로 인한 편향이 매우 커질 수 있음을 의미하며, 이는 활용에 치명적인 결함으로 보임

□ 성과외 지표 평가의견에 대한 분석

ㅇ 성과외 지표에 대해서도 성능이 1.0으로 나왔으며, 이는 성과 지표에 대한 분석결과와 마찬가지로 데이터의 등급 분류가 매우 단순하여 성능 향상과는 별개로 활용에 한계가 있음을 의미함

[표 4-19] 부스팅 계열 모델 성능 비교 (성과 외)

	XGBoost (위험단어 제거 전)	XGBoost (위험단어 제거 후)	XGBoost (하이퍼파라미터 튜닝 후)	lightGBM (기본모델)	lightGBM (하이퍼파라미터 튜닝 후)
투입	1.0	1.0	1.0	1.0	1.0
과제관리	1.0	1.0	1.0	1.0	1.0
위험요소 관리	1.0	1.0	1.0	1.0	1.0
수혜자	1.0	1.0	1.0	1.0	1.0
환류	1.0	1.0	1.0	1.0	1.0
성과분석과 환류계획의 구체성	1.0	1.0	1.0	1.0	1.0

□ 추가 실험 : 지표별 지침의 등급부여 기준 단어 활용

ㅇ 성능만을 두고 보았을 때, 위험단어 제거 전의 모델이 성능이 높았던 것을 고려하여 지표별 지침의 등급부여기준을 학습시키는 추가 실험을 진행하였으며, 이 때 학습 데이터가 등급별로 현저히 적기 때문에 적은 데이터로 학습을 해서 다량의 데이터에 대한 시뮬레이션과 예측을 수행하는 few shot learning을 시도함

- 데이터가 유의미하더라도 적은 수의 데이터만으로 학습을 진행하는 경우 모델이 충분히 등급간 차이를 학습하기 어렵기 때문에 성능이 떨어지는 경우가 많음. 때문에, 주로 많은 양의 데이터로 사전학습되어있는 딥러닝 모델에서 few shot learning을 적용하는 것이 일반적임

- 하지만, S/A의 등급분류에 머신러닝 모델을 사용하는 상황에서 단순히 S/A와 B/C를 구분하기 위해 추가적으로 딥러닝 모델을 사용하는 것은 효율적이지 못해 지침을 학습시킨 머신러닝 모델을 통해 S/A와 B/C를 분류해보고, 성능을 검토함

ㅇ 그 결과, 6개의 성과외 지표 중 [투입], [과제관리], [수혜자]에 대해 각 지표에 따른 등급을 XGBoost모델에 학습시킨 뒤 나머지 데이터에 대해 S/A와 B/C를 구분하는 과제를 수행하였을 때 투입 분류 정확도 0.02, 과제관리 분류 정확도 0, 수혜자 분류 정확도 0.07으로 매우 저조한 성적을 보임

- 성과 부문에서는 여전히 1.0의 성능을 유지했으나, 추후 활용을 고려하였을 때 지침의 평가기준 단어를 그대로 분류기준으로 사용하는 것은 모델 구축의 실효성이 떨어지고 잘못된 평가의견에서 bias가 더 클 것이라는 연구진의 판단하에 부스팅 계열 모델 활용은 아직 한계가 있음으로 결론을 지음

- 추후 상대평가 등을 통해 등급별 데이터가 고르게 분포하거나, B, C등급에 대해서도 적은 수의 데이터로나마 학습이 가능한 수준으로 축적이 가능하다면 few shot learning을 통해 부스팅 계열의 이진분류모델을 활용해 볼 수 있음을 본 실험의 시사점으로 도출함

제3절 평가 및 활용성 검증

- 모델 탐색을 통해 선별된 딥러닝, 머신러닝 계열 모델에 대한 성능비교를 위해 2022-2023 데이터를 공통된 기준으로 전처리하고 평가 정확도를 비교해본 결과, 2022년 선행연구에 비해 모두 성능이 제고된 것을 확인할 수 있었음

  - KoBERT의 경우 성과의 우수성에서 0.82, 성과외 통합모델에서 0.9의 정확도를 보이며 비교적 높은 성능을 보였으나, 성과의 핵심성 부문의 정확도가 0.73으로 차이가 있는 것으로 나타남

  - 머신러닝계열의 모델은 성과의 우수성에서 모두 80% 이상의 정확도를 보여 모델 구축 비용을 고려하지 않는다면, 서비스 제공 시 딥러닝과 머신러닝 모델 여러개를 동시에 제공하고 평가의견 특성에 따라 분석할 수 있다면 가장 바람직할 것으로 결론을 내림

[표 4-20] 딥러닝, 머신러닝 모델별 성능 비교

평가부문	'22년연구 모델성능[2]	딥러닝		머신러닝				
		KoBERT	RoBERTa	RF1	RF2[3]	RF3	SVM (ovo)[4]	SVM (rbf)
성과의 우수성	0.56[1]	0.82	0.74	0.85	0.83	0.82	0.82	0.85
성과의 핵심성		0.73	0.63	0.75	0.68	0.65	0.65	0.60
과제관리	0.88	0.9	0.68	0.77	0.67	0.67	0.65	0.63
환류계획	0.77			0.73	0.73	0.73	0.73	0.68
수혜자	0.71			0.63	0.63	0.63	0.63	0.57
위험요소관리	0.74			0.79	0.67	0.64	0.64	0.64
투입	0.68			0.74	0.74	0.67	0.63	0.79
환류	0.73			0.53	0.63	0.53	0.47	0.68

주1) 정확도 수치 0.56은 100개 validation data에 대한 등급예측 시 56개 등급을 정확히 예측했음을 의미함
주2) '22년 일부 사업(133개)의 자체평가보고서를 학습해 시도해본 모델의 성능을 의미함
주3) 랜덤포레스트2는 위험단어*를 불용어로 제거 후, 랜덤포레스트3은 2에서 하이퍼파라미터튜닝 후를 의미함
* 위험단어는 등급정보(정답데이터)를 그대로 담고 있는 단어들을 의미(예: 10점 부여, S등급 부여, 1.0 부여, 0.8 부여 등)
주4) SVM ovo는 overclassifier 옵션에 하이퍼파라미터튜닝 후를 의미함

o 모델 구축에 활용한 데이터와 비교적 유사한 2022-2023 데이터를 20개 랜덤 샘플링하여 실제 상위점검을 거쳐 확정된 평가등급과 비교해 보았을 때, KoBERT의 정확도가 모델 중 가장 높은 것으로 나타남

- 여기서 활용된 데이터는 모델 학습과 학습 후 평가 시 전혀 활용되지 않은 평가·검증용으로 신규 평가의견 데이터가 입력되었을 때 등급 예측의 정확도를 예측해보기 위한 실험임

[표 4-21] 22-23년 검증용 데이터 20개에 대한 실험 결과 성능 비교

.	KoBERT		RoBERTa		랜덤포레스트		SVM	
성과의 우수성	80%	(16/20)	60%	(12/20)	80%	(16/20)	85%	(17/20
성과의 핵심성	85%*	(17/20)	40%	(8/20)	50%	(10/20)	45%	(9/20)
과제관리	95%	(19/20)	65%	(13/20)	45%	(9/20)	50%	(10/20)
환류계획	100%	(20/20)	75%	(15/20)	75%	(15/20)	75%	(15/20)
수혜자	100%*	(20/20)	30%	(6/20)	60%	(12/20)	50%	(10/20)
위험요소관리	95%	(19/20)	50%	(10/20)	50%	(10/20)	55%	(11/20)
투입	95%	(19/20	65%	(13/20)	65%	(13/20)	75%	(15/20)
환류	100%	(19/19	68.4%	(13/19)	63.2%	(12/19)	68.4%	(13/19)

주1) *은 B 또는 C등급을 실험 중 1회 이상 분류해낸 모델을 의미 (*이 없는 모델은 B, C등급이 없거나 분류 실패)
주2) 환류는 데이터 부족으로 총 19개 데이터를 샘플링하여 검증함

o 모델 구축과 사전평가 시 활용했던 데이터와 길이, 데이터 수, 평가의견의 작성 행태 등이 전혀 다른 신규데이터로 볼 수 있는 2019-2021 데이터를 전수 활용하여 모델별 성능이 강건하게 유지되는지 확인함

o 그 결과, 상대적으로 길이가 짧고 등급을 판단하기에 불충분한 의견이 대다수 분포하고 있는 2019-2021 데이터에 대해서는 모든 모델이 등급분류에 안정된 성능을 보이지 못하는 것으로 나타남

o 그 중에서도 B, C등급에 해당하는 평가의견에 대하여 제대로 등급을 예측해 낸 모델은 극히 일부뿐이었으며, 랜덤포레스트모델의 정확도가 비교적 높게 나타났음

- 랜덤포레스트모델의 성능은 불용어 제거를 포함한 데이터의 전처리와 유효한 단어의 포함에 따라 영향을 많이 받게 됨을 고려하였을 때, 전처리가 충분히 되지 않은 평가의견 raw data를 입력할 가능성이 높은 사용자 검색 시에는 정확도의 신뢰성이 현저히 떨어질 수 있음을 시사함

[표 4-22] 19-21년 검증용 데이터 전체와 30개에 대한 실험 결과 성능 비교

	KoBERT		랜덤포레스트		SVM	
	전체	샘플링	전체	샘플링	전체	샘플링
성과의 우수성	66.3% (818/1234)	80% (24/30)	67.1%* (828/1234)	77% (23/30)	64.4% (795/1234)	80% (24/30)
성과의 핵심성	50.5%* (626/1240)	70%* (21/30)	51.1%* (634/1240)	46.7% (14/30)	53.2% (659/1240)	53.3% (16/30)

주1) '19-'21년은 성과유형별 등급이 부여되지 않아, 등급정보(정답)를 대신하여 전문가(KISTEP)가 부여한 등급과 비교를 통해 정확도를 평가

주2) *은 B 또는 C등급을 실험 중 1회 이상 분류해낸 모델을 의미 (*이 없는 모델은 B, C등급이 없거나 분류 실패)

V. 결론 및 시사점

제1절 결론

□ **본 연구는 선행연구와 NTIS 등 사례를 토대로 국가R&D 사업평가의 산출물을 토대로 인공지능 모델을 구축하고 이를 보조수단으로 활용해 부가적인 정보를 효율적으로 제공할 수 있는 세부 과업 두 가지를 선정하고, 효과적인 모델을 탐색하기 위해 다양한 실험과 검증을 수행함**

ㅇ 성과지표설정지원 서비스는 SBERT와 KeyBERT 모델을 기반으로 검색조건별 관련사업 추천 모델을 구축하였음

- SBERT 모델은 사업명(세부사업명 기준), 사업목적, 전략목표, 사업 세부내용(내역사업별 연구활동내용), 통합내용을 기준으로 사용자가 선택적으로 관련사업을 검색하고 추천받을 수 있도록 서비스 기획이 필요함

- KeyBERT 모델은 키워드를 1개에서 최대 5개까지 입력할 수 있도록 검색창을 만들어 사용자가 입력하는 키워드와 가장 유사한 전략계획서를 찾아낼 수 있어야 하며, 이 때 키워드의 diversity 파라미터값 조정을 통해 사용자가 원하는 결과를 찾아낼 수 있도록 지원할 필요가 있음

- 두 모델은 유클라디안 유사도 기준으로 어느 정도 안정적 수치의 성능을 보였으나, 추천결과에 대한 전문가 평가결과를 고려하였을 때에는 여전히 추가 보완이 필요하며, 가능한 보완사항은 추천결과의 n-gram을 다변화하여 비교·검토 후 불용어를 추가적으로 제거하거나, 사람이 유사도를 판단할 때와 마찬가지로 사업목적이나 내용이 유사할 경우 가중치를 부여하는 방식으로 모델을 고도화하거나 보완하는 것을 검토할 수 있음

ㅇ 평가등급설정지원 서비스의 경우 성능과 비용 효율성 측면에서 딥러닝 모델인 KoBERT 모델을 기반으로 성과와 성과외 부문 모두 모델을 구축하는 것이 적합해보이나, 여전히 과거 데이터의 오류로 인해 B등급과 C등급을 제대로 예측해낼 수 없다는 치명적 한계를 고려하면 내부 활용을 우선 고려하는 것이 적합함

- 특히, 2024년 이후부터 부처 단위의 상대평가가 적용되어 등급 분포가 보다 고르게 조정될 것으로 예상되는 시점에서 본 모델 활용의 시의성을 재검토할 필요가 있음

- 추후 B등급과 C등급의 평가의견 데이터가 늘어나면 few shot learning을 통해 모델의 예측 정확도를 지속적으로 점검하고 테스트한 뒤에, 데이터가 충분히 축적되고 난 뒤 본격적인 도입과 활용을 고려하는 것이 적절함

□ **본 연구의 결과와 한계를 종합하여 향후 본격적인 AI 기반 국가R&D평가지원체계 구축 시에는 다음에 대한 고려가 필요함**

ㅇ 첫 번째로, 성과지표 추천 서비스의 유용성을 제대로 검증할 필요가 있음

- 본 서비스의 목적은 전략계획서 성과목표와 성과지표 설정 시 사용자가 참고할 만한 부가정보를 제공하는 것이므로 추천된 사업이 정말 대상사업과 관련이 있다고 판단할 수 있는지 사용자, 전문가, KISTEP과 과기정통부 평가 담당자 등의 철저한 검증이 수행되어야 할 것임

ㅇ 두 번째로, 성과지표 추천 서비스의 경우 전략계획서 제도가 도입된 기간이 짧아 상대적으로 관련사업 pool이 크지 않음을 고려하여 성과목표지표계획서 등 데이터의 시계열을 확대할 것을 검토할 필요가 있음

- 다만, 이 경우 3가지에 대한 추가 검토가 선행되어야 하는데 첫째는, 부처 수요 등으로 사업 착수 이후 수정된 전략계획서 데이터를 하나로 통합하여 일관되게 관리하여야 하고, 둘째는, 사업 통합과 구조조정 등으로 사업 이력 추적이 가능해야 하며, 마지막으로는, 평가제도 개선 등으로 지속적으로 바뀐 사업내용 항목과 데이터 등을 일관된 기준으로 수집하여 데이터셋을 구축할 필요가 있음

ㅇ 세 번째로, 성과목표와 지표 데이터의 표준화가 필요함

- 본 연구는 전략계획서 데이터를 잘 분석해내서 관련사업을 잘 찾아내는 모델을 탐색하고 그 결과를 검증하는 실험을 진행하였지만, 관련사업 추천 결과 데이터인 성과목표와 지표 데이터는 표준화되어 있지 않아 입력상태 그대로 출력될 수 있음

- 현재 성과지표데이터는 동일한 성과에 대해 성과지표명, 측정대상 성과의 정의, 측정 산식과 시기, 방법, 기준과 자료출처 등이 일관되어 있지 않아 추후 사용자가 활용 시에 혼란이 있을 수 있음을 고려하여 본격적인 서비스 구축에 앞서 성과지표 데이터에 대한 표준화가 선행될 필요가 있음

- 다만, 데이터의 양을 고려하여 표준성과지표 지침에 따른 성과지표 유형이나 예시 등을 참고하여 해당 데이터만이라도 일관된 기준으로 사용자가 참고할 수 있도록 추가 작업이 필요함

- o 네 번째로, 자체평가 등급 설정지원 서비스의 활용을 위해 등급간 데이터 불균형이 해소될 필요가 있음
  - 분석결과 전체 평가의견 데이터에서 B, C등급의 데이터 개수가 현저히 적었기 때문에 근본적으로는 해당 등급의 데이터가 충분히 확보되어야 하며, 그 이후에야 보완된 모델을 구축하여 S, A, B, C 등급 다중분류가 가능한 서비스 구축이 가능할 것으로 보임
- o 다섯 번째로, 딥러닝 모델 활용을 고려하여 평가의견에 대한 정제와 전처리 기준 마련이 필요함
  - 본 연구에서는 컴퓨팅 자원의 한계로 인해 학습에 활용된 문장이 최대 512자로 제한되어 평가결과 요약 데이터를 활용하였지만 해당 데이터 안에 평가등급을 구분할 수 있는 내용이 부족하여 정확도가 낮을 수 있음을 고려하여야 함
  - 따라서 추후 후속연구에서는 평가의견 전체에서 핵심적인 단어나 어절을 중심으로 요약하는 추출적 요약기법을 활용하여 요약된 원문을 등급예측을 위한 데이터로 활용하는 등 보완을 검토할 필요가 있음
  - 다만, 이 경우 요약된 원문이 해당 평가의견을 얼마나 잘 대표하는지에 대해서도 전문가와 사용자의 검증이 추가적으로 필요함
- o 마지막으로, 연구의 한계를 고려하여 다양한 보완방법을 검토할 수 있음
  - 자체평가의견 데이터의 가장 큰 한계는 등급간 데이터 수가 매우 불균형하다는 것과 평가의견에 따라 등급을 책정한 기준이 일관되지 않아 정확한 등급예측에 한계가 있다는 점임을 고려하여 학습데이터의 라벨인 등급 신뢰도 향상을 위해 평가 의견 중 우수 의견만을 모아서 적은 데이터로 학습이 가능한 few shot learning 모델을 구현해 결과를 비교 검토해볼 수 있음
  - 또한, 등급별 불균형에 대해서는 데이터 증강 기법을 다양하게 시도해볼 수 있는데, 예를 들어 문장의 순서를 바꾸거나, 유의어로 대체하는 방법 등을 시도해볼 수 있고, 한국어 문장을 영어로 번역했다가 다시 한국어로 번역하는 역번역 방법을 활용할 수 있고, 최신의 트렌드를 반영하여 생성형 LLM(Large Language Model)을 활용해 기존의 데이터나 지침 기준에 맞는 새로운 평가의견 문장 데이터셋을 만들어내는 방법 등을 검토해볼 수 있음
  - 다만, 위의 보완방법 등에 대해서도 전문가와 사용자의 검증이 필요하며, 정성적 검증으로 인해 실제 모델의 정확도가 개선되는 효과에 대해서 면밀히 측정하기 어렵다는 한계가 있을 수 있음

- 그 외에도 평가기준과 양식 변화를 고려하여 장기간 일관된 기준으로 균등한 데이터를 쌓고 활용할 수 있는 방안에 대한 검토가 이루어지고, 데이터 특성과 활용 목적, 운영비용을 고려해 모델을 구축하고, 서비스 전 성능 안정성과 목적달성 여부에 대한 면밀한 검토와 검증이 필요함

제2절 시사점

□ 본 연구는 국가R&D사업 평가에서 AI 도입을 고려해 다양한 모델링 시도와 실질적 활용 가능성을 검토해보았다는 점에서 그 의의가 있다고 할 수 있음

- 본 연구를 통해 개발된 분류 및 예측 모델을 추후 구축하여 활용한다면 평가 업무가 일부 자동화되어 업무의 효율성이 향상될 수 있으며, 평가자가 자신의 평가 의견의 적절성을 미리 확인해보고 검토해볼 수 있어 보다 객관적인 평가를 하는 데 기여할 수 있을 것으로 기대됨

- 또한 본격적으로 구축하여 활용이 가능한 시점부터 데이터와 활용 사례가 축적되고나면, R&D 사업평가를 담당하는 KISTEP과 과학기술정보통신부 차원에서도 상위점검에서 본 서비스를 통해 점검을 효율화하고, 일관된 평가 산출물을 데이터로 축적할 수 있다는 장점이 있음

- 전략계획서 점검 시에도 특정 키워드나 문장을 입력했을 때 관련된 사업의 성과지표를 빠르고 쉽게 추천해줄 수 있다는 점에서 업무 효율의 전반적인 제고를 기대할 수 있으며, 신규사업의 부처별 담당자가 성과지표를 설정하는 과정과 지표관리 과정에서 갖는 행정부담을 완화하는 등 관련 과업 효율성이 향상될 것으로 기대할 수 있음

□ 자연어처리 모델은 최근의 연구 등을 통해 R&D사업의 평가 부문에서도 반복되는 업무 프로세스에서의 비효율을 개선하고, 효과적인 정보를 제공해주는 등 그 가능성이 기대되는 반면, 데이터로 인한 오류와 데이터 등 모델 구축 비용으로 인한 우려가 존재함

- 비용 효율성 측면에서는 데이터의 비표준화, 제도개선으로 인한 장기간 일관된 데이터 축적의 어려움 등으로 인해 자연어 처리의 효용이 떨어지거나 데이터 구축에 모델로 인한 효과 이상의 비용이 소요되어 비용 효율성이 매우 떨어질 수 있다는 한계가 있음

- 또한 데이터가 부족하거나 품질이 낮거나, 분류 간 불균형이 심할 경우 모델의 성능이 안정적이지 못하고 현저히 저하될 수 있어 실제 프로세스에 활용하기 위한 서비스 도입 시 사용자들의 불편과 애로가 급격히 증가할 수 있다는 우려가 있음

- ㅇ 선행 연구에서 살펴보았듯이 기존의 파이썬을 활용한 다방면의 자연어 처리 모델이 기 구축되어 있지만, 데이터의 특성이나 업무의 특성을 고려하였을 때 국가R&D사업 평가를 위한 별도의 모델을 식별하고, 구축하고, 학습하여 평가 및 검증하는 과정을 지속적으로 거쳐야 한다는 어려움이 있음

  - 특히 딥러닝 모델은 복잡한 구조를 가지고 있어 해석이 어려운 블랙박스 구조로 작동됨. 사람이 모델의 학습 원리를 이해하기 어렵고 모델의 예측 결과값에 대한 근거를 정당화하는 데 한계가 존재한다는 점은 도입 이후 애로사항 보완에 어려움이 있을 수 있음

- ㅇ 또한 장기간 데이터가 축적되어 데이터셋이 대규모가 되거나 모델이 점점 더 복잡해질 경우 많은 연산량과 메모리가 요구됨에 따라 모델 훈련과 예측에 더 많은 시간과 비용이 소요되며, 제한된 컴퓨팅 자원을 가진 환경에서 활용이 어려울 수 있음

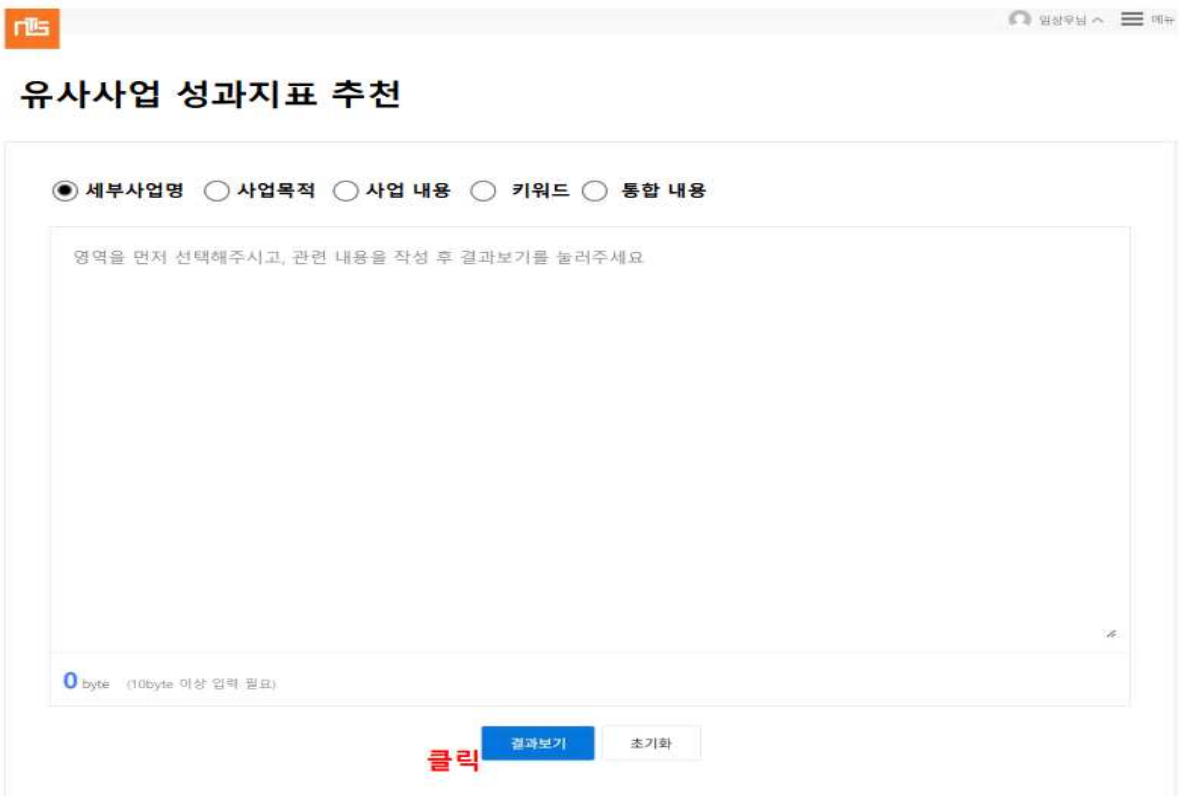
- ㅇ 또한, 국가연구개발사업 평가에는 과학기술적, 경제적, 사회적 영향 등 여러 기준을 포함하는데, 이러한 기준은 종종 주관적이며 정량화하기 어렵고, 신뢰성이나 수용성에 문제가 있을 수 있음

  - 융복합, 사회문제해결 등 사업의 다각화에 대응한 평가가 요구되는데 이러한 영역에서는 계량적 증거를 토대로 목표 달성을 확인하거나 성과지표를 효과적으로 추천해주는 데에 한계가 존재할 수 있음

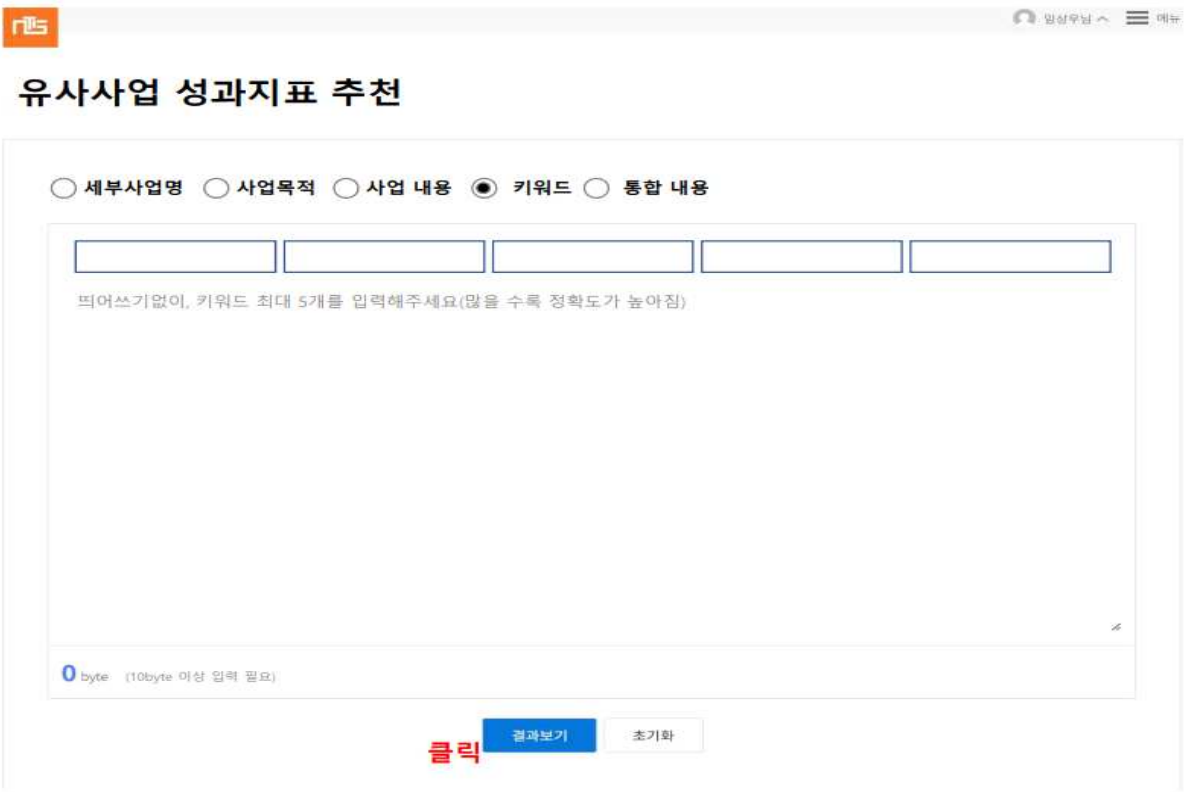
  - 평가에의 활용을 고려했을 때 결정의 신뢰성이나 전문성, 기술적 위험 및 수용성에 대한 대응 문제 등으로 인한 접근성에 제한이 있어 자연어처리 모델에 대한 신뢰도가 다방면, 여러 사용자로부터 충분히 검증되고 확인될 필요가 있음

[부록] AI 평가지원서비스 개념설계(안)

1. 성과지표 설정 지원



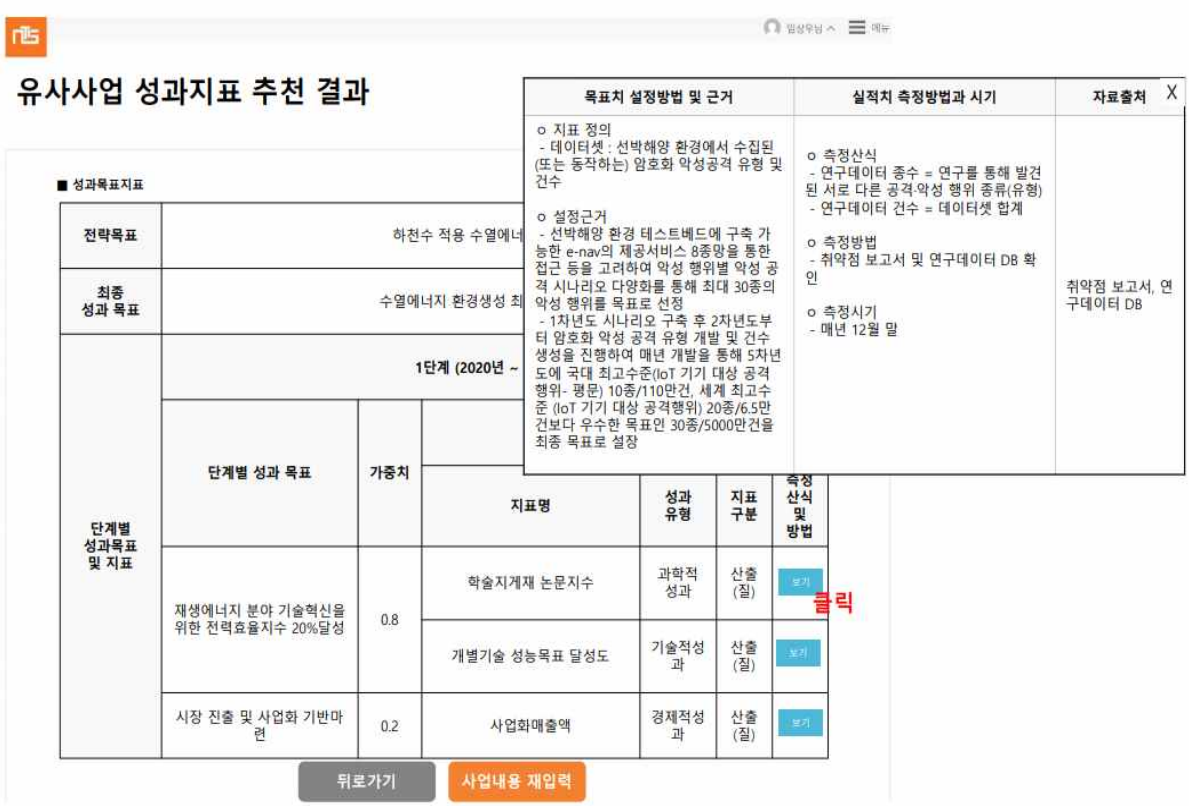
① 사업명/사업목적/사업내용/통합내용/키워드 등 검색옵션과 파라미터값, 유사도 기준치 설정



② KeyBERT 기반 관련사업 추천을 위한 키워드 입력

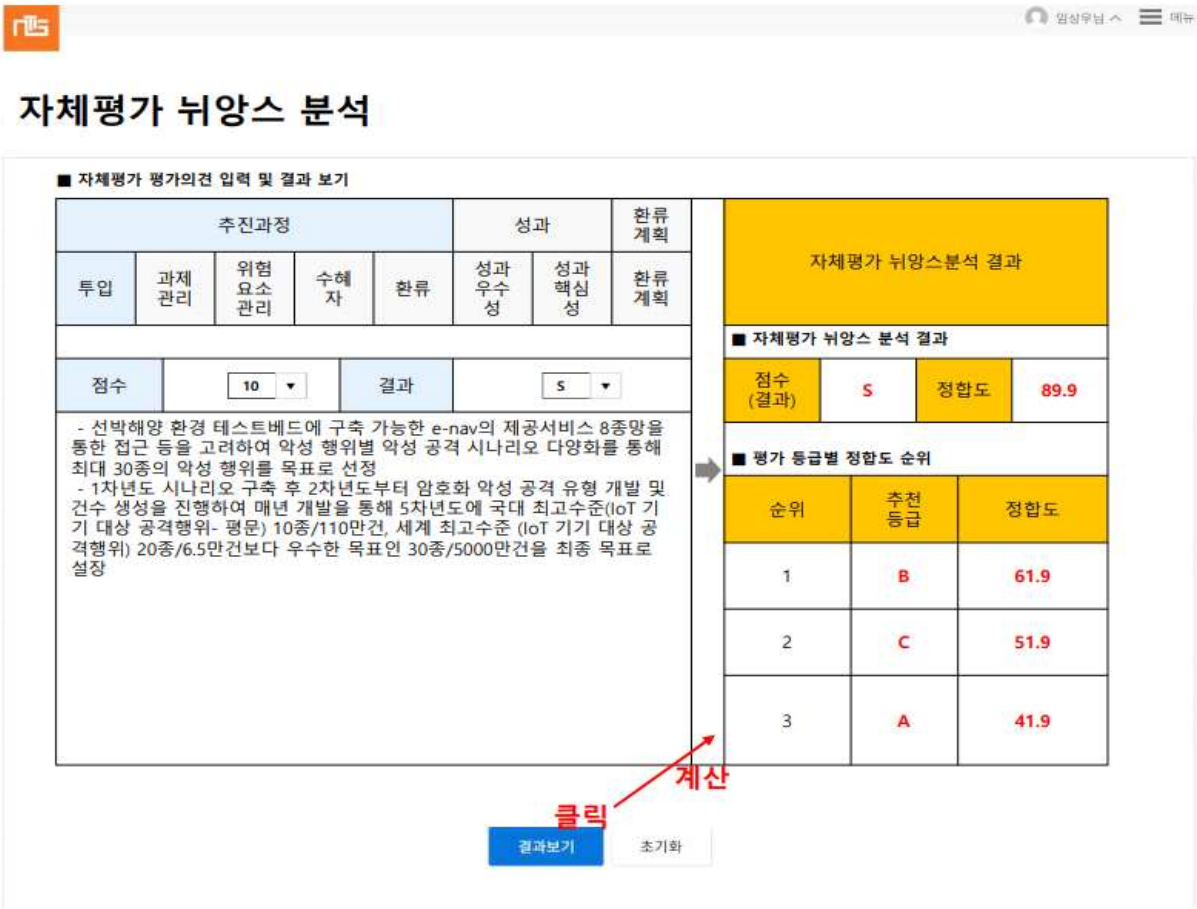
③ 관련사업 리스트 및 유사도 출력, 전략계획서 다운로드

④ 관련사업의 사업정보총괄표(PEIS) 링크



⑤ 성과지표 클릭 시 목표치 설정근거, 측정방법 및 산식, 자료출처 등 자료 팝업

2. 자체평가 등급설정 지원



(1안) 평가부문(추진과정/성과/환류계획) 및 지표(투입/과제관리/.../환류계획) 선택 후 평가의견 입력 시 등급 추천

(2안) 평가부문(추진과정/성과/환류계획) 및 지표(투입/과제관리/.../환류계획) 선택 후 평가의견, 자체평가점수, 예상등급 입력 시 등급별 정합도 제시

참 고 문 헌

[국문 참고문헌]

강윤희, & 박용범. (2004). SVM 을 이용한 디렉토리 기반 기술정보 문서 자동 분류시스템 설계. 전기전자학회논문지, 8(2), 186-194.

곽희종. (2023). “토픽모델링을 활용한 도시재생정책 이슈 분석.” 대한국토도시계획학회지 [국토계획 58.2: 22-37.

권민지. (2019). “토픽 모델링 기반 뉴스기사 분석을 통한 서울시 이슈 도출.” 한국방송미디어공학회 학술발표대회 논문집 (2019): 11-13.

권시현. “데싸노트의 실전에서 통하는 머신러닝”, 골든래빗 (2022)

김민호 and 한재필. (2022). “AI 기술, 지원정책의 효과를 높일 수 있을까?” KDI Policy Forum. No. 288.

김상일 외. (2022). “2022년 국가연구개발 성과평가 정책 수립 및 성과평가 실시”. 과학기술정보통신부·한국과학기술기획평가원: 65-66

김성진, & 안현철. (2016). 기업신용등급 예측을 위한 랜덤 포레스트의 응용. 산업혁신연구, 32(1), 187-211.

김판준. “랜덤포레스트를 이용한 국내 학술지 논문의 자동분류에 관한 연구.” 정보관리학회지 36.2 (2019): 57-77.

김현우, and 이승룡. “모바일 텍스트의 감성분류를 위한 SVM 기반 음운 커널 기법.” 정보과학회 논문지: 소프트웨어 및 응용 40.6 (2013): 350-355.

문화관광부. (1999). (21세기 세종계획)국어 기초자료 구축.

민진우, et al. “RoBERTa 를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존파싱.” 한국정보과학회 학술발표논문집 (2019): 407-409.

박은정, 조성준. (2014). KoNLPy: Korean natural language processing in Python. https://konlpy.org/ko/latest

박찬정, 성동수 & 이건배. (2010) SVM을 이용한 특허문서 분류기의 설계 및 구현. 산업기술종합연구소논문집., 38, 115-128.

박찬정, et al. “특허정보 문서를 이용한 자질선택 방법 및 분류 알고리즘의 성능비교 1.” 대한전자공학회 학술대회 (2011): 1034-1036.

서호준. (2019). “텍스트 네트워크 분석을 활용한 우리나라 과학기술정책 50 년의 주요 의제 분석-[과학기술 50 년사] 를 중심으로.” 과학기술정책 2.2 (2019): 171-201.

손정은, 고병철, and 남재열. “Random Forest 분류기와 Bag-of-Feature 특징 히스토그램을 이용한 의료영상 자동 분류 및 검색.” 정보처리학회논문지. 소프트웨어 및 데이터 공학 2.4 (2013): 273-280.

오현환 외. (2021). “2021년 국가연구개발 성과평가 정책 수립 및 성과평가 실시”. 과학기술정보통신부·한국과학기술기획평가원: 111-112

우경진, 정수현. (2019). 문장 유형에 따른 한글 형태소 분석기 비교. 한국정보과학회 학술발표논문집, 1,388-1,390.

유영호, 이용운. (2018). 은전한닢 프로젝트. https://eunjeon.blogspot.com/

유재호, 김하나, and 전의찬. (2021). “토픽모델링 기법을 활용한 녹색성장 정책 변화 분석.” 한국기후변화학회지 12.1 (2021): 67-75.

유호현. (2014). twitter-korean-text. Github. https://github.com/twitter/twitter-korean-text

윤영근. (2013). “정책증거, policy ecidence의 시차에 관한 연구: 산아제한정책사례의 적용” 행정논총 51(4)

윤태균, and 이관수. “의료진단 및 중요 검사 항목 결정 지원 시스템을 위한 랜덤 포레스트 알고리즘 적용.” 전기학회논문지 57.6 (2008): 1058-1062.

이경란, & 강창묵. (2023). 자연어처리 기계학습 기법을 이용한 공시문서의 자동분류: Confidential treatment 를 가진 8-K 문서를 중심으로. The Journal of Society for e-Business Studies, 28(2), 21-36.

이민철. (2022). Kiwi, Korean Intelligent Word Identifier. Github.

이유나, 박성미, and 박노섭. (2022). “개체명 인식과 이벤트 추출을 통한판결문 범죄사실

구성요소 및 스토리라인시각화방안 연구.” 한국정보처리학회 학술대회논문집 29.2 (2022): 490-492.

이재민, 하태현, 이민국, 박강희, 임대현, 권이남, . . . 전홍우. (2021). 과학기술 미래예측: 딥러닝 기반 특허기술 장기전략 예측. In: 한국과학기술정보연구원.

이종하, 구명완, 이경표. (2022). KoBERT 기반의 연구개발 직무 추천 모델 연구. 대한산업공학회 춘계공동학술대회 논문집, 개최지.

조한철, and 조근식. “나이브 베이지안 분류자와 메세지 규칙을 이용한 스팸메일 필터링 시스템.” 한국정보과학회 학술발표논문집 29.1B (2002): 223-225.

조희련, et al. “KoBERT, 나이브 베이즈, 로지스틱 회귀의한국어 쓰기 답안지 점수 구간 예측 성능 비교.” 한국정보처리학회 학술대회논문집 28.1 (2021): 501-504.

채정민, & 김효정. (2023). AI 기반 개인정보 침해평가 시스템: 정부입법 및 의원입법안에 대한 NLP 활용. 정보과학회 컴퓨팅의 실제 논문지, 29(12), 545-554.

최윤수, et al. “RoBERTa 를 이용한 한국어 기계독해.” 정보과학회 컴퓨팅의 실제 논문지 27.4 (2021): 198-203.

하지은, 신현철, & 이준기. (2017). RandomForest 와 XGBoost 를 활용한 한국어 텍스트 분류: 서울특별시 응답소 민원 데이터를 중심으로. 한국빅데이터학회지, 2(2), 95-104.

한채연, 김우식, and 윤동근. (2021). “토픽모델링과 네트워크 분석을 활용한 국・ 내외 재난 연구 동향 분석.” 2. 한국방재학회 논문집 21.5 (2021): 79-88.

황상흠, and 김도현. “한국어 기술문서 분석을 위한 BERT 기반의 분류모델.” 한국전자거래학회지 25.1 (2020): 203-214.

[영문 참고문헌]

Afridi, T. H., Alam, A., Khan, M. N., Khan, J., & Lee, Y.-K. (2021). A multimodal memes classification: A survey and open research issues. Paper presented at the Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications.

Ahn, H., and H. Y. Lee. “A combination model of multiple artificial intelligence techniques based on genetic algorithms for investment decision support aid: An application to KOSPI.” The e-Business Studies 10.1 (2009): 215-236.

Albadi, N., Kurdi, M., & Mishra, S. (2019). Hateful people or hateful bots? Detection and characterization of bots spreading religious hatred in Arabic social media. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-25.

Albaugh, Quinn, et al.. (2013). “The automated coding of policy agendas: A dictionary-based approach.” The 6th Annual Comparative Agendas Conference, Antwerp, Belgium.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. Paper presented at the 2017 international conference on engineering and technology (ICET).

Ali, Jehad, et al. “Random forests and decision trees.” International Journal of Computer Science Issues (IJCSI) 9.5 (2012): 272.

Angin, M., Taşdemir, B., Y ı lmaz, C. A., Demiralp, G., Atay, M., Angin, P., & Dikmener, G. (2022). A RoBERTa Approach for Automated Processing of Sustainability Reports. Sustainability, 14(23), 16139.

Ansolabehere, S., & Iyengar, S. (1995). Going negative: How attack ads shrink and polarize the electorate. (No Title).

Arunachalam, Ravi, and Sandipan Sarkar. (2013). “The new eye of government: citizen sentiment analysis in social media.” Proceedings of the IJCNLP 2013 workshop on natural language processing for social media (SocialNLP).

Azar, E. E. (1980). The conflict and peace data bank (COPDAB) project. Journal of Conflict Resolution, 24(1), 143-152.

Belgiu, Mariana, and Lucian Drăguţ. “Random forest in remote sensing: A review of applications and future directions.” ISPRS journal of photogrammetry and remote sensing 114 (2016): 24-31.

Berrar, Daniel. “Bayes’ theorem and naive Bayes classifier.” Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics 403 (2018): 412.

Blei, D., Ng, A., & Jordan, M. (2001). Latent dirichlet allocation. Advances in neural information processing systems, 14.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. (2003). "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

Bogdanowicz, A., & Guan, C. (2022). Dynamic topic modeling of Twitter data during the COVID-19 pandemic. Plos one, 17(5), e0268669.

Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.

Cabot, P.-L. H., Dankers, V., Abadi, D., Fischer, A., & Shutova, E. (2020). The pragmatics behind politics: Modelling metaphor, framing and emotion in political discourse. Paper presented at the Findings of the association for computational linguistics: emnlp 2020.

Calvo-González, Oscar, Axel Eizmendi, and Germán Reyes. (2018). "Winners never quit, quitters never grow: Using text mining to measure policy volatility and its link with long-term growth in latin America." World Bank Policy Research Working Paper 8310 (2018).

Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

Collingwood, Loren, and John Wilkerson. (2012). "Tradeoffs in accuracy and efficiency in supervised learning methods." Journal of Information Technology & Politics 9.3 (2012): 298-318.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics, 8, 439-453.

Dumais, S. T. (2004). Latent semantic analysis. Annual Review of Information Science and Technology (ARIST), 38, 189-230.

Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. Applied soft computing, 101, 107057.

Gennaro, G., & Ash, E. (2022). Emotion and reason in political language. The Economic Journal, 132(643), 1037-1059.

George, P., & Vinod, P. (2018). Composite email features for spam identification. In Cyber Security: Proceedings of CSI 2015 (pp. 281-289). Springer Singapore.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. Political Analysis, 18(1), 1-35.

Grimmer, Justin. (2013). "Appropriators not position takers: The distorting effects of electoral incentives on congressional representation." American Journal of Political Science 57.3 (2013): 624-642.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political analysis, 21(3), 267-297.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Ham, J., Choe, Y. J., Park, K., Choi, I., & Soh, H. (2020). Kornli and korsts: New benchmark datasets for korean natural language understanding. arXiv preprint arXiv:2004.03289

Hausladen, C. I., Schubert, M. H., & Ash, E. (2020). Text classification of ideological direction in judicial opinions. International Review of Law and Economics, 62, 105903.

Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their applications 13.4 (1998): 18-28.

Hiware, Kaustubh, et al. (2020). “NARMADA: Need and available resource managing assistant for disasters and adversities.” arXiv preprint arXiv:2005.13524 (2020).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Hogenboom, F., Frasincar, F., Kaymak, U., & De Jong, F. (2011). An Overview of Event Extraction from Text. DeRiVE@ ISWC, 48-57.

Honnibal, Matthew, et al. (2020). “spaCy: Industrial-strength natural language processing in python.”.

Huang, Y., & Luk, P. (2020). Measuring economic policy uncertainty in China. China Economic Review, 59, 101367.

Jang, H. (2019). A decision support framework for robust R&D budget allocation using machine learning and optimization. Decision Support Systems, 121, 1-12.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31(3), 685-695.

Jin, Z., & Mihalcea, R. (2022). Natural Language Processing for Policymaking. In Handbook of Computational Social Science for Policy (pp. 141-162). Cham: Springer International Publishing.

Jin, Zhijing, et al. (2021). “Mining the cause of political decision-making from social media: A case study of COVID-19 policies across the US states.” Findings of the Association for Computational Linguistics: EMNLP 2021.

Johnson, Kristen, and Dan Goldwasser. (2018). “Classification of moral foundations in microblog political discourse.” Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers).

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

Kim, D., Koo, J., & Kim, U. M. (2021, January). EnvBERT: multi-label text classification for imbalanced, noisy environmental news data. In 2021 15th International

Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1-8). IEEE.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. American political science review, 97(2), 311-331. research progress and challenges. AI Open 1 (2020): 22-39.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

Lee, Sangah, et al. "Kr-bert: A small-scale korean-specific language model." arXiv preprint arXiv:2008.03979 (2020).

Li, H., Yao, B., & Yan, X. (2021). Data-driven public R&D project performance evaluation: results from China. Sustainability, 13(13), 7147.

Li, Y., Li, J., Suhara, Y., Doan, A., & Tan, W.-C. (2020). Deep entity matching with pre-trained language models. arXiv preprint arXiv:2004.00584.

Liu, K., Chen, Y., Liu, J., Zuo, X., & Zhao, J. (2020). Extracting events and their relations from texts: A survey on recent research progress and challenges. AI Open, 1, 22-39.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

Lowe, W., Benoit, K., Mikhaylov, S., & Laver, M. (2011). Scaling policy preferences from coded political texts. Legislative studies quarterly, 36(1), 123-155.

Luo, Y., Card, D., & Jurafsky, D. (2020). Detecting stance in media on global warming. arXiv preprint arXiv:2010.15149.

Maarten Grootendorst. (2021). MaartenGr/KeyBERT: BibTeX (v0.1.3). Zenodo. https://doi.org/10.5281/zenodo.4461265

McClelland, C. A. (1976). World event/interaction survey codebook. In: ICPSR Ann Arbor.

Menini, S., Nanni, F., Ponzetto, S. P., & Tonelli, S. (2017). Topic-based agreement and disagreement in US electoral manifestos.

Merritt, R. L., Muncaster, R. G., & Zinnes, D. A. (1993). International event-data developments: DDIR phase II: University of Michigan Press.

Mitamura, T., Liu, Z., & Hovy, E. H. (2017). Events Detection, Coreference and Sequencing: What's next? Overview of the TAC KBP 2017 Event Track. Paper presented at the TAC.

Mitchell, T. M. (1997). Does machine learning really work? AI magazine, 18(3), 11-11.

Nanni, Federico, et al. (2022). "Political text scaling meets computational semantics." ACM/IMS Transactions on Data Science (TDS) 2.4 (2022): 1-27.

Nguyen, D.-H., Nghiem, N. V. D., Nguyen, B.-S., Le, D. T., Sabahi, S., Nguyen, M.-T., & Le, H. (2022). Make the most of prior data: A solution for interactive text summarization with preference feedback. arXiv preprint arXiv:2204.05512.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., . . . Ray, A. (2022). Training language models to follow instructions with human feedback, 2022. URL https://arxiv. org/abs/2203.02155, 13.

Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, M., & Costa da Silva, E. (2019). Local interpretable model-agnostic explanations for classification of lymph node metastases. Sensors, 19(13), 2969.

Pennebaker, James W., Martha E. Francis, and Roger J. Booth. (2001). "Linguistic inquiry and word count: LIWC 2001." Mahway: Lawrence Erlbaum Associates 71.2001 (2001): 2001.

Pisner, Derek A., and David M. Schnyer. "Support vector machine." Machine learning. Academic Press, 2020. 101-121.

Qi, Peng, et al. (2020). “Stanza: A Python natural language processing toolkit for many human languages.” arXiv preprint arXiv:2003.07082 (2020).

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2006). An automated method of topic-coding legislative speech over time with application to the 105th-108th US Senate. Paper presented at the Midwest political science association meeting.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. American Journal of Political Science, 54(1), 209-228.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Raleigh, C., Linke, r., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: An armed conflict location and event dataset. Journal of peace research, 47(5), 651-660.

Reimers, Nils and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” ArXiv abs/1908.10084 (2019)

Reimers, Nils and Iryna Gurevych. “Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation.” EMNLP (2020)

Rish, Irina. “An empirical study of the naive Bayes classifier.” IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.

Slapin, Jonathan B., and Sven-Oliver Proksch. (2008). “A scaling model for estimating time-series party positions from texts.” American Journal of Political Science 52.3 (2008): 705-722.

Sakib, A. S., Mukta, M. S. H., Huda, F. R., Islam, A. N., Islam, T., & Ali, M. E. (2021). Identifying insomnia from social media posts: psycholinguistic analyses of user tweets. Journal of medical Internet research, 23(12), e27613.

Schrodt, P. A. (2000). Pattern Recognition of International Crises Using. Political complexity: Nonlinear models of politics, 296.

Schumacher, G., Schoonvelde, M., Traber, D., Dahiya, T., & De Vries, E. (2016). EUSpeech: A new dataset of EU elite speeches.

Shaar, S., Martino, G. D. S., Babulkov, N., & Nakov, P. (2020). That is a known lie: Detecting previously fact-checked claims. arXiv preprint arXiv:2005.06058.

Sharma, P., & Li, Y. (2019). Self-supervised contextual keyword and keyphrase retrieval with self-labelling.

Shineware. (2019). KOMORAN 문서-사용자 사전 사용.

Shirodkar, M., Nimbalkar, S., Ingole, A., Mishra, P., & Sahu, S. (2020). Election result prediction using Sentiment Analysis. International Research Journal of Engineering and Technology, 7(2), 2855-2857.

Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time?series party positions from texts. American Journal of Political Science, 52(3), 705-722.

Son, T. S., et al. “A Study on the Covert Channel Detection in the TCP/IP Header based on the Support Vector Machine.” Journal of The Korea Institute of Information Security and Cryptology 14.1 (2004): 35-45.

Sundberg, R., & Melander, E. (2013). Introducing the UCDP georeferenced event dataset. Journal of peace research, 50(4), 523-532.

Thakur, N., Reimers, N., Daxenberger, J., & Gurevych, I. (2020). Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks.arXiv preprint arXiv:2010.08240.

Trappey, A. J., Liang, C.-P., & Lin, H.-J. (2022). Using machine learning language models to generate innovation knowledge graphs for patent mining. Applied Sciences, 12(19), 9818.

Trappey, A. J., Wei, A. Y., Chen, N. K., Li, K.-A., Hung, L., & Trappey, C. V. (2023). Patent landscape and key technology interaction roadmap using

graph convolutional network-Case of mobile communication technologies beyond 5G. Journal of Informetrics, 17(1), 101354.

Trappl, R. (2006). Programming for peace: computer-aided methods for international conflict resolution and prevention (Vol. 2): Springer Science & Business Media.

Vapnik, V. N. "Statistical learning theory J Wiley New York." (1998).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. Information Systems, 94, 101582.

Walker, C., Strassel, S., Medero, J., & Maeda, K. (2006). Ace 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia, 57, 45.

Wang, J., Fan, Y., Palacios, J., Chai, Y., Guetta-Jeanrenaud, N., Obradovich, N., . . . Zheng, S. (2022). Global evidence of expressed sentiment alterations during the COVID-19 pandemic. Nature Human Behaviour, 6(3), 349-358.

Xiang, W., & Wang, B. (2019). A survey of event extraction from text. IEEE Access, 7, 173111-173137.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. Paper presented at the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.

Yoo, H. S., Jung, Y. L., & Jun, S.-P. (2023). Prediction of SMEs' R&D performances by machine learning for project selection. Scientific Reports, 13(1), 7598.

Yoon, S., & Kim, N. (2023). Document Classification Methodology Using Autoencoder-based Keywords Embedding. Journal of the Korea Society of Computer and Information, 28(9), 35-46.

Yun, Tae-Gyun, and Gwan-Su Yi. "Application of random forest algorithm for the

decision support system of medical diagnosis with the selection of significant clinical test." The transactions of The Korean Institute of Electrical Engineers 57.6 (2008): 1058-1062.

[웹페이지]

자연어 처리(NLP), SAS 인사이트, "https://www.sas.com/ko_kr/insights/analytics/what-is-natural-language-processing-nlp.html" (최종접속일: 2024.3.3.)

Stanza– A Python NLP Package for Many Human Languages, Stanze, "https://stanfordnlp.github.io/stanza/" (최종접속일: 2024.3.3.)

https://github.com/SKTBrain/KoBERT (최종접속일: 2024.3.3.)

https://hleecaster.com/ml-random-forest-concept/ (최종접속일: 2024.3.3.)

https://sktelecom.github.io/project/kobert/ (최종접속일: 2024.3.3.)

https://www.spotfire.com/glossary/what-is-a-random-forest (최종접속일: 2024.3.3.)

https://sooftware.io/roberta/ (최종접속일: 2024.3.3.)

https://www.ajunews.com/view/20201011091342159 (최종접속일: 2024.3.3.)

https://github.com/SKTBrain/KoBert (최종접속일: 2024.3.3.)

https://github.com/bab2min/Kiwi (최종접속일: 2024.3.3.)

https://zdnet.co.kr/view/?no=20220725093548 (최종접속일: 2024.3.3.)

https://bit.ly/42ofdQg (최종접속일: 2024.3.3.)

https://heytech.tistory.com/149 (최종접속일: 2024.3.3.)

https://bit.ly/3SoJolV (최종접속일: 2024.3.3.)

https://komorandocs.readthedocs.io/ko/latest/manual/manual.html (최종접속일: 2024.3.3.)