Research Report

RE-092

국가연구개발사업의 SW 융합 현황 분석에 관한 연구

A Study on SW Convergence Analysis of National Research and Development Project

서영희

2020. 01.

SPRi 소프트웨어정책연구소

이 보고서는 2019년도 과학기술정보통신부 정보통신·방송연구개발사업의 연구결과로서 보고서 내용은 연구자의 견해이며, 과학기술정보통신부의 공식입장과 다를 수 있습니다.

연 구 기 관 : 소프트웨어정책연구소

과제책임자 : 서영희 선임연구원

목 차

표 목 차

그 림 목 차

요 약 문

1. 제 목 : 국가연구개발사업의 SW 융합 현황 분석에 관한 연구

2. 연구 목적 및 필요성

(1) 추진 배경

ㅇ 제4차 산업혁명의 핵심 동인은 소프트웨어(이하 SW)로써 전 산업의 생산성과 글로벌 경쟁력을 제고하고 사회 문제를 해결하여 혁신 성장을 주도하고 있음

ㅇ 글로벌 시장은 이미 SW 기업을 중심으로 재편중이며 SW가 핵심 역량으로 부각하고 있기 때문에 지능화 혁신의 주체로써 인공지능, 빅데이터, 클라우드 등 SW 신기술의 고도화 및 전 산업의 디지털 전환을 위해 기존의 SW 산업을 위한 R&D 정책에서 벗어나 현황에 맞는 정부 차원의 R&D 정책에 대한 전략 수립이 필요함

ㅇ SW 중요성의 제고로 부처별로 국가연구개발사업을 통해 SW 융합 과제를 진행하고 있으며 정성적 방법론을 활용한 SW 융합 R&D 조사가 이루어졌으나 방대한 시간과 비용이 투입되며 지속적인 분석에 어려움이 존재함

→ SW 융합 역량 제고를 위한 SW 융합 현상에 대한 부처별, 분야별 등의 유형별 통계 자료가 필요하나 현재까지는 정량적 방법을 활용하여 자동으로 SW 융합 R&D 단계를 분류하는 방법이 제시되고 있지 못한다는 것이 본 연구의 출발점임

(2) 연구의 목적

ㅇ 본 연구는 국가연구개발사업의 과제 중 SW 융합 R&D 과제를 자동으로 분류하기 위한 모델을 개발하여 SW 융합 R&D 현황을 조사하고 분석하여 SW 융합 역량 확보를 위한 SW 융합 R&D 정책 방향을 제시하는 것임

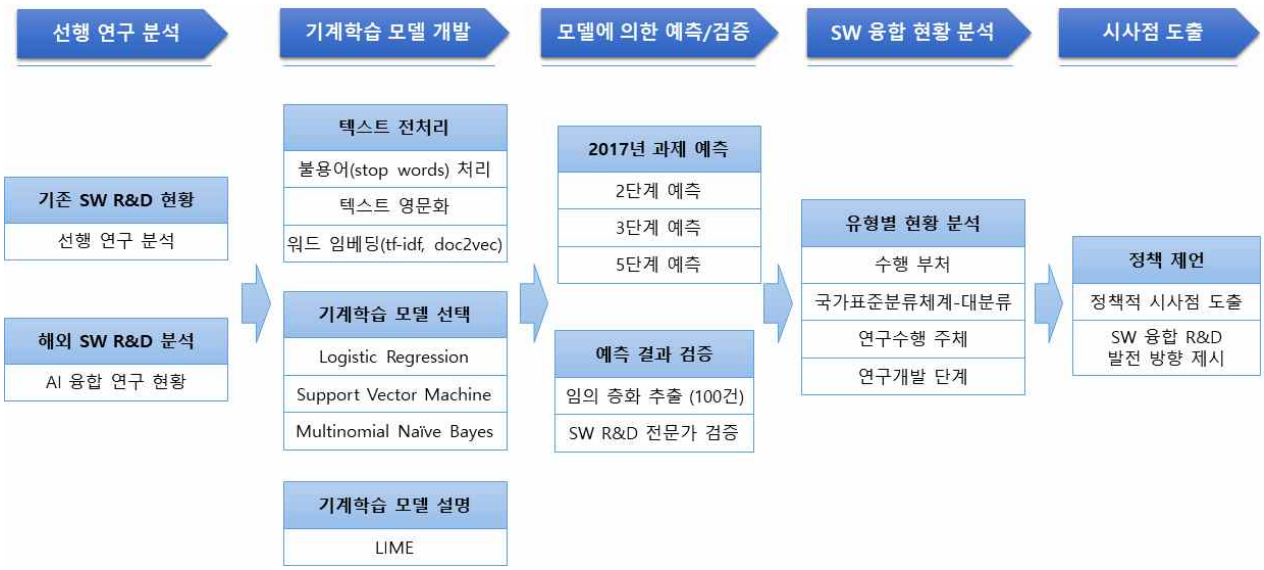
- 2018년에 수립된 판단 기준에 따라 전문가 델파이 조사의 결과 값을 입력 데이터로 활용하여 기계학습을 활용한 자동 분류 모델을 개발하고 2017년 연구개발과제에 대한 유형별 SW 융합 R&D 관련한 기초 자료를 확보하고자 함

3. 연구의 구성 및 방법

(1) 연구의 구성

ㅇ 본 연구의 구성 및 진행 방법은 아래와 같이 총 5단계로 구성됨

[그림 1] 연구 진행 단계



① (1단계) 정량적으로 SW 융합 R&D 현황을 분석하기 위해 기존에 수행된 연구에 대한 정리 및 SW 융합 R&D 기준에 대해 알아보고, 국외 SW 융합 R&D 정책에 대한 현황을 정리함

② (2단계) SW 융합 R&D의 자동 분류를 위한 기계학습 모델을 개발하기 위해 학습 데이터인 국가연구개발과제의 연구 요약문 관련 텍스트에 대한 전처리 과정을 거쳐 세 가지의 기계학습 모델에 대한 실험을 통해 모델을 선택함

③ (3단계) 선택된 분류 모델을 활용하여 2017년 국가연구개발과제를 대상으로 총 3가지(2/3/5단계) 형태로 SW 융합 R&D에 대한 예측 및 예측 결과에 대한 신뢰도 검증 수행

④ (4단계) 2017년 국가연구개발과제의 예측 결과를 토대로 수행 부처, 국가표준분류체계(대분류), 수행 주체, 연구개발 단계라는 총 4가지의 유형별 SW 융합 R&D 현황 분석을 위해 시각화 도구를 활용하여 분석을 수행함

⑤ (5단계) 3단계의 SW 융합 R&D 현황을 토대로 향후 SW 융합 활성화를 위한 정책적 시사점 및 발전 방향을 도출함

(2) 연구의 방법

ㅇ SW 융합 R&D 과제에 대한 자동 분류 모델 개발을 위해 선행 연구 분석 및 다양한 분류 모델에 대한 실험 및 검증을 거쳐 SW 융합 R&D 전문가 자문 회의 등을 통해 SW 융합 R&D 현황 분석 및 시사점을 도출함

① 텍스트 마이닝 방법

- 중요 단어의 추출 시에 대표적으로 활용되며 텍스트 내에 특정 단어의 상대적 중요도를 계산하는 통계적인 수치인 tf-idf 방법과 문서를 벡터로 표현하는 방법론인 doc2vec을 활용하여 각 결과 값을 비교・분석함

② 기계학습 모델 선택

- 로지스틱 회귀(Logistic Regression)와 Support Vector Machine(SVM), 다항 나이브 베이즈(Multinomial Naive Bayes)를 비교・분석하여 최상의 모델을 선택함

기계학습 모델	설명
로지스틱 회귀	종속 변수가 0이나 1인 이진형 변수에서 자주 사용하는 분류 방식
Support Vector Machine(SVM)	입력 공간과 관련된 비선형 문제를 고차원 공간의 선형문제로 사상(projection)시켜 나타냄
다항 나이브 베이즈	신뢰도가 높고 효율적인 문서 분류기로 널리 사용되는 기분 문서 분류 방법

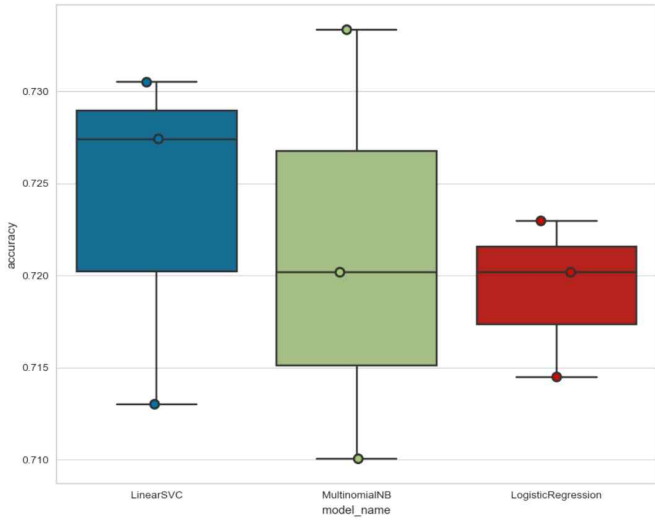
4. 연구 내용 및 결과

(1) SW 융합 R&D 분류 모델 예측 결과 및 단계 소개

○ SW 융합 R&D 단계의 구분을 위해 2가지의 텍스트 마이닝 기법과 3가지의 모델을 실험을 수행하였으며, 모델의 성능은 정확도(Accuracy)와 정밀도(Precision), 재현율(Recall)을 활용하여 측정하고 결과를 비교함

- 정확도는 입력 단계를 그대로 정확히 맞추어야 정답으로 인정하는 **정매칭**과 중간 3개 단계를 결합하여 하나의 단계로 간주하는 **결합 매칭**, 실제 결과와 1단계 떨어진 예측 결과를 보이면 맞은 것으로 간주하는 **인접 매칭**으로 나누어 실험을 진행함

[그림 2] 모델별 정확도 실험 결과



(2) 텍스트 마이닝 결과

○ 본 연구에서 활용한 텍스트 형태의 입력 자료에 대한 벡터화 결과는 <표 1>과 같이 doc2vec 보다 tf-idf가 평균적으로 예측력이 높은 것으로 나타남

- doc2vec 알고리즘은 일반적으로 학습 데이터가 많은 경우에 보다 더 좋은 성능을 보이기 때문에 학습 데이터가 상대적으로 적은 현 실험에서는 tf-idf가 상대적으로 더 바람직한 알고리즘으로 판단됨

- doc2vec은 실험을 수행하는 과정에서 입력 단어에 대한 추론(inference)

으로 인해 입력 데이터 전체를 벡터화하기 위해 요구되는 시간이 tf-idf에 비해 현저하게 많이 필요하다는 단점이 존재함

<표 2> tf-idf/doc2vec의 학습 결과 비교(5단계 & Linear SVM 기준)

구분	단계 (Scale)	정밀도 (Precision)	재현율 (Recall)	f1-score
tf-idf	High	0.59	0.76	0.67
	Med High	0.17	0.06	0.09
	Medium	0.18	0.07	0.09
	Med Low	0.22	0.06	0.09
	Low	0.83	0.95	0.89
	정확도 (Accuracy)	Accurate matches: 72.78% Combined matches: 75.31% Adjacent matches: 84.50%		
doc2vec	High	0.57	0.75	0.64
	Med High	0.21	0.09	0.12
	Medium	0.10	0.05	0.07
	Med Low	0.07	0.01	0.03
	Low	0.82	0.94	0.88
	정확도 (Accuracy)	Accurate matches: 70.11% Combined matches: 73.19% Adjacent matches: 82.57%		

(3) 2017년 국가연구개발과제의 예측 결과

1) Linear SVM 모델의 벡터화 알고리즘 및 단계별 예측 결과 비교

○ Linear SVM 모델과 tf-idf/doc2vec 알고리즘에 대한 실험 결과, tf-idf 방식의 2단계 예측 값이 2016년 학습 데이터와 가장 유사한 결과를 도출[1]하였음

- tf-idf방식의 2단계 예측 값으로 살펴본 결과, SW 융합 R&D가 차지하는 비중은 전체의 16.8%인 것으로 판단

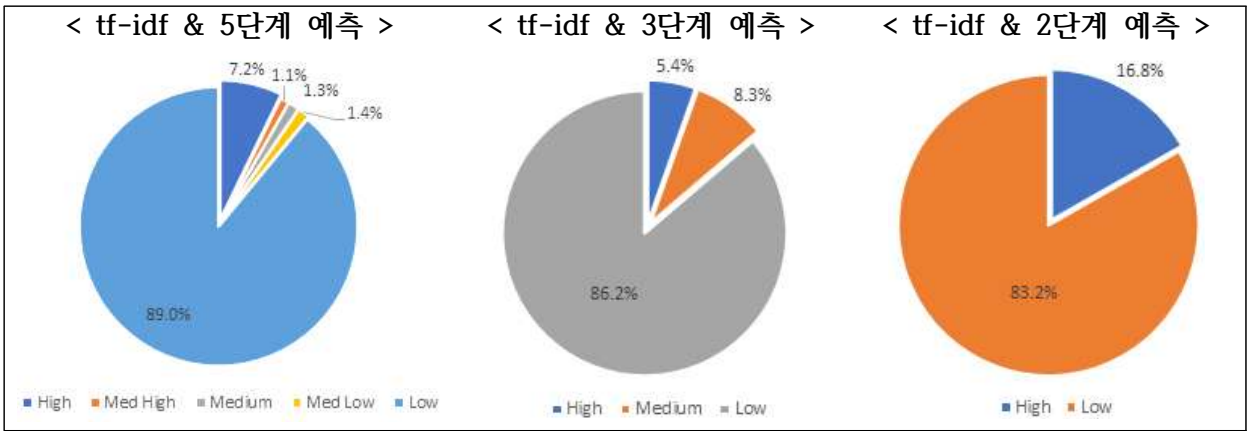
○ doc2vec은 각 단계별 치우침 현상이 상대적으로 적게 발생하여 5단계와 2단계 사이의 Low 단계의 비율 차이가 2.75%인 비교적 안정적인 모델로 나타남(tf-idf는 5.76% 차이를 보임)

1) 2018년의 선행 연구 결과, 과제수 기준으로 15.9%, 총 연구비 기준으로 16.3%가 SW 융합 R&D 과제인 것으로 조사됨(국가연구개발사업에서의 SW융합에 관한 연구, SPRi, 2018)

<표 3> 2017년도 국가연구개발과제 분류 결과 요약 (Linear SVM)

	5단계			3단계			2단계		
tf-idf	High	4,012	7.22%	High	3,022	5.44%	High	9,325	16.78%
	Med High	587	1.06%	Medium	4,623	8.32%	Low	46,241	83.22%
	Medium	746	1.34%	Low	47,920	86.24%			
	Med Low	776	1.40%						
	Low	49,445	88.98%						
doc2vec	High	3,499	6.30%	High	3,005	5.41%	High	8,107	14.28%
	Med High	668	1.20%	Medium	4,463	8.03%	Low	48,678	85.72%
	Medium	1,070	1.96%	Low	48,098	86.56%			
	Med Low	1,313	2.36%						
	Low	49,017	88.21%						

[그림 3] Linear SVM & tf-idf 방식의 단계별 예측 분포



2) Multinomial NB 모델의 벡터화 알고리즘 및 단계별 예측 결과 비교

○ 해당 모델은 <표 3>과 같이 5단계 예측에서 tf-idf와 doc2vec 모두 중간의 세 단계(Med High, Medium, Med Low)의 예측을 거의 하지 못하며, 특히 Med High 단계의 과제를 거의 분류하지 못한 것으로 나타났음

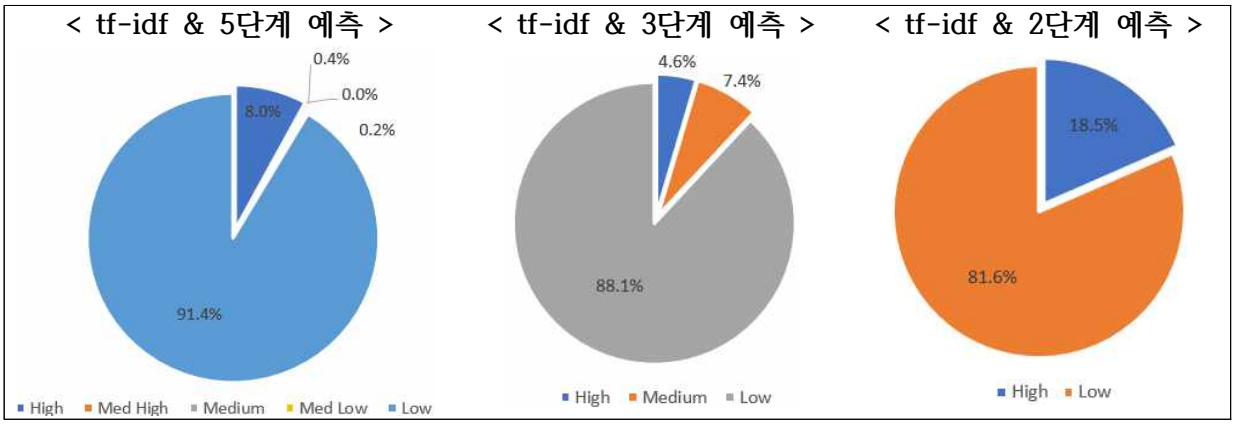
○ 두 알고리즘에서 모두 5단계와 3단계 구분에서 Low 단계로 구분한 과제가 약 90% 초반과 약 88% 정도의 수치를 보이고 있어 2016년 학습 데이터에 비해 Low 단계로의 예측에 대한 쏠림 현상이 존재함

- tf-idf의 결과에서 2단계와 5단계의 Low 단계에 대한 비율 차이는 9.87%이고, doc2vec은 10.75%의 차이를 보임

<표 4> 2017년도 국가연구개발과제 분류 결과 요약(Multinomial NB)

	5단계			3단계			2단계		
tf-idf	High	4,513	7.95%	High	2,590	4.56%	High	10,477	18.45%
	Med High	**9**	**0.02%**	Medium	4,185	7.37%	Low	46,308	81.55%
	Medium	**241**	**0.42%**	Low	50,010	88.07%			
	Med Low	**109**	**0.19%**						
	Low	51,913	91.42%						
doc2vec	High	3,736	6.58%	High	2,534	4.46%	High	10,129	17.84%
	Med High	0	**0.00%**	Medium	3,976	7.00%	Low	46,656	82.16%
	Medium	89	**0.16%**	Low	50,275	88.54%			
	Med Low	200	**0.35%**						
	Low	52,760	92.91%						

[그림 4] Multinomial NB & tf-idf 방식의 단계별 예측 분포



(4) 예측에 대한 전문가 검증 결과

○ 100건의 표본 추출 후, SW 융합 R&D 전문가의 검증을 수행하여 모델의 예측 결과와 비교한 결과, 종합적으로 자동 분류 모델이 예측한 결과는 SW R&D 전문가가 분류한 결과와 86%가 일치하는 것을 확인함

- High 단계는 두 결과가 75%, Med High와 Medium 단계는 16.7%가 일치하며, 이는 학습 데이터가 가지는 특성 때문에 Med High, Medium과 Med Low 단계는 전문가와 분류 모델이 일치하는 결과가 적다고 판단할 수 있음

○ 본 연구의 분류 모델이 중간(Med*) 단계에 해당하는 과제를 상대적으로 과소하게 예측하는 경향을 보이며 High와 Low 단계의 구분에 더 특화된 것을 알 수 있음

<표 5> 전문가 예측 기준 정 매칭, 인접 매칭, 비 매칭 수와 비중

전문가 예측	개수	정 매칭		인접 매칭		비 매칭	
		개수	비중	개수	비중	개수	비중
High	8	6	75.0%	0	0.0%	2	25.0%
Med High	6	1	16.7%	1	16.7%	4	66.7%
Medium	6	1	16.7%	0	0.0%	5	83.3%
Med Low	2	0	0.0%	2	100.0%	0	0.0%
Low	78	78	100.0%	0	0.0%	0	0.0%
합계	100	**86**	**86.0%**	3	3.0%	11	11.0%

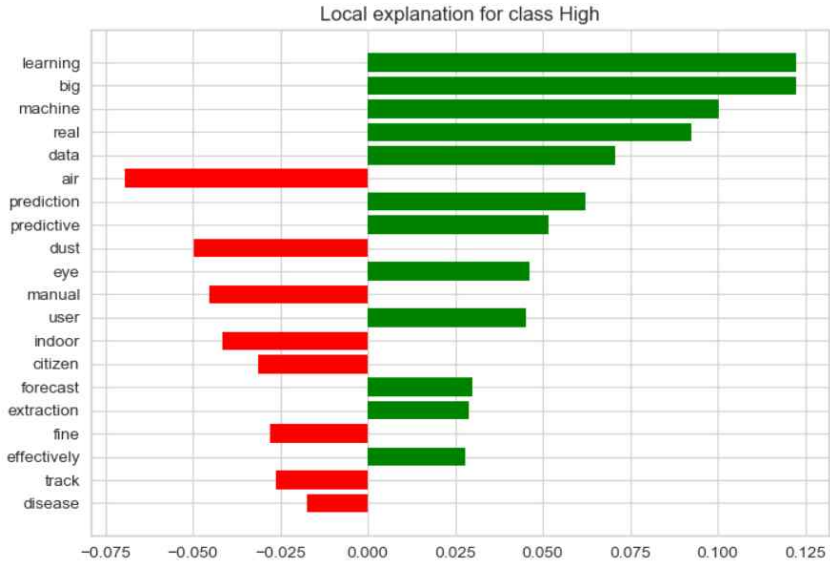
(5) 모델 예측에 대한 설명

ㅇ 분류 모델의 활용도 및 신뢰도를 높이기 위해 분류의 근거가 되는 설명을 제공하고자 하며, Lime 알고리즘을 활용하여 과제의 분류 근거를 설명함

- Lime은 입력 데이터를 변경시켜 가면서 예측이 어떻게 변하는지를 확인함으로써 특정 예에 대한 분류 모델의 결정에 대한 설명을 시도하는 블랙박스에 대한 설명자임

- [그림 3]은 High 단계로 분류된 데이터에 대한 분류 근거를 설명한 것으로 “learning”, “big” 등이 긍정적 영향을 미치고, “air”, “dust” 등이 부정적 영향을 미쳤으나 긍정의 기여도가 높아 High 단계가 결정됨

[그림 5] High 단계 판단에 대한 각 단어의 가중치 그래프

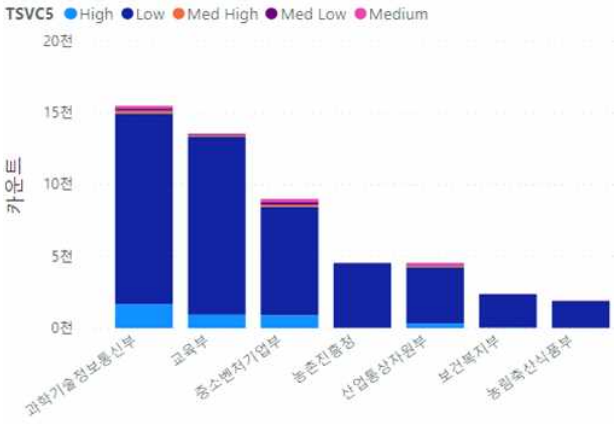


(6) SW 융합 R&D의 유형별 분석 및 시사점

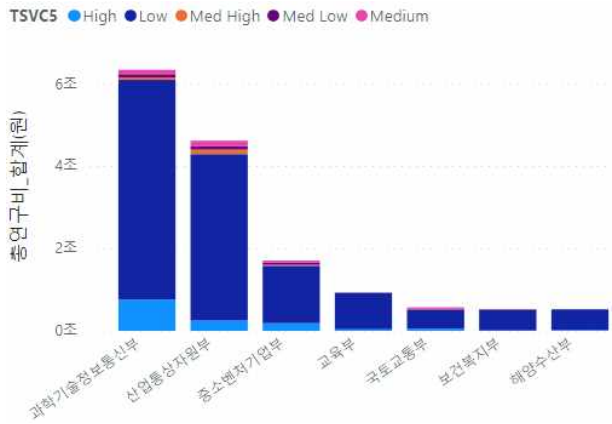
○ 자동 분류기를 활용하여 2017년 국가연구개발과제 중 예측 결과가 존재하는 총 55,566건의 예측 결과를 총 4가지(수행 부처, 수행 주체, 연구개발 단계, 과학기술표준분류체계)의 유형별로 현황을 분석하고자 함

① 수행 부처별

[그림 6] 부처별 5단계 구분 결과(과제 수 기준, 상위 7개)



[그림 7] 부처별 5단계 구분 결과(총 연구비 기준, 상위 7개)



○ **(현황)** 5단계 구분 및 총 연구비 기준으로 SW 융합 R&D를 가장 많이 수행한 주체는 과학기술정보통신부(이하 과기정통부), 산업통상자원부(이하 산업부), 중소벤처기업부(이하 중기부)로 나타났으며 과제 수 기준으로는 과기정통부, 교육부, 중기부 순으로 조사됨

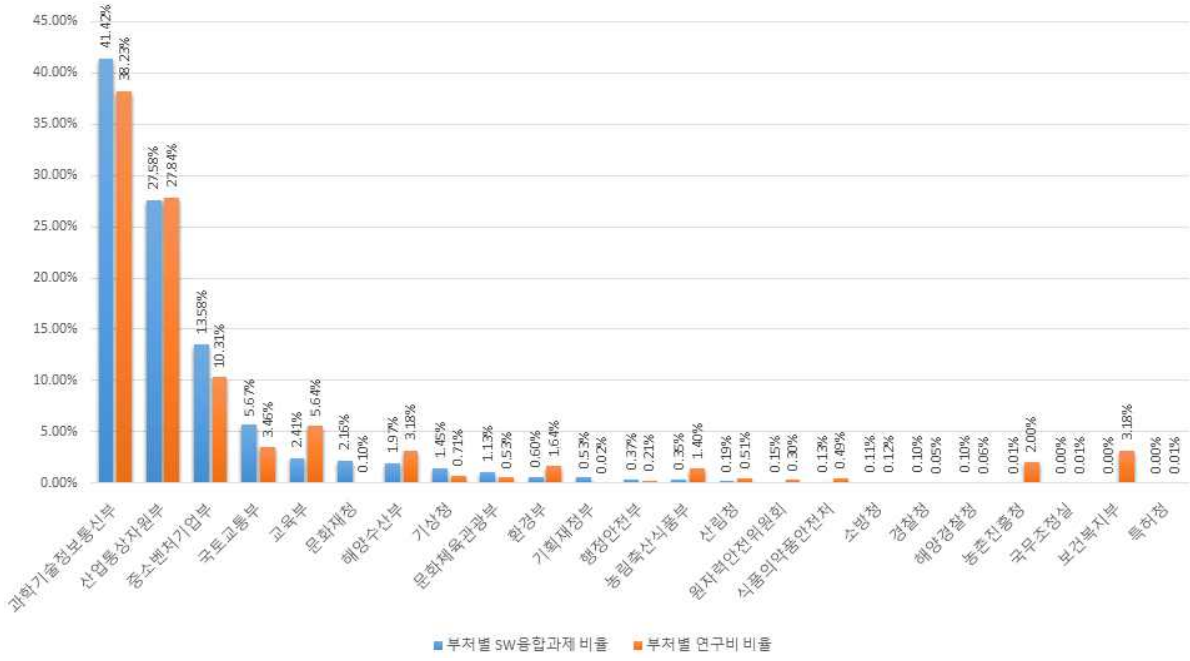
- (과기정통부) 과기정통부가 타 부처에 비해 SW 융합 R&D 활동을 80% 이상 포함하는 단계인 High 단계의 과제를 가장 많이 활발하게 수행하는 주체로 나타남

  . High 단계에서 과제 수 기준으로 교육부(2번째)에 비해 약 1.8배 정도 많은 과제를 진행하였으며, 총 연구비 측면에서는 산업부(2번째)에 비해 약 3배 정도 큰 금액을 차지하고 있음

- (중기부) 과제 수 기준으로 SW 융합 R&D 활동이 20~60% 정도(Medium, Med Low 단계)인 과제를 가장 많이 수행하는 부처로 나타났음

- (산업부) 총 연구비 기준으로 SW 융합 R&D 활동이 40~80% 사이(Med High, Medium 단계)의 과제를 가장 활발히 수행하는 부처로 조사되었음

[그림 8] 부처별 연구비 및 SW 융합 과제 현황



○ **(시사점)** SW 융합 R&D 과제는 국가연구개발사업*에 비해 상대적으로 부처별로 균형 있게 수행되지 않아 SW 융합 활성화를 위한 전략 마련이 필요

- 총 연구비 기준으로 과기정통부, 산업부, 중기부가 차지하는 비중의 합이 82.58%이나 그 외 20개 부처는 합계가 10% 미만으로 매우 저조한 상황

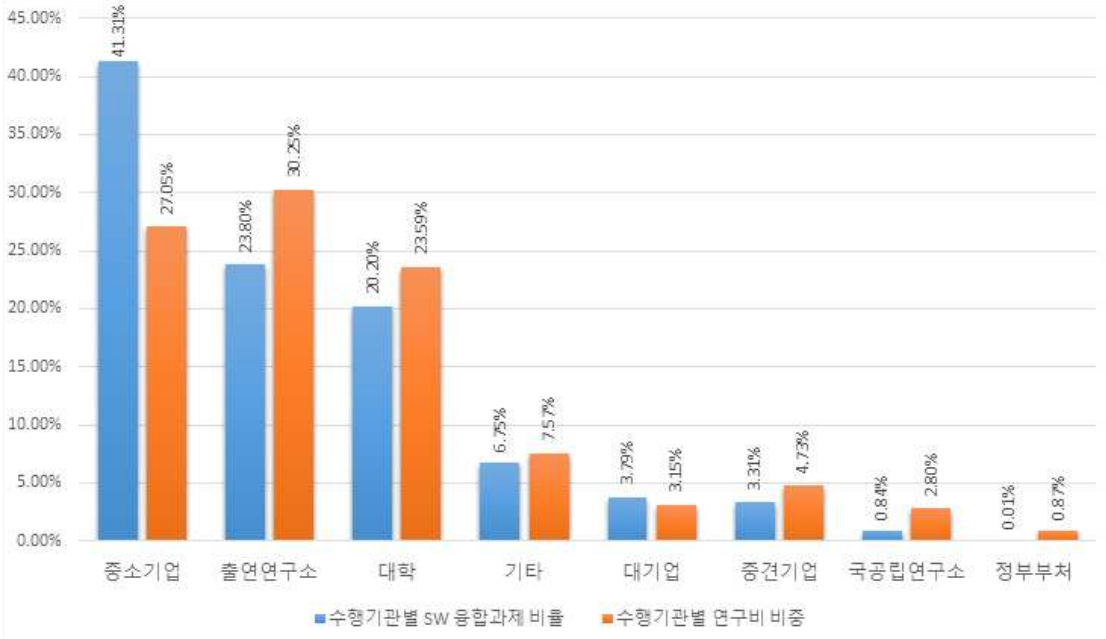
* 국가연구개발사업 전체의 총 연구비에서 위 3개 부처가 차지하는 비중은 76.38%임

- 국가연구개발사업에서 SW 융합이 부처별로 균형 있게 이루어지지 않고 있다는 점에서 SW 융합 제고를 위한 부처별 SW 융합 R&D 기획* 및 범부처 차원의 SW 융합 과제의 연계 기획 및 지원 등의 정책 수립이 요구됨

* 보건복지부, 농림축산식품부 등의 각종 통계 처리 및 현황 분석이나 기상청의 기상 예측 모델 등에 SW를 융합한 R&D의 시범 적용 및 확대 등

② 수행 주체별

[그림 9] 수행 주체별 연구비 및 SW 융합 과제 현황

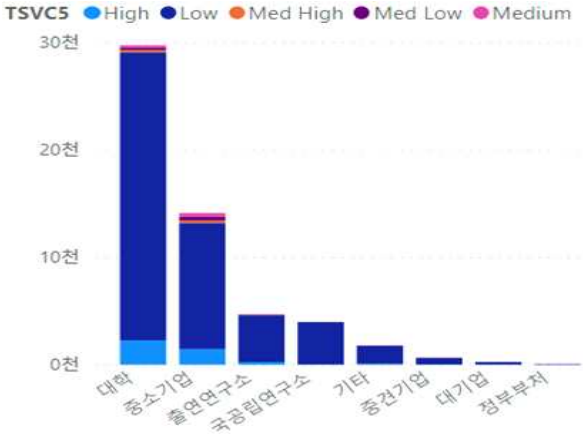


○ **(현황)** SW 융합 과제를 주로 수행하는 주체는 중소기업, 출연연구소, 대학 순으로 나타나고 있으나 대기업, 중견기업, 국공립연구소 등은 SW 융합 과제에서 차지하는 비중이 5% 미만으로 매우 저조한 것으로 나타남

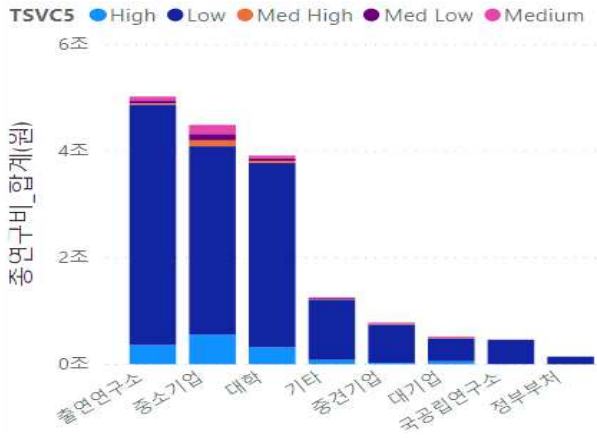
- 중소기업은 총 연구비 기준에서 Low 단계를 제외한 모든 SW 융합 R&D 단계에서 과제를 가장 많이 수행하고 있는 주체로 나타났으며, 대학은 과제 수 기준으로는 High와 Med High 단계에서 SW 융합 R&D 과제를 가장 많이 수행하는 주체로 조사됨

- 출연연구소는 총 연구비 기준으로 전체 국가연구개발과제 중 30% 정도를 수행하고 있으나 이에 비해 SW 융합 R&D 과제를 적극적으로 수행하는 주체가 아닌 것으로 나타남

[그림 10] 연구수행 주체-5가지 단계(과제 수)



[그림 11] 연구수행 주체-5가지 단계(연구비)



○ **(시사점)** 국가연구개발과제에 대한 총 연구비 비중은 출연연구소, 대학, 중소기업 순이나 SW 융합 과제는 이의 역순으로 조사되어 출연연구소의 SW 융합 R&D 활성화를 위한 노력이 필요함

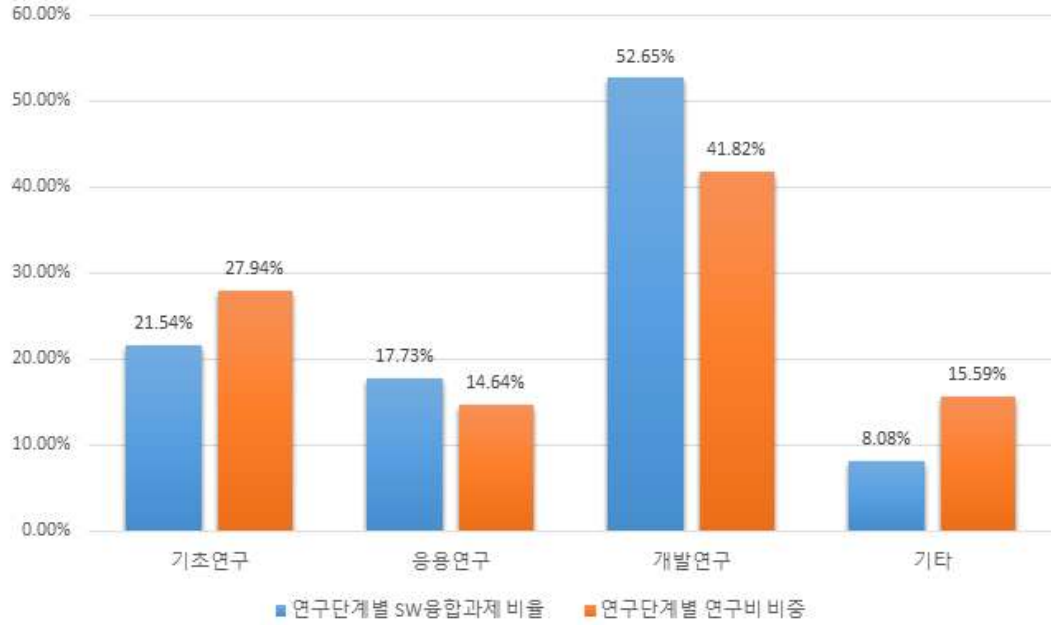
- 국가연구개발과제의 30% 이상의 연구비를 사용하는 출연연구소의 SW 융합 R&D를 통한 경쟁력 확보 및 수행한 연구 성과가 SW 융합 과제의 핵심 주체인 중소기업을 통해 사업화로 연결되기 위한 제도적 지원이 요구됨

③ 연구개발 단계

○ **(현황)** 국가연구개발과제의 총 연구비에 대한 단계별 비중은 개발, 기초, 응용연구 순으로 높게 나타났으며, SW 융합 R&D도 동일한 것으로 나타남

- SW 융합 R&D는 개발연구가 국가연구개발에 비해 다소 높은 비중인 52% 이상을 차지하고 있으며, 이는 SW R&D의 특성과 일치하는 현상인 것으로 판단됨

[그림 12] 연구단계별 연구비 및 SW 융합 과제 현황



○ **(시사점)** SW R&D는 기초 단계에 해당하는 새로운 사실의 발견보다는 제품·서비스의 창출을 목표로 하는 특성을 가지고 있기 때문에 개발 단계의 비중이 일반 연구개발보다 높은 것은 이를 반영한 현상인 것으로 판단됨

④ 국가과학기술표준분류체계의 대분류

○ **(현황)** 정보/통신과 전기/전자와 같은 일부 분야는 SW 융합 비율이 높으나 생명과학, 농업 관련 분야는 SW 융합 정도가 미흡한 것으로 조사되어 균형있는 SW 융합을 위한 정책 수립이 필요함

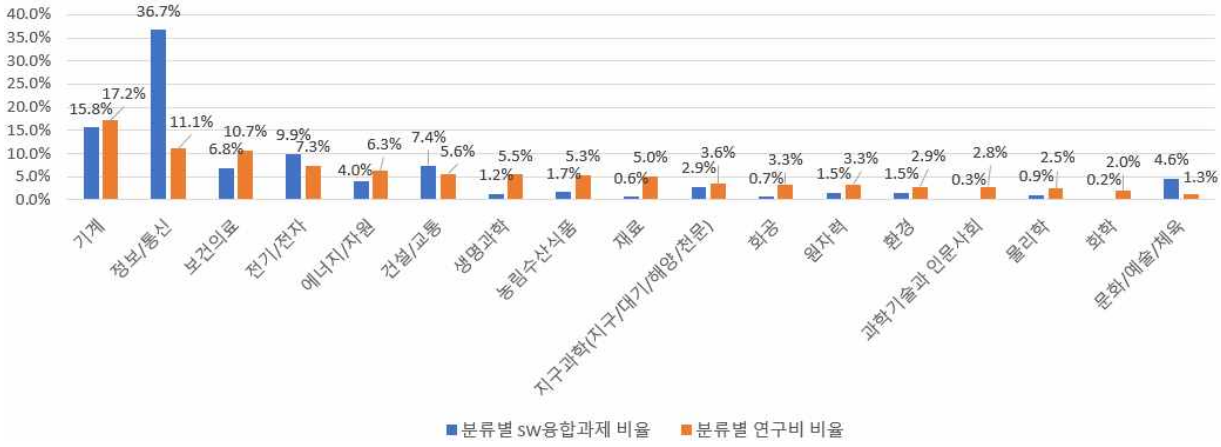
- 정보/통신 분야는 국가연구개발사업에서 차지하는 비중(11.1%)에 비해 SW 융합 R&D 과제 전체의 연구비에서 차지하는 비중이 3배 이상(36.7%) 높게 나타났으며 국가연구개발사업과 가장 큰 차이를 보임

- 전기/전자와 건설/교통, 문화/예술/체육 분야 역시 국가연구개발사업에서 차지하는 분야별 비율에 비해 SW 융합 R&D 비율이 높은 것으로 나타나 SW 융합이 높은 것으로 조사됨

- 생명과학과 농림수산식품, 재료, 화공 등의 분야는 국가연구개발의 총 연구

비에서 차지하는 분야의 비중에 비해 SW 융합 정도가 현저히 떨어지는 것으로 나타남

[그림 13] 과학기술표준분류체계별 연구비 및 SW 융합 과제 현황



○ **(시사점)** 분야별 SW 융합 정도의 차이가 존재하여 분야별 글로벌 경쟁력 강화를 위한 전략적 포트폴리오의 수립이 요구됨

- 융합이 저조한 생명과학과 농림수산식품, 재료, 화공 등의 분야에서 SW 융합을 통한 효율성 제고 및 경쟁력 확보를 위해 SW 융합 R&D 과제의 기획 및 관련 성과에 대한 홍보 등의 전략 마련이 필요함

5. 정책적 활용 내용

○ 본 연구의 결과물인 SW 융합 R&D 자동 분류 모델을 통해 국가 연구개발 사업 전체에 대한 SW 융합 현황에 대한 분석이 가능하며 기존의 전문가를 활용한 분석 방법에 비해 투입되는 자원과 시간을 절약하고 매년 데이터에 기반한 추이 분석을 할 수 있음

- 이를 통해 구체적인 SW 융합 R&D 현황 및 근거 자료를 확보하여 향후 중장기 SW 융합 R&D 전략 및 정책 수립 시 기초 자료로 활용이 가능함

- 본 연구에서는 기존의 정성적 방법이 아닌 정량적인 방법으로 SW 융합 R&D 판단 기준을 수립하고, 2017년 NTIS에서 제공한 과제 정보를 기반

으로 조사를 수행하여 현황 및 현안을 도출하고, 향후 SW 융합 R&D 활성화에 대한 정책적 개선 방향을 제시하였음

- 본 연구에서 수행한 국외 SW 관련 R&D 정책 분석과 국내 국가연구개발과제의 SW 융합 R&D 현황 분석을 통하여 SW 기술이 핵심 기반이 되고 있는 제4차 산업혁명 시대에 맞춰 전반적인 국가연구개발과제의 SW 융합 정도를 적절한 수준으로 향상하기 위한 정책 수립의 근거를 제시하였음

6. 기대 효과

○ 국가연구개발사업 전반에서 일어나고 있는 SW 융합 현상을 지속적으로 파악하고 추이를 비교·분석할 수 있는 SW 융합 R&D 분석의 발판을 마련함

- 인공지능 기술을 활용하여 기존의 많은 인력과 비용, 시간을 투입하여 표본 추출을 통해 전체의 과제를 추정하는 기존 연구의 제약점을 극복하여 매년 국가연구개발사업에 대한 연도별 SW 융합 확산 현상의 비교·분석을 빠르고 정확하게 수행할 수 있음

- 또한, NSF나 DARPA, EU의 Horizon 2020과 같은 국외 연구기관에서 수행되는 국가연구개발사업 관련 텍스트 데이터를 확보하여 자동 분류 모델을 적용할 수 있으며, 해당 결과를 국내 현황과 비교하는 추후 연구를 통해 시사점을 도출할 수 있음

SUMMARY

1. Title: Analysis of SW Convergence Status of National Research & Development Projects

2. Purpose and Necessity of the Research

(1) Background of research

ㅇ The key driver of the fourth industrial revolution is software(SW), which enhances the productivity and global competitiveness of all industries and leads innovation growth by solving social problems.

ㅇ Global markets are already reorganizing around SW businesses. As SW is emerging as a key capability, it is necessary to establish a strategy for government-level R&D policies that are tailored to the current situation, moving away from the existing R&D policies for SW industries to develop and upgrade SW technologies such as artificial intelligence, big data and cloud as the main players for intelligent innovation.

ㅇ Because importance of SW is increasing, each ministry is carrying out task of convergence of SW through national R&D projects. In addition, although SW convergence R&D survey was conducted using qualitative methodology, it takes a lot of time and money and there are difficulties in continuous analysis

→ Statistical data for SW convergence phenomena are required by type, such as departmental and sector, to enhance SW convergence capabilities, but methods for automatically classifying SW convergence R&D steps have not been presented so far using quantitative methods. This study began to overcome this point

(2) Purpose of research

ㅇ This research is to develop a model for automatically classifying SW convergence R&D tasks among national R&D projects to investigate and analyze SW convergence R&D status to present SW convergence R&D policy direction for securing SW convergence capabilities.

- Based on the criteria established in 2018, the results of the expert Delphi survey will be used as input data to develop an automatic classification model using machine learning and to secure basic data related to SW convergence R&D by type for the 2017 R&D project.

3. Composition and method of research

(1) Composition of research

The method of composition and progress of this study consists of a total of five stages.

① (Stage 1) To analyze the current status of SW convergence R&D in quantity, the existing research performed is organized and the SW convergence R&D criteria are explored, and the status of SW convergence R&D policies are summarized.

② (Stage 2) To develop a machine learning model for automatic classification of SW convergence R&D, the model is selected through experimentation on three machine learning models after preprocessing of the text related to the research summary of the national R&D task, which is learning data.

③ (Stage 3) Using the selected classification model, the national R&D project is carried out in 2/3/5 stages to verify the reliability of the prediction and results for SW convergence R&D.

④ (Stage 4) Based on the prediction results of the 2017 national R&D project, the analysis is carried out using visualization tools to analyze the current status of SW convergence R&D by type of department, national standard classification system, performance principal, and stage of execution.

⑤ (Stage 5) Based on the SW convergence R&D status of the three types of steps, policy implications and direction of development are derived for the activation of SW convergence in the future.

(2) Method of research

○ To develop an automatic classification model for SW convergence R&D tasks, SW convergence R&D status analysis and implications are derived through SW convergence R&D expert consultation meetings, after prior research analysis and experimentation and verification of various classification models.

① Text mining method

- Each result value is compared and analyzed using tf-idf method, which is a representative value used to extract important words, and a statistical value that calculates the relative importance of a specific word in the text, and doc2vec, which is a methodology for expressing a document as a vector.

② Select machine learning model

- A total of three models are compared. Support Vector Machine (SVM), which refers to logistic regression, a type of classification frequently used in binary variables with a dependent variable of zero or one, and nonlinear problems associated with input space as linear problems in high-dimensional space, is the method of mood classification widely used as reliable and efficient document classification. Compare and analyze these models to finally select the best model.

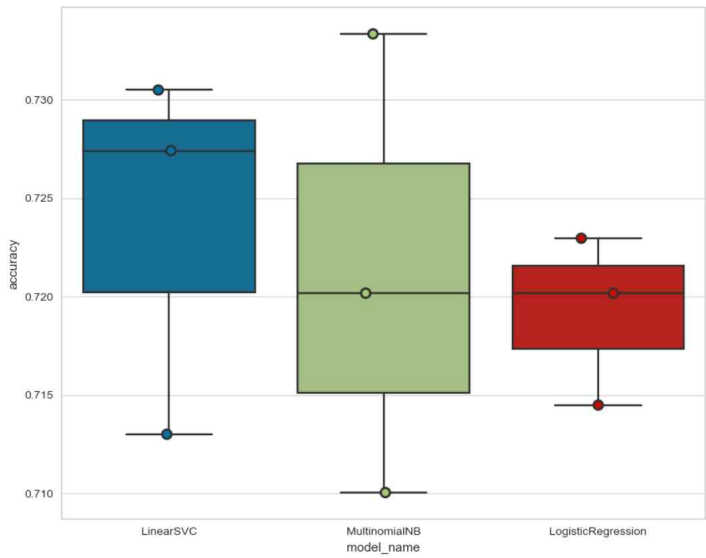
4. Main Contents and Results

(1) Introduction of SW Convergence R&D Classification Model Prediction Results and Steps

○ Two text mining techniques and three models were experimented to distinguish SW fusion R & D stages. The performance of the model was measured using Accuracy, Precision, and Recall and the results were compared.

- Accuracy is conducted by combining the correct and correct matching and the middle three steps together to consider one step, and by adjacent matching that is considered correct if actual results and predictions that fall by one step are shown.

[Picture 1] Model Specific Accuracy Experiment Results



(2) Text mining results

○ The results of vectoring for text-type inputs used in this study show that tf-idf is on average more predictable than doc2vec, as shown in Table 1.

- Because doc2vec algorithms generally perform better in cases where learning data is high, tf-idf is considered to be a relatively preferable algorithm in current experiments where learning data are relatively small.

- The disadvantage of doc2vec is that in the course of the experiment, the time required to vector the input data as a whole is significantly greater than tf-idf.

<Table 1> Comparison of tf-idf/doc2vec (5 Stage & Linear SVM)

Algorithm	Scale	Precision	Recall	f1-score
tf-idf	High	0.59	0.76	0.67
	Med High	0.17	0.06	0.09
	Medium	0.18	0.07	0.09
	Med Low	0.22	0.06	0.09
	Low	0.83	0.95	0.89
	Accuracy	Accurate matches: 72.78% Combined matches: 75.31% Adjacent matches: 84.50%		
doc2vec	High	0.57	0.75	0.64
	Med High	0.21	0.09	0.12
	Medium	0.10	0.05	0.07
	Med Low	0.07	0.01	0.03
	Low	0.82	0.94	0.88
	Accuracy	Accurate matches: 70.11% Combined matches: 73.19% Adjacent matches: 82.57%		

(3) 2017 R&D Project Prediction Results

1) Comparing the results of the Linear SVM Model

○ Experimental results for the linear SVM model and the tf-idf / doc2vec algorithm showed that the two-stage predicted value of the tf-idf method was most similar to the training data.

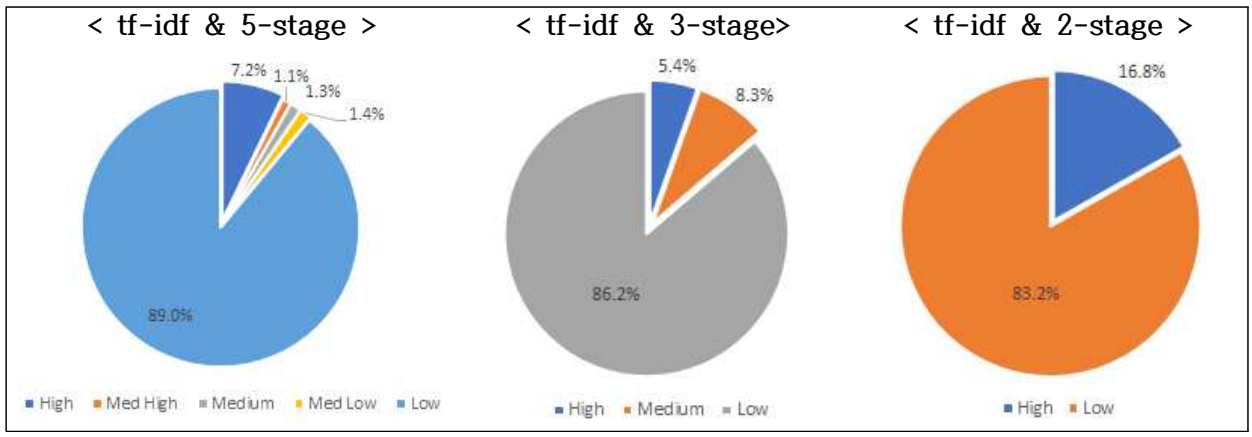
○ The doc2vec method produces a relatively stable model with a relatively

small skew of each stage, with a 2.75% difference in the ratio of the low stage between stages 5 and 2. (tf-idf shows a 5.76 percent difference)

<Table 2> Classification result summary of 2017 National R&D projects(Linear SVM)

	Stage 5			Stage 3			Stage 2		
tf-idf	High	4,012	7.22%	High	3,022	5.44%	High	9,325	16.78%
	Med High	587	1.06%	Medium	4,623	8.32%	Low	46,241	83.22%
	Medium	746	1.34%	Low	47,920	86.24%			
	Med Low	776	1.40%						
	Low	49,445	88.98%						
doc2vec	High	3,499	6.30%	High	3,005	5.41%	High	8,107	14.28%
	Med High	668	1.20%	Medium	4,463	8.03%	Low	48,678	85.72%
	Medium	1,070	1.96%	Low	48,098	86.56%			
	Med Low	1,313	2.36%						
	Low	49,017	88.21%						

[Picture 2] Linear SVM & tf-idf's Stage estimating distribution



2) Comparing the results of the Multinomial NB Model

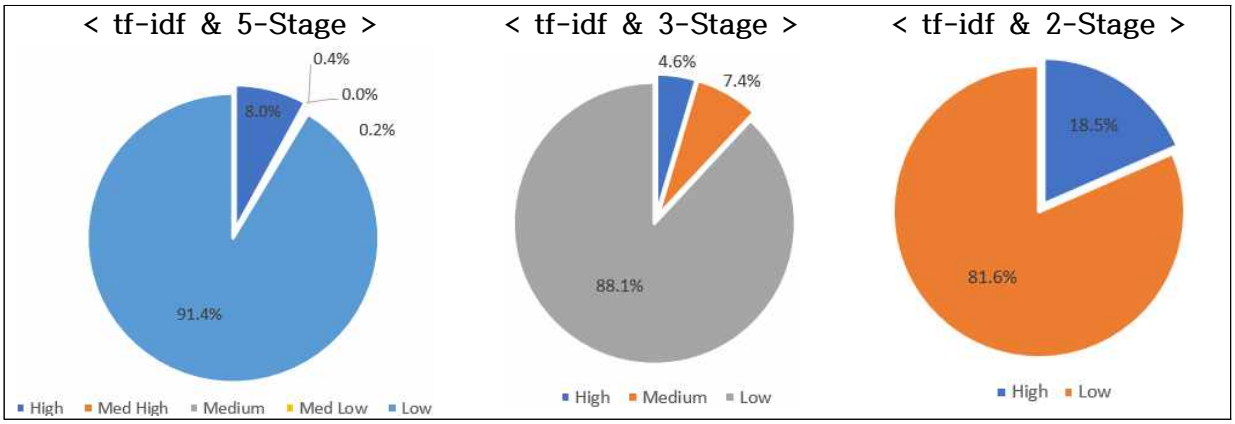
○ As shown in Table 3, tf-idf and doc2vec rarely predicted the middle three stages(Med High, Medium, Med Low), and in particular, the Med High stages task was hardly classified.

○ In both algorithms, the tasks classified into the low stages in the 5th stage and the 3rd stage show about 90% of the early stages and about 88% of the tasks.

<Table 3> Classification result summary of 2017 National R&D projects(Multinomial NB)

	Stage 5			Stage 3			Stage 2		
tf-idf	High	4,513	7.95%	High	2,590	4.56%	High	10,477	18.45%
	Med High	**9**	**0.02%**	Medium	4,185	7.37%	Low	46,308	81.55%
	Medium	**241**	**0.42%**	Low	50,010	88.07%			
	Med Low	**109**	**0.19%**						
	Low	51,913	91.42%						
doc2vec	High	3,736	6.58%	High	2,534	4.46%	High	10,129	17.84%
	Med High	0	**0.00%**	Medium	3,976	7.00%	Low	46,656	82.16%
	Medium	89	**0.16%**	Low	50,275	88.54%			
	Med Low	200	**0.35%**						
	Low	52,760	92.91%						

[Picture 3] Multinomial NB & tf-idf's Stage estimating distribution



(4) Results of expert verification of predictions

○ In comparison to the model's forecast results by performing 100 samples and validating them by SW convergence R&D experts, 86% of the results that SW R&D experts categorize collectively and the results from the automatic classification models

- The high level matches the two results by 75% and the Med High and Medium levels by 16.7%. Because of the nature of the learning data, the Med High, Medium, and Medium Low levels can be determined that experts and classification models do not match.

○ The classification model in this study tends to under-predict the tasks in

the middle (med*) phase and is more specialized in the distinction between high and low levels.

<Table 4> Accurate/Adjacent Match, Non-Matching' s Number and Percentage

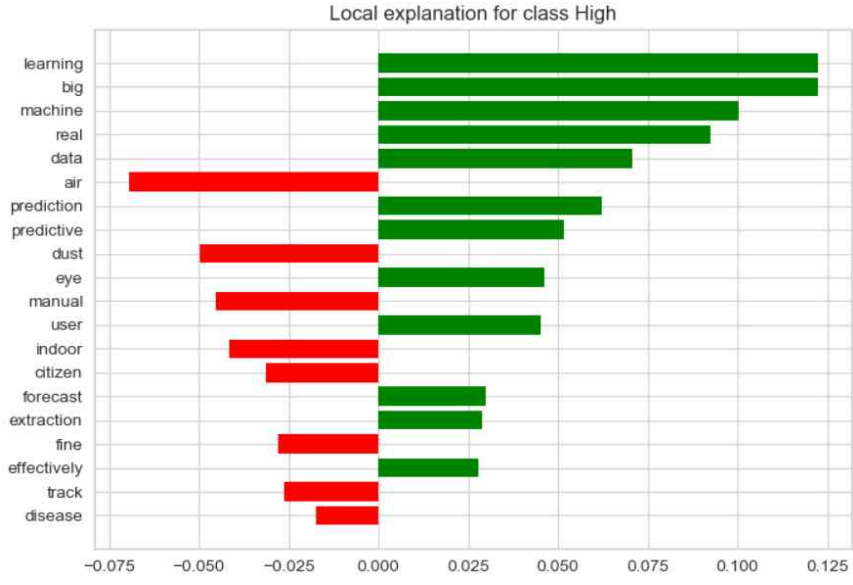
Expert Prediction	Count	Accurate Matching		Adjacent Matching		Non-Matching	
		Count	Percent	Count	Percent	Count	Percent
High	8	6	75.0%	0	0.0%	2	25.0%
Med High	6	1	16.7%	1	16.7%	4	66.7%
Medium	6	1	16.7%	0	0.0%	5	83.3%
Med Low	2	0	0.0%	2	100.0%	0	0.0%
Low	78	78	100.0%	0	0.0%	0	0.0%
합계	100	86	86.0%	3	3.0%	11	11.0%

(5) Description of model prediction

○ This study seeks to provide a basis explanation for classification in order to improve utilization and reliability of classification models, and uses the Lime algorithm to explain the basis for classification of tasks.

- [Figure 2] describes the basis for classification of data classified in the high phase, which is determined at the high level because "learning" and "big" have a positive effect and "air" and "dest" have a negative effect.

[Figure 4] Weighted graph of each word for high step judgment



(6) Analysis and Implication of SW Convergence R&D by Type

○ A total of 55,566 prediction results from the 2017 national R&D project will be analyzed for each of the four types.

① Performing ministries

○ (Status) The most frequently conducted SW convergence R&D based on the five-stage classification and total research costs were the Ministry of Science, Technology, Information and Communication (the Ministry of Science and Technology), the Ministry of Trade, Industry and Energy (the Ministry of Industry and Energy), and the Ministry of Small and Medium-sized Venture Enterprises (the division). And based on the number of assignments, they were surveyed in the order of the Ministry of Science and Technology, the Ministry of Education, and the Mid-term Department.

- The Ministry of Science and Technology is the most active person in performing the high level tasks, which include more than 80% of SW convergence R&D activities, compared to other ministries

. At the high level, the ministry conducted about 1.8 times more tasks than the Education Ministry (2th) based on the number of tasks, and in terms of total research costs, it accounted for about three times larger than the Ministry of Industry and Energy (2th).

- The Small and Medium Business Venture Department is the department that carries out the most tasks with 20 to 60% SW convergence R&D activities based on the number of tasks.

- The Ministry of Trade, Industry and Energy (MOTIE) has found that SW convergence R&D activities are most actively carrying out tasks between 40 and 80% (Med High and Medium levels) based on total research costs.

○ **(Implication)** Since SW convergence is relatively not carried out in a balanced manner compared to national R&D projects*, strategies for revitalizing SW convergence are needed.

- The combined portion of the Ministry of Science, ICT and Medium Business accounts for 82.58 percent of the total research costs, while 20 other ministries and agencies are very under 10 percent.

* The above three ministries account for 76.38 percent of the total research costs for the entire national R&D project.

- Given that SW convergence has not been balanced by different ministries in national R&D projects, policies such as planning for SW R&D by each ministry to enhance SW convergence and planning and support for SW convergence projects at the pan-ministerial level are required*.

** Various statistics such as the Ministry of Health and Welfare, the Ministry of Agriculture, Food and Rural Affairs, and the Korea Meteorological Administration's weather forecast models include the pilot application and expansion of R&D incorporating SW.

② By Project performer

○ **(Status)** The main subjects of SW convergence are small and medium-sized businesses, government-funded research institute, and universities. However, large companies, mid-sized companies and government-funded research institute showed that their portion of SW convergence projects was very low at less than 5 percent.

- Small and medium-sized enterprises are the ones that perform the most tasks in all SW convergence R&D phases, except for the low level, based on the basis of total research costs. And universities were found to be the ones who performed the SW convergence R&D tasks most frequently in the high and medium high stages by task count

standards.

- The government-funded research institute conducted about 30 percent of all national R&D tasks based on total research costs, but it was found that it was not the main body that actively carried out SW convergence R&D tasks.

○ **(Implication)** The proportion of total research costs on national R&D tasks is investigated in reverse order by the government-funded research institute, universities, and small and medium enterprises, and SW convergence tasks in order to promote SW convergence of the participating institutes.

- Efforts should be made to revitalize SW convergence at government-funded research institute that use more than 30 percent of research funds for national R&D projects. In addition, institutional support should be strengthened to link the research achievements of the government-funded research institute to commercialization through small and medium-sized enterprises, a key player in the SW convergence project.

③ Research and development stage

○ **(Status)** The proportion of the total research cost of the national R&D project was high in the order of development, basic and applied research, and SW convergence R&D was the same.

- SW Convergence R&D is believed to be a phenomenon consistent with SW R&D characteristics, with development research accounting for 52% or more, which is somewhat higher than national R&D.

○ **(Implication)** It seems that SW R&D has characteristics that aim to create products and services rather than to discover new facts that fall into the basic stage.

④ National S&T(Science and Technology) Standard Classification System

○ (Status) While some areas such as information/communications and electricity/electronics have high SW convergence rates, the degree of SW convergence in life sciences and agriculture is found to be insufficient, which necessitates the establishment of policies for balanced convergence.

- The information/communication sector is the biggest difference, with 36.7 percent, or more than three times higher than the 11.1 percent share in national R&D projects, in total research costs for SW convergence R&D projects.

- Electricity/electronics, construction/ transportation, culture/art/sports are also found to have a higher ratio of SW convergence R&D compared to the ratio of each field in national R&D projects.

- The fields of life science and agriculture, fisheries, materials and chemical engineering show a significant drop in SW convergence compared to the total research costs of national research and development.

○ (Implication) There is a difference in level of SW convergence between different fields, which calls for the establishment of strategic portfolios to strengthen global competitiveness in each sector

- It is necessary to establish strategies for promoting SW convergence by planning tasks through SW convergence R&D and promoting related achievements in life science, agriculture, fisheries, materials and chemical engineering.

5. Policy use

○ The SW convergence R&D automatic classification model, which is the

result of this study, enables analysis of SW convergence status for the entire national R&D project. In addition, data-based trending analysis can be performed annually, saving resources and time spent compared to traditional methods of analysis using experts.

- Through this process, specific SW convergence R&D status and evidence data can be secured and used as basic data for the establishment of mid- to long-term SW convergence R&D strategies and policies in the future.

- In this study, the criteria for SW convergence R&D determination were established in a quantitative manner rather than a conventional qualitative method, and the research was conducted based on task information provided by NTIS in 2017 to derive the status and issues. It also presented a policy improvement direction for the future activation of SW convergence R&D.

6. Research Implication and Expected Effects

○ This research provided a stepping stone for SW convergence R&D analysis to continuously identify SW convergence phenomena taking place across national R&D projects and compare and analyze trends.

- Using artificial intelligence technology, we can quickly and accurately compare and analyze the spread of SW convergence on national R&D projects every year by overcoming limitations in existing research that estimate the overall challenges through sampling.

- In addition, text data related to national R&D projects carried out by foreign research institutes such as NSF, DARPA, and EU Horizon 2020 can be obtained to apply an automatic classification model, and the implications can be derived from further studies comparing the results with the domestic situation.

CONTENTS

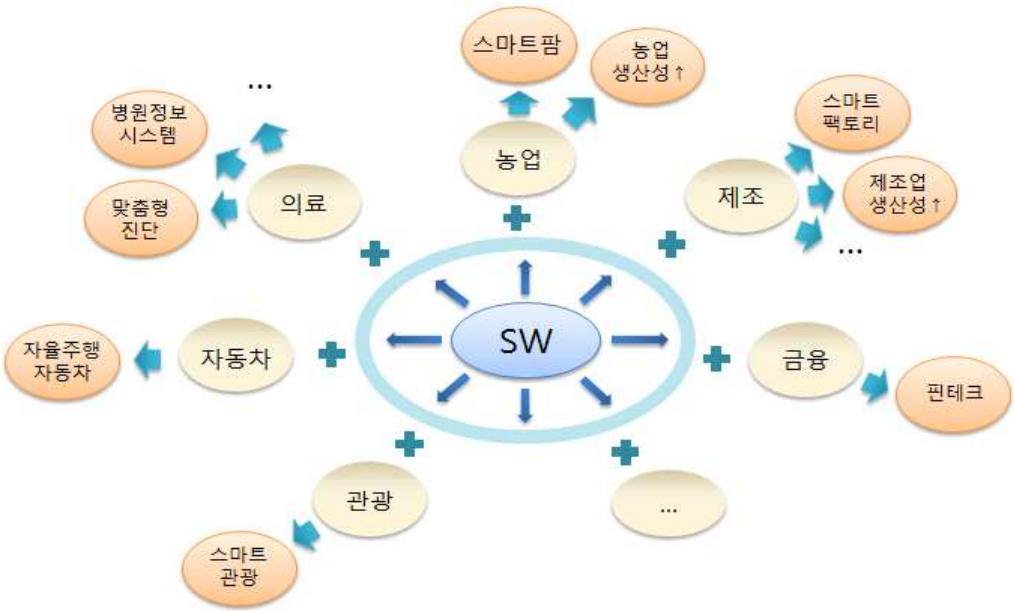
제1장 서론

제1절 연구의 배경 및 목적

1. 연구의 배경

제4차 산업혁명은 초연결 기반의 지능화 혁명으로 각 산업을 넘어서서 사회, 삶 전반의 혁신적인 변화를 유발하고 있다. 제4차 산업혁명을 가능하게 하는 핵심 동인(Enabler)은 소프트웨어로(이하 SW)써 산업의 생산성과 글로벌 경쟁력을 제고하고, 고질적 사회문제 해결을 통해 삶의 질을 높이고 성장 동력으로 연결할 가능성을 제공한다. SW는 [그림 1-1]과 같이 금융, 제조, 의료, 관광, 농업 등 이미 여러 영역에서 활용되며, 각 분야에서 핀테크, 스마트공장, 정밀의료, 스마트관광, 스마트팜이라는 파괴적 혁신을 야기하고 있다.

[그림 1-1] SW + 각 산업의 융합 확대



출처 : SW 융합 R&D 현황 분석 및 시사점, SPRi, 2018

특히, 지능화 혁신의 주체로 인공지능, 빅데이터, 클라우드 등의 SW 신기술 개발 및 고도화를 통해 다양한 산업의 혁신이 가능하다. 이러한 신기술과 산업

의 큰 환경변화에 대응하는 유효한 정책적 수단 중 하나는 연구개발(Research and Development, 이하 R&D)이다. 정부는 국가 차원의 주요 핵심 기술 개발 및 SW 역량 제고를 위해 국가연구개발사업의 로드맵을 제시하고 R&D의 우선순위를 조정하며 글로벌 경쟁력 확보를 위한 정책적 대응을 수행하고 있다.

기존의 국가연구개발사업에서 SW 관련 연구개발은 SW 산업 자체의 발전을 목표로 기획되고 관리되었으나 이제는 SW가 하나의 산업을 넘어 전 산업에서 융합되며 기존 산업의 경쟁력을 높이는 핵심 요소가 되고 있기 때문에 보다 차별화된 방식의 SW R&D 기획 및 관리 방식이 요구된다.

글로벌 시장은 이미 SW 기업을 중심으로 재편되고 있으며, 전 산업에서 AI와 SW가 핵심역량으로 부각하고 있다. 시가 총액 기준으로 글로벌 10대 기업 중 SW 플랫폼 기업은 2008년에 1개였으나 2018년에는 7개로 확대되었다(한국경제연구원, 2018.). 또한, SW기술 발전으로 인해 산업 간 경계가 무너지는 빅블러(Big Blur)현상[2]으로 인해 전 산업의 SW화가 확산되고 있다. 온라인 서점으로 시작한 아마존은 전자상거래, 미디어 유통, 배송 및 클라우드 서비스 기업으로 자리매김하였고, 검색 서비스로 시작한 구글은 모바일 플랫폼, 광고, 자동차 산업에서 선도적으로 활약하고 있다.

<표 1-1> 시가 총액 상위 10위 기업

2008년	
순위	기업명
1	페트로차이나
2	엑손모빌
3	GE
4	중국이동통신
5	**MS**
6	중국공상은행
7	페트로브라스
8	로얄더치셸
9	AT&T
10	P&G

2018년	
순위	기업명
1	**애플**
2	**알파벳**
3	**MS**
4	**아마존**
5	**페이스북**
6	**텐센트**
7	**알리바바**
8	버크셔헤더웨이
9	중국공상은행
10	JP모건체이스

출처 : 한국, 글로벌 시총 500대 포함기업 수... 10년 간 제자리, 한국경제연구원, 2018, S&P Capital IQ 재가공

2) 인공지능, 빅데이터, 핀테크 등 SW 발전에 따른 혁신적 변화로 기존 산업 간, 온-오프라ㄹ인 간 경계가 모호해 지는 현상

국가연구개발사업을 통해 글로벌 경쟁력 제고의 기반이 되는 SW 융합 역량을 확보하기 위해서는 우선 현재 진행되고 있는 SW 융합에 대한 정확한 통계자료가 요구된다. 이러한 현황 데이터를 토대로 SW 융합 역량을 제고하기 위한 국가 차원의 전략 마련이 가능할 것이다.

현재 각 정부 부처 및 분야마다 SW 융합 R&D가 추진되고 있는 것으로 보이나, 정확한 융합 현황을 파악할 수 있는 객관적이고 정량적인 방법은 없는 것이 현실이다. 그러므로 30여 명의 SW R&D 전문가를 활용하여 국가연구개발사업 중 일부를 표본으로 추출하여 정성적 방법론인 델파이 방법을 통해 SW 융합 현황을 파악하였다(공영일, 2018). 그러나 해당 방법은 국가연구개발사업 전체를 파악한 것이 아니라 임의 추출된 표본을 통해 모수에 대한 추정을 하는 방식으로써 이를 수행하기 위해서는 방대한 비용과 시간과 노력이 투입된다. 또한, 연도별 비교를 위한 추이 분석을 위해서는 매년 같은 과정을 반복해야 수행해야 한다는 단점이 존재한다. 이를 극복하기 위해 본 연구에서는 이미 정답이 존재하는 과제 정보를 활용하여 자동으로 SW 융합 R&D 단계를 판단하는 모델을 개발하여 기존의 정성적인 방법보다 효율적인 방식으로 현황 파악을 수행하고자 한다.

2. 연구의 목적

전 산업의 경쟁력을 제고하기 위해 SW 융합 역량을 확보하는 것이 국가연구개발사업의 핵심적인 과업으로 볼 수 있다. 국가적 차원에서 SW 융합 역량을 높이기 위해서는 국가연구개발 사업에서의 SW 융합 정도에 대한 정확한 판단과 현황 분석 데이터의 확보가 필요하다.

본 연구의 목적은 우리나라의 SW 융합 역량 제고를 위해 국가연구개발사업에서의 SW 융합 현상에 대해 지속적이고 정량적으로 분석할 수 있는 분류 모델을 개발하고, 이를 통한 현황을 조사・분석하여 SW 융합 촉진을 위한 SW R&D 정책 방향을 제시하는 것이다.

본 연구의 연구 질의(Research Question)는 다음과 같다.

- 국가연구개발사업에서의 SW 융합 현상을 자동으로 분류하여 지속적으로 현상을 분석할 수 있는 방법이 있는가?

- 구체적으로 주요 부처별, 주요 분야별로 SW 융합이 어느 정도 이루어지고 있으며, 그 특성은 어떠한가?

- SW 역량 제고를 위한 SW 융합 촉진 정책 방향은 무엇인가?

본 연구는 이러한 연구 목적을 달성하기 위해 2018년에 수행한 선행 연구에서 전문가 델파이 수행을 거쳐 SW 융합 R&D 단계가 결정된 총 2,000개의 표본 추출 데이터를 활용하여 SW 융합 R&D의 단계를 구분하는 분류 모델을 개발한다. 해당 분류 모델에 의해 2017년 국가연구개발사업 전체 과제를 대상으로 부처별, 과학기술표준분류체계의 분야별, 연구수행 주체 등 유형별 SW 융합 R&D 현황 분석을 통해 SW 융합 활성화 정책을 마련하기 위한 기초 자료를 확보하고자 한다.

3. 연구의 구성

본 연구의 구성은 다음과 같다.

제1장은 서론으로 연구의 배경과 목적을 설명한다. 제2장은 SW R&D 규모에 관한 기존에 이루어진 연구에 대한 검토를 진행하고, 국외에서 수행되고 있는 SW 융합 연구개발 관련 추진 정책 방향에 대해 인공지능 기술을 중심으로 살펴본다. 그리고 본 연구에서 개발한 SW 융합 R&D 자동 분류 모델을 설명하기 위한 설명 가능한 인공지능에 대한 선행 연구 내용을 소개한다. 제3장에서는 본 연구에서 활용하는 자료에 대해 설명한다. 특히, 본 연구에서 활용하는 학습 데이터의 이해를 위해 데이터의 분포와 비중에 대한 분석을 진행하며 본 연구의 진행 단계에 대해 소개한다. 제4장에서는 기계학습 모델에 대한 실험 결과를 제시하고 본 연구에서 활용하는 텍스트 마이닝 방법과 학습 데이터에 대한 비중 및 기계학습 모델을 선정한다. 제5장은 개발한 분류 모델을 활용하여 2017년 국가연구개발과제에 대한 SW 융합 R&D 예측 결과를 제시하고 수행 부처 및 연구개발 단계, 수행 주체 등의 유형별 분석을 진행한다. 또한, 해당 결과에 대한 SW 융합 R&D 전문가의 분석 결과를 제시하여 연구 결과를 검증하며, 개발한 분류 모델이 어떠한 이유로 각 과제를 예측한 것인지에 대해 설명한다. 마지막으로 제6장에서는 수행된 SW 융합 R&D의 유형별 분석 결과를 토대로 SW 융합 활성화를 위한 시사점을 도출하고 SW 융합 R&D 정책 방향을 제시한다.

제2장 선행 연구 분석

제1절 SW 융합 R&D 관련 분석 연구

SW 융합 R&D와 관련된 선행 연구는 본 연구의 1차년도 연구인 “국가연구개발사업에서의 SW 융합에 관한 연구, SPRi(2018)” 가 있다. 해당 연구에서는 SW 융합 R&D의 현황을 파악하기 위해 SW R&D 전문가가 방대한 시간과 비용을 투입하여 수행하는 정성적 방법론인 델파이 방식으로 과제를 판단하였다. 우선 전 부처 및 분야에서 수행된 SW 융합 R&D 현황을 파악하기 위해 아래 <표 2-2>와 같이 SW 융합 R&D 판단 기준을 제시하고, 30명의 전문가를 대상으로 4개월에 걸쳐 델파이 분석을 수행하였다. 판단의 대상이 되는 과제는 2016년 국가연구개발과제인 총 54,827건 중 2,005건을 표본으로 층화 추출하여 3라운드에 걸친 전문가 델파이 분석을 진행하여 2,000건의 과제에 대해 합의를 이루었다. 이를 통해 전체 2016년 국가연구개발과제에 대한 규모 및 유형별 현황을 추정하였다.

<표 2-1> SW 융합 R&D 판단 기준

Scale	Description
High	• 연구개발 활동의 대부분이 SW를 연구개발하는 과제 (80% 이상 ~ 100%) • EX. 새로운 Word Processor 개발, 스마트폰App(예: T Map) 개발, 새로운 알고리즘, Program Source 등
Med High	• 많은 부분이 SW를 연구개발하는 활동이며, 이와 함께 부분적으로 비SW의 개선·개발을 추진 (60% 이상 ~ 80% 미만) • EX 1. SW의 전반적인 개선과 부분적 비SW 기능 추가 혹은 대체 • EX 2. 기존 내비게이션 시스템에 신규 기능(예 : AR/VR)을 주도적으로 추가 개발하고, 일부 출력용 HW 모듈을 함께 개발
Medium	• SW와 비SW를 동시에 연구개발을 추진 (40% 이상 ~ 60% 미만) • EX. 기존내비게이션을 3D화하기 위한 관련 HW와 SW를 동시에 개발
Med Low	• 부분적으로 SW를 연구개발하며, 대부분 새로운 비SW 연구개발을 추진 (20% 이상 ~ 40% 미만) • EX 1. 새로운 HW 개발에 따른 부분적 SW 수정 혹은 개발 • EX 2. 새로운 고속 처리 내비게이션 HW를 개발하고, 기존 내비게이션 SW를 일부 수정하여 포팅
Low	• 기존 SW를 단순 활용하고, 대부분 새로운 비SW의 연구개발을 추진 (0% 이상 ~ 20% 미만) • EX 1. 안드로이드OS 포팅, SI (요구사항에 의한 단순 코딩/활용) 등 • Ex 2. 새로운 고성능 내비게이션 디바이스를 개발하고, 기존 내비게이션 SW를 단순히 포팅하는 과제

출처 : 국가연구개발사업에서의 SW 융합에 관한 연구, SPRi, 2018

본 연구에서는 위의 SW 융합 R&D 기준에 따른 구분 결과를 입력 데이터로 활용하였다. 서영희(2018)는 “SW 융합 R&D 현황 분석 및 시사점”에서 SW 융합 R&D를 “SW 연구개발 활동이 포함된 연구개발을 수행하는 것”이라고 정의하였다. 즉, 연구개발을 수행하는 과정에서 단순히 SW를 사용하는 것을 제외하고 어떤 형태로든 어느 정도의 SW 관련 연구개발 활동을 포함하는 것을 뜻한다. 여기서 SW[3]는 “컴퓨터, 통신, 자동화 등의 장비와 그 주변장치에 대하여 명령 · 제어 · 입력 · 처리 · 저장 · 출력 · 상호작용이 가능하게 하는 지시 · 명령의 집합과 이를 작성하기 위하여 사용된 기술서나 그 밖의 관련 자료”이다.

위의 1차년도 연구 결과, 2016년 수행된 총 54,827건의 국가연구개발과제에서 표본으로 추출된 2,000건 과제의 총 연구비[4](7,510억 원) 중 SW 융합 R&D 활동을 일부라도 포함하는 비중은 16.3%(1,220억 원)로 나타났다. 그러므로 2016년에 수행된 국가연구개발사업 총 연구비인 약 22조 4천억 원[5]중의 SW 융합 R&D는 약 3조 7천억 원으로 추정할 수 있다. 또한, 과제 수 기준으로 약 15.9%가 SW 융합 R&D 과제인 것으로 파악되었다.

이에 본 연구에서는 위 연구데이터를 기반으로 SW 융합 R&D를 총 다섯 가지(High, Med High, Medium, Med Low, Low) 단계로 구분한 2,000건의 데이터를 활용하여 해당 단계에 속하는 과제들의 연구 제목 및 요약문의 텍스트 정보를 학습하여 자동으로 SW 융합 R&D의 단계를 분류하는 모델을 만들고자 한다.

3) 소프트웨어산업진흥법 제2조 2항
4) 총 연구비는 정부연구비와 매칭펀드 금액을 합한 금액으로, 항목별로는 현금과 현물에 해당하는 인건비, 직접비, 간접비, 위탁연구비를 포함함
5) NTIS에서 제공한 2016년 54,827개 과제의 총 연구비의 합

제2절 국가별 SW 융합 기술 발전 정책 및 특징

1. 개요

ICT 기술 발전 가속화에 따라 제4차 산업혁명 시대가 다가왔으며, 주요 선진국들은 이와 관련한 국가 차원의 전략을 수립 및 추진 중이다. 2016년 1월 개최된 다보스포럼에서도 제4차 산업혁명에 대한 주제 논의가 있었고, ICT 강국들은 제4차 산업혁명에 대비한 다양한 전략들을 앞다투어 발표했다. 전문가들은 제4차 산업혁명 시대의 핵심기술로 인공지능, 빅데이터, IoT 등 SW 융합 중심의 ICT 기반 기술들을 언급하고 있으며, 이러한 기술들을 중심으로 전략을 수립 및 추진하고 있다.

미국은 2011년부터 스마트제조에 관한 필요성을 지속적으로 언급하고 있으며 국가 제조업 혁신 네트워크(National Network for Manufacturing Innovation, NNMI), ICT 연구개발 기본계획 NITRD(Networking and Information Technology Research and Development) 등 다양한 ICT 관련 정책을 추진하고 있다. 독일은 하이테크 전략 2020과 함께 국가 차원의 전략인 인더스트리 4.0을 발표하면서 정부 중심으로 제4차 산업혁명에 적극적으로 대응할 것을 피력하고 있다. 일본은 세계 최첨단 IT 국가 창조선언 등 정부 차원의 ICT 전략을 추진하고 있으며 일본이 강점으로 하는 로봇기술을 적극 활용하는 제4차 산업혁명 대응전략을 발표하였다. 중국 역시 ICT를 기반으로 중국 제조업을 한 단계 업그레이드할 수 있는 중국 제조 2025를 발표하였으며, 그 밖에 인터넷 플러스 등 ICT와 관련한 다양한 정책을 발굴하여 추진 중이다.

이러한 전략을 기반으로 미래 기술경쟁력 확보를 위한 글로벌 경쟁이 진행되고 있으며 특히 최근에는 자율주행, 스마트홈 및 스마트팩토리 등 기술 개발이 가속화되면서 대표적인 SW 융합 기술인 인공지능 기술 개발을 위한 전략을 앞다투어 발표하고 있다. 본 절에서는 국내외 선행 연구에서 분석한 내용을 기반으로 SW 융합 기술의 대표기술인 인공지능 관련 국가별 정책과 전략을 중심으로 정리하고자 한다.

2. 미국의 SW 융합 (인공지능) R&D 정책

가. 배경 및 목적

미국의 인공지능 기술 전략은 사회, 경제적 편익을 추구하는 것으로부터 출발한다. 인공지능은 엄청난 사회, 경제적 편익을 약속하는 변형 기술(transformative technology)로 생활, 업무, 학습, 발견, 소통방법에 대변혁을 일으킬 잠재력 보유하고 있다고 보며, AI 연구로 인해 경제적 번영 확대, 교육기회 및 삶의 질 개선, 국가 및 국토 안보 강화를 포함한 국가적 우선 과제를 더욱 성공적으로 진행 가능한 것으로 보고 있다.

백악관은 “인공지능의 미래를 준비합니다”를 공지하며, 국가과학기술위원회(National Science and Technology Council, NSTC) 산하에 ‘머신러닝 및 AI 분과위원회’를 설치(2016.05.)하였고, 위 분과위원회는 네트워킹 및 정보기술 R&D 분과위원회(NITRD) 국가조정사무국(National Coordination Office, NCO)에 본 보고서 발간을 지시(2016.06.)하여 NSTC 산하 NITRD 분과위원회는 「국가 AI R&D 전략계획」 수립을 발표(2016.10.)하였다.

나. 미국의 인공지능 R&D 전략 계획

1956년에 있었던 역사적인 회합으로 정부 및 산업 차원의 AI 연구의 장이 마련되었고, 오늘날 경제 분야 형성 및 사회복지에 커다란 혜택 제공하게 되었다. 지난 25년 동안 통계적 및 확률론적 방법 채택, 대규모 데이터 유용성, 증가된 컴퓨터 처리능력으로 AI 접근 방식의 중요성이 신장되었으며, 2015년, 미국 정부가 기밀로 취급하지 않는 AI-기술 R&D에 투자한 금액은 약 11억 달러이며, 이로써 중요한 새로운 과학 기술을 주도하게 되었다.

AI의 변혁적인 효과를 인식하고, 백악관 과학기술정책국(Office of Science and Technology Policy, OSTP)은 이 보고서와 함께 ‘AI의 미래에 대한 준비’ 보고서를 발표(2016.10.)하였다. 또한, NSTC는 머신러닝 및 AI 분과위원회를 신설(2016.05)하였고, 해당 분과 위원회가 NITRD NCO에 국가 AI R&D 전략계획 수립의 과제를 부여(2016.06.)하였다. AI R&D 전략계획에는 광범위한 자원들이 투입되었으며 본 계획은 AI의 미래에 대한 몇 가지의 가정을 설정하고 있다.

첫째, 정부 및 산업에 의한 AI R&D 투자에 힘입어 AI 기술이 지속해서 더욱 정교해지고 어디에나 존재할 것이다. 둘째, AI가 미국의 경제 성장뿐만 아니라 고용, 교육, 공공안전 및 국가안보를 포함하여 사회에 미치는 영향이 지속적으로 증가할 것이다. 셋째, AI에 대한 산업투자가 지속해서 증가하는 동시에 연구의 중요한 영역 중의 일부는 산업부문으로부터 충분한 투자를 받지 못할 것이다. 마지막으로 AI 전문 인력에 대한 수요가 산업계, 학계, 정부 내부에서 지속해서 증가하여 공공 및 민간 부문의 인력난을 초래하는 것이다.

AI에 있어서 최근의 진보는 AI의 잠재력에 관해 의미 있는 낙관론을 심어주었고, 그 결과 강력한 산업 성장과 AI에 대한 접근의 상업화를 야기하고 있다. 연방정부는 민간 산업부문이 추구하지 않는 중요한 사회적 사안들을 다루기 위해 소비자 시장이 겨냥하지 않는 사회적으로 중요도가 큰 영역에 투자하는 것을 원칙으로 하며, AI R&D 전략계획은 중요한 기술적, 사회적 과제들을 다루는 AI에 대한 장·단기적 지원을 위한 전략적 우선 과제를 확인하여 다양한 분야에 국가적 우선순위 추진을 위한 비전을 수립하여 추진하고 있다.

다. 미국 인공지능 R&D 전략

미국의 인공지능 R&D는 산업계에서 대응할 개연성이 낮아 연방정부의 투자로 편익을 확보할 가능성이 높은 영역에 초점을 두고 7대 AI R&D 전략의 방향성 제시하고 있으며, 지각, 자동화된 추론/기획, 인지 시스템, 머신러닝, 자연언어처리, 로봇공학 및 관련 분야의 모든 AI 하위분야에 공통되는 수요를 포함하고 있다.

<표 2-2> 미국의 인공지능 AI 전략 7가지

구분	상세 내용
전략 1	고위험, 고보상 기초연구에 대한 연방정부 투자 대부분은 혁신적인 기술(인터넷, GPS, 스마트폰 음성인식, 심장 모니터, 태양광 패널, 첨단 배터리, 암 치료법 등)로 이어졌으며, AI 분야에서 글로벌 리더십을 유지하기 위해 최우선 순위의 기초적이고 장기적인 AI 연구에 투자 집중
전략 2	인간과 AI 간의 상호작용 및 협업을 위한 효과적인 방법을 개발하기 위한 연방정부의 본질적 연구(사용하기 쉽고 유익한 방법으로 인간과 함께 일할 AI 시스템을 어떻게 가장 잘 만들어 낼 것인가를 연구) 필요
전략 3	윤리적, 법적, 사회적 목표들과 보조를 같이하는 AI 시스템을 개발함으로써 전 영역에서 AI가 미치는 영향을 이해 및 대처하기 위한 연구 필요 ※ AI의 진보는 사회에 많은 긍정적 혜택을 제공하며 미국의 경쟁력을 증가시키고 있으나 직업, 경제, 안전, 윤리적 및 법적 문제에 이르는 몇몇 영역에서 위험 표출

전략 4	사용자가 신뢰 및 수용할 수 있는 방법으로 일을 수행하며, 사용자의 의도대로 작동하는 것을 보장할 수 있는 설명 가능하고 투명한 시스템 필요 ※ 인간 사용자 및 환경과 풍부한 상호작용이 가능한 AI 시스템의 잠재적 역량과 복잡성으로 인해 AI 기술의 보안과 제어를 강화하는 연구에 투자하는 것이 중요
전략 5	AI 연구의 진전을 촉진하고 대안적 해법을 보다 효과적으로 비교할 수 있도록 AI 훈련 및 시험용 공유 공공 데이터 세트에 투자할 것을 연방정부에 촉구
전략 6	특정 문제와 도전적 과제에 대한 진전을 정의하고, 차이를 좁히고, 혁신적인 해법을 추진하기 위한 R&D에 표준 및 벤치마크가 어떻게 초점을 맞출지 논의 ※ 인표준 및 벤치마크는 AI 시스템을 측정하고 평가할 뿐만 하니라 AI 기술이 기능성과 정보처리 상호운용이라는 중대한 목표달성을 보장하는데 필수적
전략 7	AI 기술에 대한 심층적 기술을 가진 핵심 AI 과학자 및 기술자 등 역량을 갖춘 인재 파이프라인을 충분히 확보할 수 있도록 조치

3. 독일의 SW 융합 (인공지능) R&D 정책[6)]

가. 배경 및 목적

유럽은 독일의 인더스트리 4.0 발표 이후 제4차 산업혁명에 대한 관심을 크게 고조시켰고, 주요 선진국을 중심으로 제4차 산업혁명에 대한 다양한 대응전략이 발표되고 있다. 유럽은 독일을 중심으로 하이테크 전략 2020으로 분야를 초월하여 기술 혁신을 가져올 수 있는 다양한 정책을 지원하고 이를 추진하는 목적으로 추진하고 있다. 하이테크 2020 전략은 2011년 독일 자국 내에 「Industry 4.0(이하 인더스트리 4.0)」 개념이 소개되면서 2012년 결정된 하이테크 전략 2020에 인더스트리 4.0이 새롭게 편입되어 5대 중점분야와 10대 미래 프로젝트로 구성되어 추진되고 있다.

나. 인공지능 분야 구분

인공지능 관련된 개념이 일관적이고 보편적으로 정립되지 않았음을 감안하여 혼용되고 있는 인공지능의 의미를 강 인공지능과 약 인공지능으로 구분할 수 있다. 강 인공지능은 인간과 동일하거나 이를 능가할 수 있는 인공지능을 개발하는 데 중점을 두고 있는 반면, 약 인공지능은 수학과 컴퓨터과학을 통한 방법론으로 특정 응용문제를 해결하는데 초점을 두고 있다. 이를 기반으로 자기 최적화(self-optimization)가 가능한 시스템을 개발하는 것을 목적으로 한다. 여기에 추가로 인간의 지능을 모델링하고 기술하거나 인간의 생각하는 방식을 시뮬레이

6) 독일 정부의 인공지능(AI) 선도 전략(한국산업기술진흥원, 2019)을 중심으로 정리

션하는 기술도 이 범주에 포함된다고 할 수 있다. 독일 정부는 이러한 구분 중 약 인공지능에 방점을 두고 이를 위한 방안을 개발하고자 하며, 구체적으로는 5대 분야로 구분하고 있다.

<표 2-3> 독일의 인공지능 중점 5대 분야

분야	적용 예
추론 시스템, 기계 기반 근거 처리	논리적 표현에서 형식적 진술을 파생시키며, HW 및 SW의 정확성을 입증하는 시스템
지식 기반 시스템	지식 모델링 및 수집 방법, 시뮬레이션용 SW, 인간 전문기술 및 전문가 지원(이전에는 대개 전문가 시스템으로 불림), 심리학 및 인지과학
패턴 분석 및 인식	추론 분석 기법, 그 중에서도 기계학습
로봇 공학	로봇 시스템의 자율적 제어. 자율 시스템
지능형 멀티모드	인간-기계 상호작용, 언어분석 및 이해(언어학과 연계), 이미지, 제스처 및 기타 인간 상호작용

출처 : 독일 정부의 인공지능(AI) 선도 전략, 한국산업기술진흥원 유럽사무소, 2019.06

다. 예산 지원 계획 및 기대효과

독일은 이미 인공지능 분야에서 탁월한 위치를 점하고 있다고 볼 수 있다. 독일 정부의 전략은 기존의 강점을 강화하는데 그치는 것이 아니라 이를 확대하여 잠재력이 낮은 분야에서도 연쇄적인 효과를 가져오는데 그 목표를 두고 있다. 이에 2019년에는 초기 예산으로 5억 유로의 예산을 편성하였고, 이후 유사한 규모의 예산을 지속적으로 투입할 계획으로 2025년까지 인공지능 관련 정책 전략 실현에 집중하기 위해 총 20억 유로를 투자할 예정이다. 이를 통해 민간뿐만 아니라 국공립 연구소 및 유관 기관 및 지자체는 AI 관련 예산 규모를 2배까지 확대 가능해질 것이다.

라. 독일 정부의 SW 융합 선도 전략

독일은 2018년 11월에 인공지능전략 관련 보고서를 발표하고 정부 차원의 인공지능 분야 연구개발 정책을 디자인하고 관련 분야 혁신을 촉진하고 있다. 특히 인공지능 선도국으로서 위상을 확고히 하고, 국제 경쟁력을 확대하여 경제 발전에도 기여하는 것은 물론이고, 다양한 응용 가능성을 개발하여 시민과 다양한 사회 영역에 긍정적인 변화를 유도하려는 노력을 기울이고 있다. 이를 통해

삶의 질 향상과 환경 보전이라는 목표에 초점을 맞추고 이와 동시에 모든 사회 구성원 그룹과 밀도 있는 의사소통을 실현하려는 의지를 피력하고 있다.

4. 중국의 SW 융합 (인공지능) R&D 정책

가. 배경 및 목적

중국은 정부의 양적 투자를 바탕으로 G2로 성장하였으나, 최근에는 경제 산업 구조의 질적 향상과 지속성장을 위해 AI 등의 ICT 혁신기술에 대한 관심 증가하고 있다. 특히, 중국 정부는 새로운 기술을 창출하고 다양한 산업과 융합되어 중국 경제구조의 변환을 가져올 것으로 기대되는 일련의 AI 투자・지원 정책을 연이어 발표하고 있다. 중국은 1970년대부터 학계, 산업계를 중심으로 AI를 도입하여 발전시켜왔으며, 2014년부터 AI를 산업고도화 수단으로 인식하고 정부 차원에서 공식적 접근 시작하였고, 주요 성장동력인 제조업의 침체 위기를 극복하고 산업 강대국의 입지를 공고히 하기 위한 방안으로 「중국제조 2025」을 발표하고, 이후 AI 관련 다양한 정책들을 제시하고 있다. 최근(2017.07.)에는 중국의 글로벌 선도 전략으로 AI를 채택하여 개발 방향성을 설정하는 「차세대 AI 발전 계획」을 발표하면서 중국의 AI 정책은 산업 고도화 수단에서 주요 산업, 그리고 미래를 위한 국가전략으로 진화하고 있다.

본 보고서에서는 2000년 이후 ICT 기술 대응 단계부터 최근 중국의 AI 관련 정책 및 전략의 변화를 살펴보고 국내 전략 수립에 참고할 수 있는 자료를 제시하고자 한다.

나. 중국의 SW 및 AI 발전과 현황

1) 중국의 AI 발전 과정

① AI 기초 개발 단계: 인터넷 업계 활성화 (2000~2012년)

2000년대 초 중국은 기계학습기술 기반의 검색엔진 전문기업이 등장하고, 다양한 지능형 로봇과 게임 산업이 발전하면서 인터넷 업계가 활성화되면서, 기계

학습기술의 획기적인 발전과 추천시스템 등 광범위한 응용 프로그램을 기반으로 새로운 비즈니스 모델 등장하기 시작했고 지능형 산업 로봇, 인간-컴퓨터 간 경쟁 기반 게임, 필체 식별 등 중국 AI 관련 기술의 성공적 개발로 산업계 활성화 되었다.

② AI 성장 단계: 컴퓨터 능력 향상에 따른 기반 구축 공고화 (2012~2015년)

이 시기 알파고와 같이 처리속도가 빠르고 서버 입력 시 전력 소모가 적은 CPU・GPU 모드가 보편화되고, 주요 중국 ICT 기업 중심으로 AI 개발 및 투자 활성화를 가져왔고, 바이두, 알리바바, 화웨이 및 기타 주요 중국 기업 중심으로 AI 분야 적극적 개발이 시작되면서, 중국 정부는 2014년부터 AI에 대한 정부 차원의 공식적 전략을 마련하기 시작하였다.

글로벌 경쟁에서 최고 위치에 오르기 위해서 과학기술에 대한 혁신 및 독립적 능력이 필요함을 강조하였다. 이에, 시진핑 주석은 중국과학원 제17대 원사대회에서 새로운 혁신기술인 AI를 발전시키기 위한 계획이 필요하며 이를 착실히 추진해야 함을 강조하면서 중국은 AI를 전통 제조업과 융합하여 중국의 경제구조를 새롭게 전환시킬 수 있는 도구적 측면으로 접근하였다.

③ AI 개발 촉진 단계: 정책과 산업계 지지 (2015년~현재)

중국은 알파고 바둑 대국 이후 주요한 산업이자 상품인 AI 산업 분야 육성 필요성을 강조하였고, 중국은 AI 기술을 세계를 이끌어나갈 주요 동력 및 중국 사회를 선도할 본격적인 국가 전략수단으로 제시하였다.

시진핑 주석은 2017년 1월 처음으로 참석한 세계경제포럼에서 AI가 세계 경제 성장을 주도할 것으로 예측하고, 환경변화에 대응하기 위한 계획의 필요성을 강조하였다. 2017년 전국인민대표회의에는 AI가 국가의 주요 의제로 선정되었고, 최고위 회의에 AI가 주요 안건으로 최초 포함하면서 AI 관련 정책 변화를 가져갔다.

2) 중국의 AI 정책 변화

AI에 대한 지속적인 관심과 정책적 강조에 따라 2015년 이후 AI 관련된 다양

한 정책을 연이어 수립하고 있다.

<표 2-4> 중국 AI 관련 정책 내용

계획	구체적 내용
「중국제조 2025 (2015.05.)	국무원은 독일의 4차 산업혁명을 참조하여 제조업을 고도화하는 방안으로 제조 全 산업에 AI를 결합하는 중국 최초의 AI 정책 수립
로봇산업발전계획 2016~2020 (2016.03.)	공업정보화부, 국가발전개혁위원회는 「중국제조 2025」와 연계하여 중국 산업 구조를 개선하기 위한 방안 수립
인터넷+ AI 3개년 실천방안 (2016.05.)	2015년에 발표된 「인터넷+」 전략에 AI 관련 내용을 구체적으로 제시하고, AI 산업 발전에 필요한 정부 지원 계획과 방향성을 제시
소비재 표준 및 품질향상 계획 2016~2020 (2016.09.)	국무원은 중국 재화의 소비 촉진을 위해 지능형 소비재의 표준화를 포함하는 계획 수립
차세대 AI 발전계획 (2017.07.)	국무원은 중국 미래를 위한 주요 국가 전략으로 AI 기술을 채택하고 개발 방향성을 설정

3) 중국의 AI 현황

중국의 산업 규모는 100억 6천 위안(약 1조 6,599억 원)으로 2016년 대비 43.3% 성장하는 등 가파른 증가 추세이다. 중국의 AI 기업은 709개로 미국(2,905개)에 이어 두 번째로 많은 숫자로 바이두・알리바바・텐센트 등 거대 ICT 기업들을 중심으로 AI 산업 발전을 선도하고 있다. 또한, 연구개발 분야에 있어 중국은 정부의 적극적인 인재영입 지원 및 인력양성 정책을 바탕으로 약 10만 명의 AI 전문 연구 인력을 확보하여 AI 연구 및 특허 개발에 집중하고 있다.

다. 중국의 SW 및 AI 주요 정책

1) 중국제조 2025

제조 분야에 대한 포괄적 지능화를 통해 건국 100주년인 2049년까지 일본, 독일 등 경쟁국을 선도하는 세계 최고 제조국가 지위 획득을 목표로 3단계 전략인 「중국제조 2025」 발표하였다(2015.05.).

<표 2-5> 중국제조 2025의 3단계 전략 내용

단계 및 시기	상세 내용
1단계 (~ 2025년)	2020년까지는 제조업 전반에서의 ICT 활용 수준 개선 및 핵심 경쟁력 제고, 2025년까지 제조업의 전반적 품질과 에너지 소비를 세계 선진 수준에 도달
2단계 (~ 2035년)	중국 제조업 수준을 글로벌 제조 강국의 중간수준까지 개선하고, 중국이 강점을 지니는 산업의 글로벌 경쟁력 확보 * 2단계 기간 동안 세계 제조업 제2그룹 대열에 진입을 목표로 추진
3단계 (~ 2049년)	세계 최고의 기술 및 산업시스템을 구축하여 글로벌 시장 선도 * 3단계 기간 내 제1그룹 대열에 진입을 목표로 추진

중국의 제조업 고도화를 위한 제조의 스마트화 달성을 위해 5대 기본원칙 제시하여, 산업 규모는 세계 1위이지만 통합 제조 시스템 구축 등 산업 전반의 고도화 수준은 글로벌 경쟁력 수준에 비해 뒤처지는 중국의 제조업을 고도화하기 위해 혁신능력 제고, 품질 강화, 녹색개발, 구조 최적화, 인재양성을 포함하는 5대 기본원칙을 마련하였다.

또한, 전체 산업의 지능화 및 혁신능력 제고 정책추진과 함께 아래와 같은 10대 핵심 분야 선정하여 추진 중이다.

<표 2-6> 중국의 지능화 10대 핵심 분야

핵심 분야	상세 내용
① 차세대 정보기술 산업	집적 회로 및 특수 장비, ICT 장비, 운영체제 및 산업 SW 분야
② 고정밀 수치제어 및 로봇	고급 디지털 선반, 로봇의 표준화·모듈화 발전 촉진 및 시장 응용범위 확대 등을 통한 신제품 개발 등을 포함
③ 항공 우주장비	항공 장비, 우주 장비 등을 포함
④ 해양 공학 장비 및 첨단 기술 선박	심해탐사, 자원 개발·이용, 해상작업 안전장비와 시스템 및 전용설비 발전, 해저정거장, 대형 부유식 구조물 구축 등
⑤ 고급 철도 장비 등 선진 교통 장비	신소재 개발과 시스템 보안, 환경 보호, 지능형 디지털 네트워크 기술 등의 주요 연구 분야 등
⑥ 에너지 절약 및 신에너지 자동차	엔지니어링, 산업용 모터, 효율적인 내부 연소 엔진, 고급 변속기, 경량 소재, 지능형 제어 등
⑦ 전력 설비	고성능 전력 전자부품, 고온 초전도소재 등 핵심 기술 개발과 더불어 대형 발전 장치의 상용화 및 시범사용, 신재생에너지 설비와 첨단 에너지 저장 장치 등
⑧ 농기계 장비	첨단 농업 장비를 이용하여 일반 식량 및 경제작물의 파종·재배·수확·운반·저장 등
⑨ 신소재	특수 금속 기능성 재료, 고성능 구조재, 기능성 고분자 재료, 특수비금속 무기질 재료, 첨단복잡소재 등
⑩ 의학 및 의료장비	중증 질환 대상 화학·중의학·생명공학·바이오의약품 개발, 의료기기 등

2) 인터넷+ AI 3개년 실천 방안

중국은 양적 확대 중심의 제조업에서 벗어나 질적 성장을 이끌어 나갈 주요 핵심동력인 AI 육성을 위한 실천 방안을 수립하였으며, 주요 내용은 다음과 같다. 문헌, 음성, 이미지, 영상, 지도 등 대용량 디지털 라이브러리 및 기초 자원 서비스 플랫폼 개발을 가속화하여 대규모 학습능력 컴퓨팅 클러스터 구축과 신흥 산업 육성에 집중한다. AI 기술과의 융합을 통해 가정용품, 자동차, 무인시스템, 보안 설비 등 중점 지원 분야를 선정하여 스마트 상품혁신 추진. 지능형 단말기 기술 관련 R&D 및 상용화를 촉진하여 다양한 이동형·부착형 스마트 단말기, 시뮬레이션 등 제품 공급량을 확대한다. 그리고 AI 활용과 스마트화를 위해 자금 지원, 표준체계 확립 등 6대 주요 대책 제시 등의 내용을 포함하고 있다.

3) 차세대 AI 발전 계획

중국은 국가와 사회 전반의 변혁을 이끌어나갈 주요 동력인 AI를 국가 전략화하기 위해 「차세대 AI 발전계획」을 발표(2017.07.)하고 2030년까지 AI 이론, 기술, 적용의 전 분야에서 글로벌 선도국 지위에 도달하기 위한 3단계 전략 목표를 설정하였다. 중국은 글로벌 AI 시장 선도를 위한 전략 방향으로 다음과 같은 6대 중점 과제를 제시하고 있다.

<표 2-7> 중국 차세대 AI 발전계획의 6대 중점 과제

6대 중점 과제	내용
① 개방·협동형 AI 과학기술 혁신 시스템 구축	• 기초연구, 핵심 공통기술, 혁신채널, 우수 인재 등에 대한 혁신시스템 구축 강화
② 최첨단·고효율의 스마트 경제 육성	• AI 신흥 산업 발전, AI 산업 스마트화 촉진, AI 혁신 기지 조성
③ 안전하고 편리한 스마트 커뮤니티 건설	• 고효율 스마트 서비스 발전, 사회관리 스마트화 제고, AI를 통한 공공안전보장 능력 강화, 사회의 소통·공유·공조 신뢰 촉진
④ AI 영역의 군민융합 강화	• AI 기술 촉진을 통한 군민 협력, 군과 민이 가진 AI자원의 공동 활용방안 구축
⑤ 안전, 고효율의 스마트 인프라 시스템 구축	• 인터넷 네트워크, 빅데이터, 고성능 컴퓨팅 등 인프라 설비 발전
⑥ 차세대 AI 주요 항목 구성	• 과학기술 과제전망과 차세대 AI 특징에 대한 기초이론, 공통의 중요기술 정체 현상*에 대응하기 위해 전반적인 통합계획 강화 * 기술개발 속도가 AI 산업발전 속도를 따라가지 못하는 현상 발생 • 차세대 AI 중대 과학 기술 과제를 핵심으로 한 현재와 미래 연구개발 과제 기반의 AI 과제 통합 체계화

출처: 해외 ICT R&D 정책동향 2017-05호, 정보통신기획평가원

이 외 중국은 글로벌 AI 시장을 선도할 수 있도록 재정, 금융, 사회적 자본 등의 다양한 자원을 적재적소에 투입하고 있으며, 중국 내 AI 발전을 위한 관련 법제 및 규범을 제정하고, 새로운 정책을 수립하고 있다. 여기에는 AI의 전면적, 장기적 계획 실행을 위한 조직・시스템・로드맵 구축, 여론 형성 등의 내용을 포함하고 있다.

5. 일본의 SW 융합 (인공지능) R&D 정책

가. 배경 및 목적

전 세계적으로 제4차 산업혁명의 주요 핵심동력인 AI에 대한 관심이 증대되면서, 일본 또한 산업경제 관점의 AI 정책으로 「일본재흥전략」, 과학기술 관점의 AI 정책으로 「제5기 과학기술기본계획」 등을 통해 미래 지능사회 동력 확보를 위한 AI 정책 방향성을 설정하고 실행 전략으로는 「AI 산업화 로드맵」을 발표하였다. 본 연구에서는 일본의 AI 정책과 실행전략 분석을 통해 우리나라에 유의미한 시사점을 도출하고 한국의 SW 융합 특히 인공지능 정책 수립에 도움이 되는 자료를 제공하는 것을 목적으로 한다.

나. 일본재흥전략 - 산업경제 관점의 일본 AI 정책

아베 정부는 장기화된 일본 경제의 부진을 탈피하고 미래 시대를 준비하기 위해 「일본재흥전략」을 매년 발표하고 있다. 특히 2015년부터는 저출산・고령화로 인한 생산 가능인구 감소와 제4차 산업혁명이라는 내부 및 외부의 도전 과제를 해결하기 위해 미래 투자를 통한 생산성 혁명을 주요 전략으로 채택하였다. 주요 시책으로는 범부처 차원의 AI, IoT, 빅데이터 등의 혁신기술 활용 극대화 방안을 제시하고 있다.

2016년 재흥전략은 2015년의 목표를 재확인하고, ‘4차 산업혁명을 향해’라는 부제에 부합되도록 AI, IoT, 빅데이터 등을 활용한 「민관 전략 프로젝트 10」을 핵심시책으로 발표하였다. 2016년 일본재흥전략에서는 자율주행, 스마트 공장, 소형 범용 로봇 등의 신기술 도입을 통해 2020년까지 30조엔에 이르는 고부가가치 창출을 목표로 범부처, 민관협력을 강조하고 있다. 일본재흥전략은 전

략적 목표를 일부 제조업에 국한하지 않고 인구감소에 대응한 사회 전체로 확대한다는 것이 그 특징이다.

1) 범부처 차원에서 4차 산업혁명에 대응

일본 정부는 제4차 산업혁명을 도전으로 간주하고 범부처 차원의 입체적이고 적극적인 대응전략 구상하여, 특히 4차 산업혁명 대비가 상대적으로 지체되었다는 지적과 함께 생산성 혁명을 위한 설비, 기술, 인재부문에 투자를 강조하고 있다. 환경 정비, 공공과 민간이 지향해야 할 방향성 공유를 위한 민관 대화, 벤처 창조의 선순환 등을 제시하고, 국제 혁신 벤처 창출 거점 형성을 위한 새로운 대학・대학원 창설, 실리콘밸리와 일본의 가교 프로젝트 등을 통해 적극적으로 4차 산업혁명에 대응할 것을 제안하고 있다.

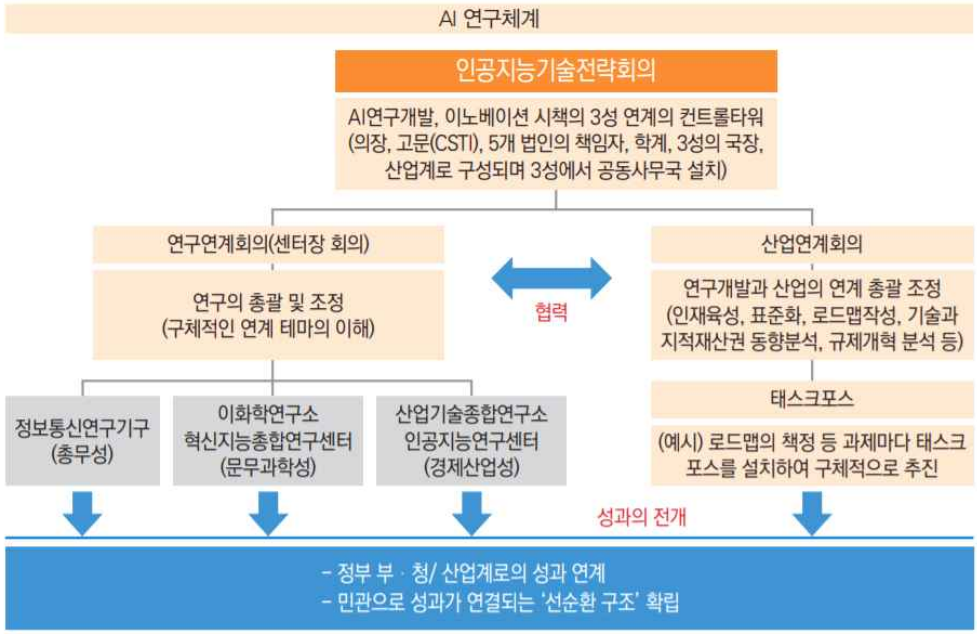
2) 대규모 국가프로젝트 시작으로 AI 제시

부처 간 장벽을 초월한 기술 인텔리전스 강화와 기술 동향을 고려한 중요 분야(기술기반인 AI, 로봇, 바이오, 에너지, 환경 기술 등의 융합연구)에서의 기술 전략을 수립하여 AI R&D와 이노베이션 시책의 관리를 총괄할 총무성, 문무과학성, 경제산업성 3성 연계의 사령탑인 인공지능기술전략회의를 설치하여 AI 연구개발 관리 및 로드맵 구축하여 다음과 같은 업무를 추진하고 있다.

<표 2-8> AI 연구개발을 위한 회의 및 조직

구분	상세 설명 및 역할
인공지능기술전략회의	산학연의 지식과 혜안 결집을 위해 칸막이를 배제하여 설치한 강력한 컨트롤 타워 역할 수행 * 월 1회 정도 개최되며 정기적으로 인공지능전략회의에 관련 내용을 보고(NEDO AI포탈)
연구연계회의	인공지능기술전략회의 산하에서 AI R&D를 종합·조정하는 역할 수행
산업연계회의	R&D와 산업의 연계조정을 통해 AI 기술과 R&D 성과의 사회적 탑재를 가속화하는 역할 수행 * 연구연계회의와 마찬가지로 월1회 정도 개최, 인공지능전략회의에 관련 내용 정기 보고(NEDO AI포탈)
이화학연구소	기업과 공동연구를 추진하며, 연구인재 100명을 선발하여 AI 핵심거점 연구기관인 혁신지능 통합연구센터(AIP센터)를 개설

[그림 2-1] 일본 정부의 인공지능 연구체계도



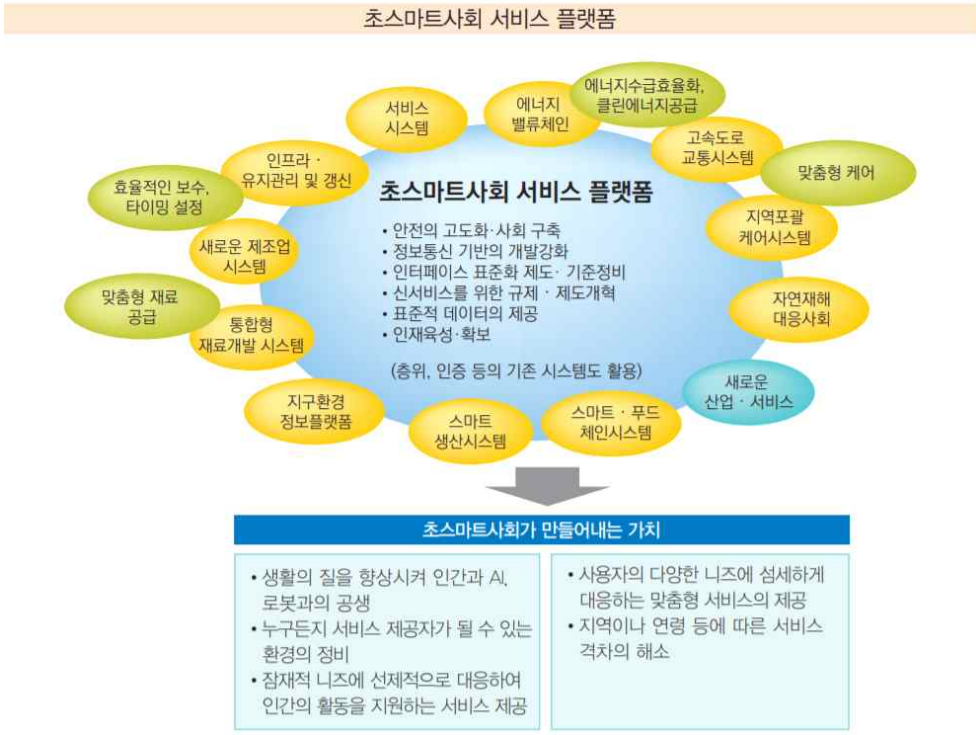
출처 : 일본의 인공지능 정책동향과 실행전략, 정보통신기획평가원, 2017

다. 제5기 과학기술 기본계획 - 과학기술 관점의 AI 정책

제5기 과학기술 기본계획은 과학기술 정책의 거시적 방향성을 민관협력을 통한 4차 산업혁명 대응으로 설정하고 있다. 세부적으로는 지속적인 성장과 지역사회의 자율적인 발전, 국가 및 국민의 안전·안심과 삶의 질 확보, 글로벌 규모 과제에 대한 대응 및 세계 발전에 기여, 지적 자산의 지속적인 창출을 목표로 하고 있다. 지식과 가치 창출 프로세스가 크게 변모하고 산업 구조가 급속히 변화하는 대변혁의 시대를 맞아 미래에 과감히 도전하는 R&D와 인력 강화를 제시하여 획기적이지만 리스크가 높은 연구에 대해서도 과감한 도전을 강조하고 있다.

선진국 간 ICT 기반의 미래사회 선도 경쟁이 치열해지고 있어 제4차 산업혁명에 대응하기 위한 민관 차원의 협력과 새로운 가치 창출을 강조하면서 일본 정부는 이러한 도전적인 R&D와 새로운 가치·서비스가 연속으로 창출되는 미래를 '초스마트 사회(society 5.0)'로 보고 강력하게 대응하고 있다.

[그림 2-2] 일본의 초스마트 사회 서비스 플랫폼



출처 : 일본의 인공지능 정책동향과 실행전략, 정보통신기획평가원, 2017

초스마트사회에서는 맞춤형 서비스 제공이나 잠재적 니즈 충족 등 삶의 질 향상을 위해 AI를 통한 분석이 요구되며, 대용량 데이터 처리를 위한 고성능 AI 전용 슈퍼컴퓨터 등의 '장치기술'이 필요하다. 이에 초스마트사회 서비스 플랫폼에 필요한 기반 기술과 新 가치 창출을 위한 핵심 강점 기술을 전략적으로 개발할 것을 제안하고 있다.

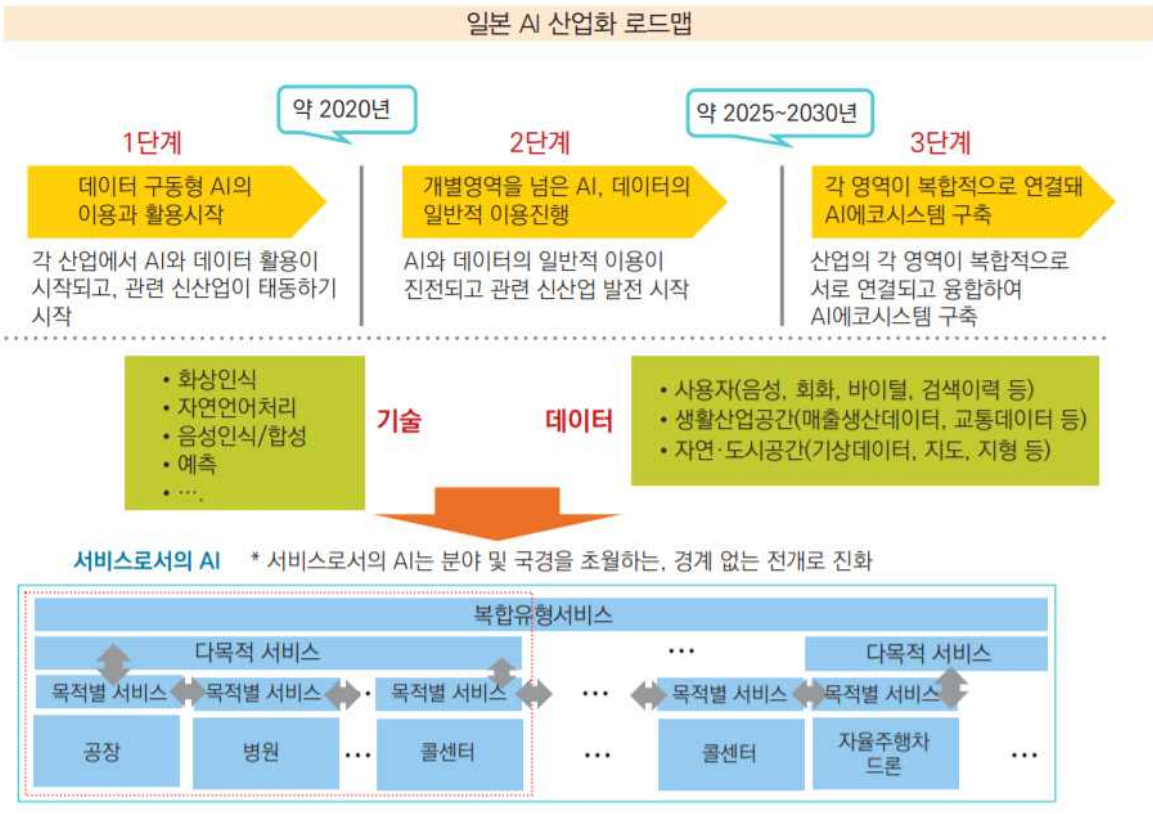
국내외 우수 인재의 참여를 통한 균형 있는 연구를 통해 AI 및 기반 기술의 강화가 가능하다고 보고 중·장기적인 관점에서 기초연구와 실용연구의 균형을 맞추기 위해 노력하고 있다. 또한, 세계적으로 우수한 인재, 지식, 자원을 도입하여 R&D 및 인재 육성을 진행하고 있다. 그리고 인문·사회과학 및 자연과학 연구자 간 연계를 통해 AI R&D의 발전이 가져올 미래 사회 변화에 대한 통찰력 확보도 강조하고 있다.

라. AI 산업화 로드맵 - AI 정책의 실행전략

인공지능기술전략회의는 일본 경제발전과 초스마트사회 실현을 위한 AI 정책 실행전략인 AI 연구 개발 목표 및 산업화 로드맵을 발표하였다(2017.02). 「일본 재흥전략」과 「제5기 과학기술기본계획」에서 제시했던 글로벌 선도를 위한 일본의 강점과 AI의 융합 필요성을 다시 확인하고, 일본과 세계가 직면한 사회문제에 대해 일본이 가진 강점과 AI의 결합을 통해 대응할 때 일본의 글로벌 경쟁력 확보가 가능하다는 것을 제시하고 있다.

AI 기술 수준 및 사회문제 관점에 따라 아래 그림과 같이 향후 AI 산업화를 3단계로 구분하고, 모든 영역이 경계 없이 융합함으로써 AI 생태계가 구축되는 초스마트사회를 목표로 설정하고 AI 기술을 서비스로 접근하여 사회의 사용자, 산업, 도시 공간과 같은 모든 영역에서 다양하게 활용할 수 있는 방안을 모색하고 있다.

[그림 2-3] 일본 AI 산업화 로드맵



출처 : 일본의 인공지능 정책동향과 실행전략, 정보통신기획평가원, 2017

6. 주요국의 SW 융합 (인공지능) R&D 정책 특징 정리

가. 개요

전 산업과 SW의 융합, 특히 인공지능 기술과의 융합은 제4차 산업혁명 시대를 대비하기 위해 필수적으로 요구되는 것으로 국외 주요 국가는 각자의 기술 수준 및 환경 등 당면한 상황에 맞춰 SW 융합 기술을 활용하여 자국의 문제를 해결하면서 제4차 산업혁명에 대비한 인공지능 기술을 선도적으로 확보하는데 맞춰져 있다. 이에 위에서 살펴본 내용을 토대로 주요국의 SW 융합 관련 AI 기술을 중심으로 R&D 전략의 특징들을 요약하고자 한다.

<표 2-9> 주요 국가의 AI 융합 전략 특징 (요약)

국가	주요 특징	중점 추진 분야
미국	- 산업계의 대응 개연성이 낮은 분야 - 고위험 도전적 과제 추진 - 윤리적/법적/사회적 대처 병행 - 필요 인력 확보 방안 병행	- 지각, 자동화된 추론/기획 - 인지시스템 - 머신러닝, 자연어 처리 - 로봇공학
독일 (유럽)	- 제조분야 글로벌 선진국 지위 유지 - 제조+인공지능 융합으로 신하이테크 전략 추진	- 추론시스템, 기계기반 처리 - 지식기반 시스템 - 패턴분선 및 인식 - 로봇공학 - 지능형 멀티모달
일본	- 강점분야(로봇, 제조업) 융합기반 AI R&D - 사회문제 해결 기술을 통하여 글로벌 경쟁력 확보	- 생산성 향상 - 건강(의료, 간호) 분야 - 공간이동 - 정보보안
중국	- 독일 Industry 4.0을 벤치마킹 - 제조업 고도화 방안에 AI를 접목	- 차세대 정보기술, 수치제어 및 로봇 - 항공우주, 해양공학, 철도교통 - 에너지, 전력설비 - 농기계 장비, 신소재, 의학 등

나. 미국의 AI R&D 전략 특징

미국은 구글, 아마존, 페이스북 등 민간분야의 AI 관련 글로벌 선두 기업들이 다수 존재하여 민간 중심으로 AI 기술 개발이 활발히 진행되고 있다. 따라서 미국 정부의 AI R&D 전략은 산업계에서 대응할 개연성이 낮으나 연방정부 투자로

편익을 확보할 수 있는 가능성이 높은 분야 중심으로 전략적 국가 AI R&D 투자 방향 수립과 전문 인력 확보를 위한 로드맵을 가지고 있다. 정부 부처, 산업계, 학계 등으로부터 다양한 요구사항을 수렴하여 국가가 전략적으로 집중 지원해야 하는 AI R&D 투자 방향을 도출하여 AI 활용으로 산업적, 사회적 편익이 예상되는 분야를 선별하여 공통 요소 기술 및 차세대 핵심기술을 도출하고 장·단기적 관점의 R&D 추진 방향을 제시하고 있다.

AI R&D 추진에 있어 고려해야 하는 법적·사회적 이슈 사항을 종합적으로 검토하여 시스템 설계, 인프라 구축 및 생태계 조성에 공통으로 반영하여 AI의 윤리적, 법적, 사회적 영향을 예측하고 확장성과 유연성을 고려, 보편적 규범에 부합하면서 신뢰할 수 있는 AI 시스템 설계 가이드라인을 마련하고 있다.

다. 독일의 AI R&D 전략 특징

독일은 지속적인 기술개발 전략을 연계하여 제조 분야에서 글로벌 경쟁력을 갖추고 강점을 유지해 오고 있다. 2014년 이후 제조 분야 하이테크 글로벌 선진국 확보와 유지를 지속하기 위해 AI 기술을 접목하여 일관된 혁신 정책 추진과 시스템 구축하여 범부처 차원의 전략 추진으로 다양한 주체에 오픈하여 문제 해결형의 사회 혁신으로 추진하고 있다.

그동안 추진해 온 하이테크 전략에 이어 2014년 이후 신하이테크 전략으로 연계하여 6대 우선 과제(디지털 경제와 사회, 지속적인 경제와 에너지, 혁신적 노동시장, 건강한 삶, 지능형 이동, 시민 안정)에 AI 기술을 접목하고 있다.

라. 일본의 AI R&D 전략

일본은 자신의 장점인 로봇, 제조업 융합에 기반을 두고 일본과 세계가 직면한 사회문제에 AI 기술을 결합하여 문제 해결과 더불어 글로벌 기술경쟁력을 확보한다는 전략을 제시하고 있다.

마. 중국의 AI R&D 전략

중국은 최근 급격한 경제 성장을 하고 있으나 상대적으로 떨어지는 기술 경쟁력 확보를 위해 독일의 Industry 4.0 모델을 참조하여 제조업을 고도화하는 방

향으로 제조 전 산업에 AI 기술을 결합하는 내용으로 중국제조 2025 전략을 수립하여 추격에 나서고 있다.

제3절 설명 가능한 인공지능(eXplainable AI, XAI)

1. 설명 가능한 인공지능의 개념

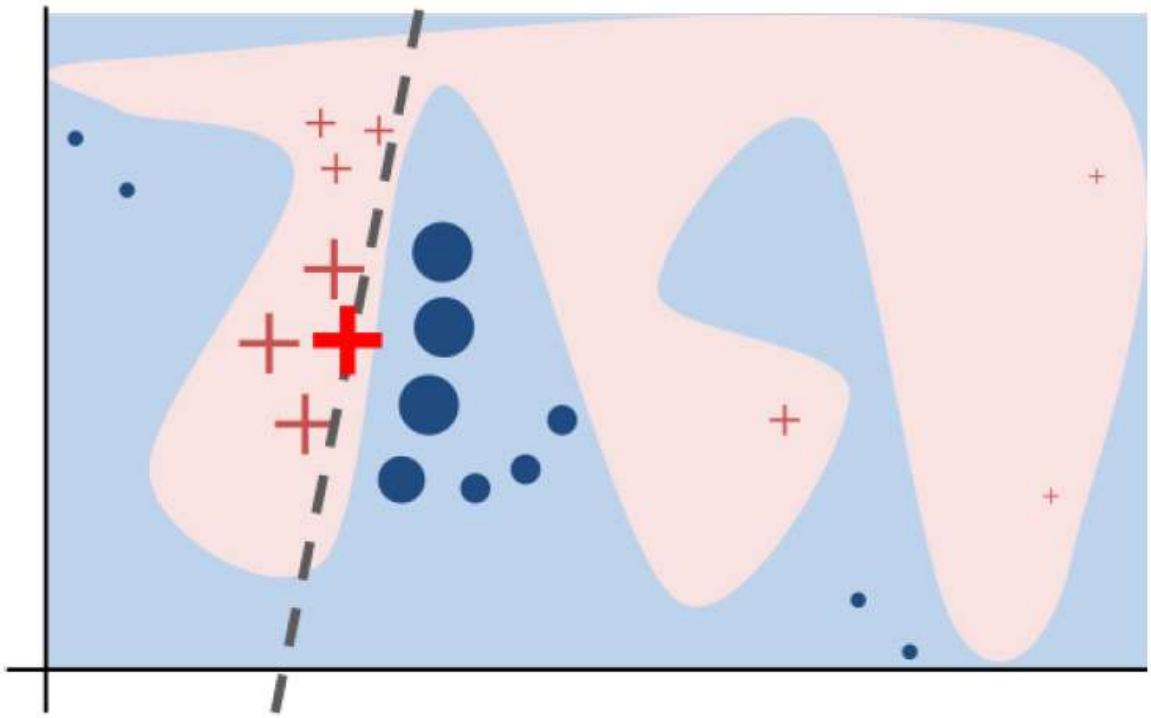
현재 인공지능 기술은 학술적 연구 단계를 뛰어넘어 각 산업 분야에 적용 가능한 수준으로 빠르게 발전하고 있으며, 현재 수준에서 시각 지능과 청각 지능에서 이미 인간의 수준을 초월한 것으로 알려져 있다. 또한, 2030년 이후에는 인간의 지능과 동등 또는 그 이상으로 진화할 것으로 전망되며, 향후 더 많은 영역에 활용될 수 있다. 그러나 금융보안원(2018)에 의하면, 현재의 인공지능 기술은 의사결정, 예측, 인지 등에 활용되는 경우, 블랙박스 형태로 결과 정보만 제시될 뿐 도출된 결과의 근거나 이유, 과정의 타당성 등은 설명할 수 없는 점이 한계로 지적되었다. 그러므로 인공지능 기술을 의료, 국방, 금융 분야 등 신뢰를 기반으로 하는 시스템에 적용할 경우, 효율성은 제고할 수 있지만 신뢰성과 공정성, 정확성을 담보하기는 어려운 것이 현실이다. 이에 따라 인공지능 기술의 결과에 대한 사용자 및 사회의 수용과 신뢰가 우려되면서 설명 가능한 인공지능(eXplainable AI, 이하 XAI)에 대한 관심이 높아지는 추세이다. 유럽연합(EU)은 2018년부터 개인정보 보호 규정인 General Data Protection Regulation(GDPR)을 통해 알고리즘에 의해 자동으로 결정된 사안에 대해서 정보 주체들이 어떻게 결과를 도출하는지에 대한 설명을 요구할 수 있다(Bryce Goodman, 2017). 또한, 2018년부터 미 국방성 산하 방위고등연구계획국(Defense Advanced Research Projects Agency, 이하 DARPA)에서는 2017년부터 XAI 학습 모델 개발 및 시험 관련 프로젝트를 추진하고 있다(DARPA, 2016).

2. LIME(Local Interpretable Model-agnostic Explanations) 알고리즘 개요

본 연구에서 개발한 기계학습 모델의 예측 결과에 대한 이유를 설명하기 위해 LIME(Local Interpretable Model-agnostic Explanations) 알고리즘을 활용하였다. LIME(Marco T. R., 2016)은 인공지능 예측 모델에 대한 결과를 신뢰할 수 있

는 방법으로 해석하는 알고리즘이다. LIME(Marco T. R., 2016)은 데이터 공간 전체에서 어떤 모형을 설명하는 것은 어렵지만 설명하고자 하는 데이터와 가까운 거리에 있는 국소적인 데이터 공간에서는 의미 있는 모형으로 설명할 수 있다는 것이 핵심 원리라고 볼 수 있다. 아래 그림에서 검은 점선은 국소 영역에서 신뢰할 수 있는 학습된 설명(explanation)이고 두꺼운 빨간 십자가와 파란 점은 설명하고자 하는 데이터이다. 전체 공간에서 살펴보면 검은 점선이 파란 점과 빨간 십자가를 모두 설명할 수 없지만, 국소적인 공간에서는 두 개를 구분할 수 있다.

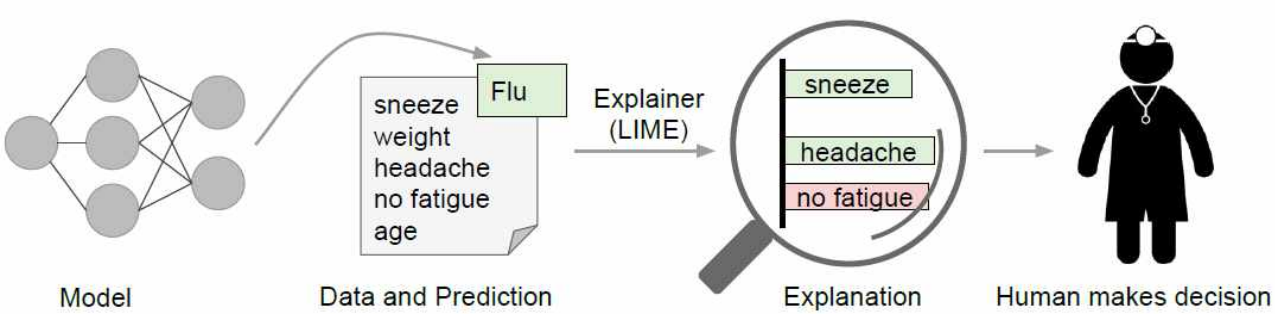
[그림 2-4] Local Fidelity 예제



출처 : Marco T. R. and Sameer S., “Why Should I Trust You?” Explaining the Predictions of Any Classifier, 2016.

LIME에서 예측 결과는 설명자(Explainer)에 의해 설명될 수 있으며, 모델에서 가장 중요한 몇 가지 증상을 설명자를 통해 강조하여 나타낼 수 있다. 아래 그림과 같이 LIME을 통해 해당 모델에 대한 이유가 의사에게 제공된다면, 의사는 인공지능 모델의 도움을 통해 더 나은 결정을 내릴 수 있다.

[그림 2-5] LIME 알고리즘의 활용 - 개별 예측 설명



출처 : Marco T. R. and Sameer S., "Why Should I Trust You?" Explaining the Predictions of Any Classifier, 2016.

종합하면, LIME은 입력 값을 조금 바꿨을 때 모델의 예측값이 크게 바뀐다면 해당 변수를 중요한 변수라고 판단하고 해당 입력 값이 미치는 영향에 대한 상대적 가중치를 제공하는 것이다. 본 연구에서는 개발한 분류 모델을 LIME 방법론을 활용하여 설명하고자 한다.

제3장 연구 프레임워크[7)]

제1절 연구 자료

1. 개요

본 연구에서 활용한 연구 자료는 국가과학기술지식정보서비스(National Science & Technology Information Service, 이하 NTIS)[8)]에서 제공하는 2016년 국가연구개발사업의 과제 정보이다. NTIS 과제 정보 중 과제 제목과 국가과학기술표준분류체계[9)], 연구 요약문에서는 연구 내용과 연구 목표, 총 과제비 등을 확보하였다. 그 중 분류 모델 개발을 위해 활용한 데이터는 2018년 연구[10)]에 의해 SW 융합 R&D 단계(총 5개 단계)가 구분된 2,000개의 과제 정보이다. 각 과제를 다섯 단계로 구분하기 위해 SW R&D 관련 산・학・연 전문가 30명으로 구성된 연구반을 통해 정성적 분석 방법론 중 하나인 델파이 조사를 수행하였다. 델파이 조사는 2018년 9월부터 12월까지 4개월간 총 3라운드에 걸쳐 표본 추출된 2,005개 과제 중 5개를 제외한 총 2,000개의 과제에서 합의가 이루어졌다.

하지만 학습 데이터에 해당하는 총 2천 개의 과제에서 SW 융합 R&D가 아니라고 구분되는 Low 단계가 총 1,681건을 차지하고 있어 전체 데이터의 약 84%의 비중을 차지하고 있어 학습을 위한 충분한 데이터로 판단하기는 어렵다. 그러므로 Low 단계에 속하지 않는 SW 융합 R&D 과제를 확보하기 위해 비중을 확보하기 위해 추가적으로 분류 작업을 수행하였다. 추가 분류할 데이터는 과기정통부의 「2019년도 정부 연구개발 투자방향 및 기준」에서 SW R&D로 구분하고 있는 국가표준분류체계의 “대분류-정보통신”, “중분류-소프트웨어, 정보보호, 정보이론”에 해당하는 1,124개의 과제를 추출하였고, 해당 과제를 SW R&D

---

7) 본 연구의 결과는 “기계학습을 활용한 SW 융합 R&D 자동 분류 및 현황 분석에 관한 연구, 서영희(2019)”의 내용을 일부 인용하여 추가 정리하였음

8) 사업, 과제, 인력, 연구시설장비, 성과 등 국가연구개발사업에 대한 정보를 한 곳에서 서비스하는 국가 R&D 정보 지식포털

9) 국가과학기술 표준분류체계는 과학기술기본법 제27조 및 동법 시행령 41조에 명시되어 있으며, 과학기술 분야에서 정보의 관리 및 유통, 인력 관리의 효율화, 연구개발사업의 효율적 기획・관리를 위한 국가 표준 분류 틀로, 연구 분야와 적용 분야의 독립적인 2차원 분류체계임

10) 2018년에 수행한 국가연구개발사업에서 SW융합 연구의 현황과 분석(소프트웨어정책연구소)의 Raw Data를 활용하였음

전문가에게 분류 작업을 요청하여 입력 자료를 확보하였다. 물론 기계학습 모델 개발을 위해서 입력 데이터가 많이 확보될수록, 예측력이 높아지고 정밀해지겠지만, SW 융합 R&D 단계에 대한 전문가 판단을 하기 위해 많은 비용과 노력이 발생된다는 점을 고려하여 총 3,124개의 과제를 입력 데이터로 활용하는 것으로 결정하였다.

총 5가지(High, Med High, Medium, Med Low, Low)의 SW 융합 R&D 단계가 예측 모델 개발을 위한 종속변수이며, 개별 R&D 과제는 아래 <표 3-1>의 판단 기준에 따라 한 개 단계에 할당되어 있다. High 단계로 분류된 과제는 연구개발 활동의 80~100% 정도가 SW를 연구개발하는 과제이고, Low 단계는 전체 연구개발 과정에서 SW 연구개발 활동이 20% 미만인 과제로, SW 융합 R&D가 아닌 비 SW 융합 R&D로 구분한다. 또한, SW 융합 R&D 활동이 20~40%인 경우는 Med Low로, 40~60% 이내인 경우는 Medium 단계, 60~80%에 속하면 Med High로 해당 비중에 따라 20%의 격차를 두고 각 단계로 구분한다.

<표 3-1> SW 융합 R&D 단계

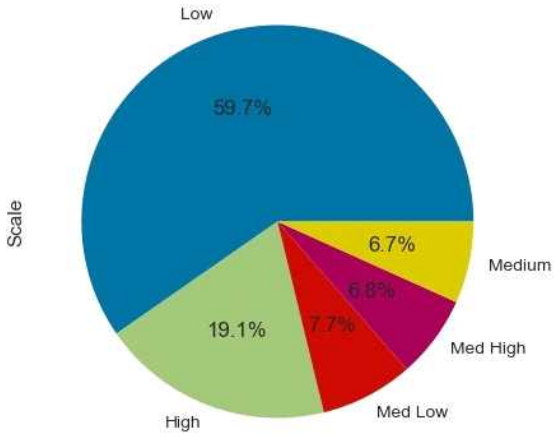
Scale	Description
High	• 연구개발 활동의 대부분이 SW를 연구개발하는 과제 (80% 이상 ~ 100%)
Med High	• 많은 부분이 SW를 연구개발하는 활동이며, 이와 함께 부분적으로 비SW의 개선·개발을 추진 (60% 이상 ~ 80% 미만)
Medium	• SW와 비SW를 동시에 연구개발을 추진 (40% 이상 ~ 60% 미만)
Med Low	• 부분적으로 SW를 연구개발하며, 대부분 새로운 비SW 연구개발을 추진 (20% 이상 ~ 40% 미만)
Low	• 기존 SW를 단순 활용하고, 대부분 새로운 비SW의 연구개발을 추진 (0% 이상 ~ 20% 미만)

본 연구에서 SW 융합 R&D 과제의 예측을 위한 분류 모델을 개발하기 위해 활용한 입력 데이터는 텍스트 형태의 연구 요약문의 연구 내용과 연구 목표이다. 해당 텍스트는 한글로 작성되어 있어 형태소 분석과 품질의 속도 성능이 영문에 비해 낮다는 한계점이 있다. SW 융합 R&D 분류 모델을 개발하기 위해 수없이 많은 실험이 필요하기 때문에 한글 자연어 처리 보다는 영문으로 번역하여 실험을 수행하는 것이 효율적이다. 또한, 영문 데이터를 학습 데이터로 사용하게 되면, 국외에서 제공하는 국가연구개발과제에 대한 정보를 수집하여 개발된 분류 모델을 적용하여 구분한 뒤, 국내 국가연구개발과제와 비교할 수 있다는 장점이

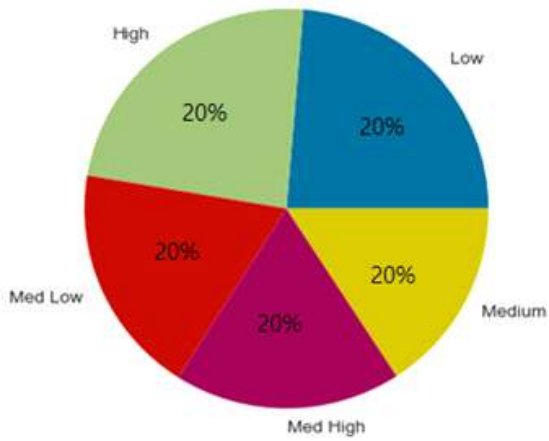
있다. 현재 국립과학재단((National Science Foundation, 이하 NSF), Horizon 2020은 홈페이지를 통해 과제 제목과 과제의 요약문을 제공하고 있기 때문에 국가간 비교 분석이 가능하다. 추후 개발된 분류 모델을 활용하여 국외 연구개발 과제에 적용하여 추이를 분석하고 국내 연구개발 과제와 비교하는 것은 향후 정책 개발을 위한 유의미한 분석이라고 판단된다. 이에 본 연구에서는 텍스트 분석을 영문으로 번역하여 진행하는 것으로 결정하였다.

학습 데이터의 단계별 구성을 살펴보면, 총 5단계 중 약 60%가 Low 단계가 약 60%를 차지하고 있기 때문에 모델 전체에 큰 영향을 미칠 수 있는 가능성이 존재한다. 그러므로 각 단계 중 데이터가 가장 적은 Scale을 기준으로 단계별로 동일한 개수의 데이터가 추출될 수 있도록 다운 샘플링을 수행하여 비중의 조정 전후의 결과를 비교해보고자 한다.

[그림 3-1] 비중 조정 전 학습 Data의 단계별 구성



[그림 3-2] 비중 조정 후 학습 Data의 단계별 구성



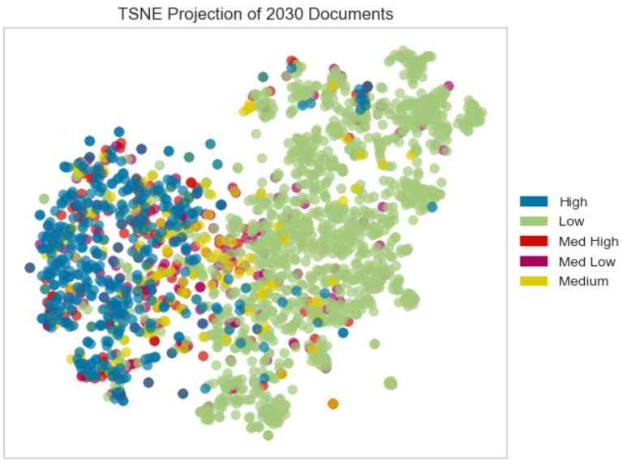
2. 학습데이터의 분포 및 비중 분석

본 연구에서는 분류 모델의 개발을 위해 전체 데이터의 65%에 해당하는 데이터를 학습용(training) 데이터로 활용하였으며, 나머지 35%는 검증용(validation) 데이터로 사용하였다. 총 3,124개의 과제를 학습 데이터로 사용하였으므로 학습용 데이터는 2,030개고, 검증용 데이터는 1,094개의 과제로 구성된다. 반복 무작위 부표본 검증(repeated random sub-sampling validation)을 사용하여 학습용 데

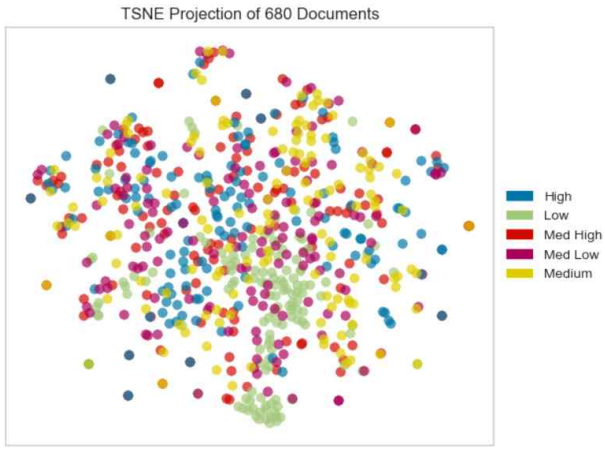
이터를 추출하였으며, 매 실험마다 다른 표본이 추출될 수 있도록 설정하였다. 모델의 실험은 각각 10번씩 수행한 뒤 평균값을 구해 기재하였다.

분류 모델을 개발하기 전에 가지고 있는 학습 데이터에 대한 형태와 분포를 미리 알기 위해 t-SNE(T-distributed Stochastic Neighbor Embedding, t-분포 확률적 임베딩) 알고리즘을 활용하여 벡터 시각화를 수행하였다. 아래 [그림 3-3]은 총 5단계 구분에서 학습 데이터에 해당하는 2,030개(65%)의 데이터를 벡터화하여 과제 정보와 단계별 상관관계를 도식화하였으며, [그림 3-4]는 5단계 구분에서 데이터 수가 가장 적은 단계의 과제 수에 맞추어 비중을 조정한 뒤 데이터의 상관관계를 도식화 한 것이다. 두 그림에서 알 수 있듯이 중간 단계에 속하는 Med High와 Medium, Med Low 단계는 데이터가 뒤섞여 있어 구분하기가 매우 어렵다. 특히 단계별로 비중을 조정하기 전은 Low 단계에 해당하는 자료가 많기 때문에 비교적 High와 Low 단계의 구분이 명확하지만 비중을 조정한 뒤에는 High와 Low 단계도 구분하기가 어려운 것을 알 수 있다.

[그림 3-3] Scale별 분포(비중 조정 전)



[그림 3-4] Scale별 분포(비중 조정 후)



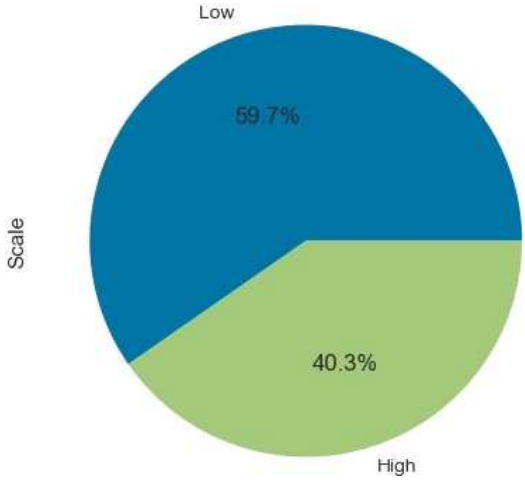
분류 모델의 개발을 위해 본 연구에서는 학습 데이터를 세 가지 형태로 구분하여 각각 수행하였다. 단계별 데이터의 구분 및 분포는 아래와 같다.

① 2단계(High, Low) 구분 및 분포

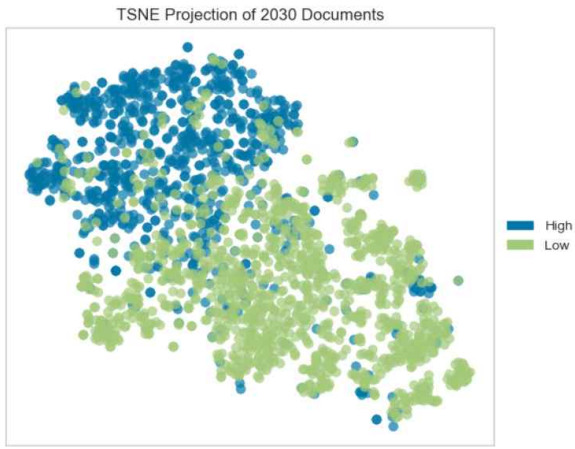
2단계 구분은 SW 융합 R&D 비중이 80%가 넘는 High 단계와 Med High 단계(60~80%), Medium(40~60%), Med Low(20~40%)를 모두 합쳐서 High 단계로 구

분하는 것이다. Low 단계는 20% 미만이 SW 융합 R&D 비중을 차지하는 것으로 비 SW 융합 R&D로 판단하고, 20% 이상 SW R&D 활동을 포함하면 SW 융합 R&D로 간주한다. 이러한 구분을 통해 정책입안자는 과제별 SW 융합 R&D 여부에 따라 유형별 현황을 파악하고 향후 발전 방향을 도출할 수 있다.

[그림 3-5] 2단계 학습 데이터 구성



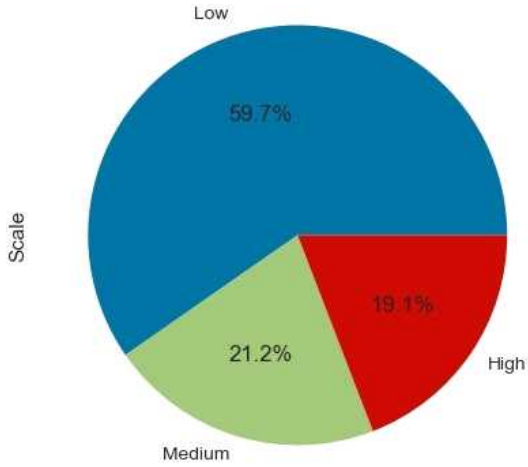
[그림 3-6] 2단계 과제 분포



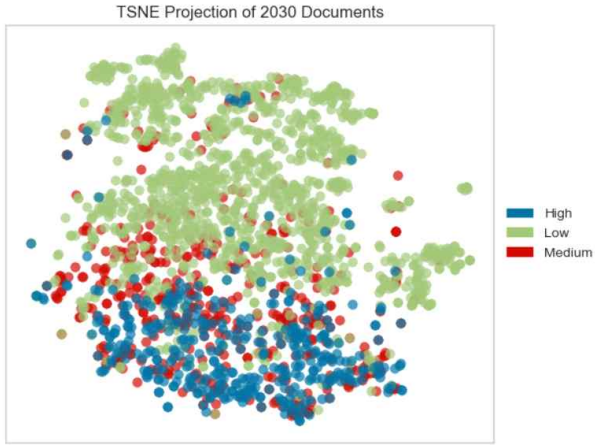
② 3단계(High, Medium, Low) 구분 및 분포

3단계는 Med High, Medium과 Med Low(이후 Med*) 단계를 묶어서 Medium 단계로 본다. 2018년에 수행한 SW R&D 전문가 델파이 분석 시에도 Med High와 Medium, Med Low 단계는 동의가 원활히 수행되지는 않은 단계로 [그림 3-3]과 같이 Med* 단계의 자료가 섞여 있기 때문에 단계 구분이 어렵다는 것을 알 수 있다. 그러므로 중간 단계(Med*)를 하나의 구간으로 간주하여 SW 융합 R&D가 20~80%에 해당하는 과제를 SW 연구개발 활동이 일부 포함되는 것으로 파악하고, 80~100%에 해당하는 High 단계는 SW 융합이 매우 높은 단계로, 0~20%에 해당하는 Low 단계는 SW 융합이 이루어지지 않는 비 SW R&D로 간주하여 총 3가지 형태로 SW 융합 R&D 현황을 분석하는 것이다. 이를 통해 학습 데이터가 가지는 한계점을 극복할 수 있고, 단계를 상/중/하 형태로 구분하여 각 유형에 맞는 SW 융합 R&D 전략을 마련하는 경우에 이러한 구분이 유의미하게 활용될 수 있다고 판단된다.

[그림 3-7] 3단계 학습 데이터 구성



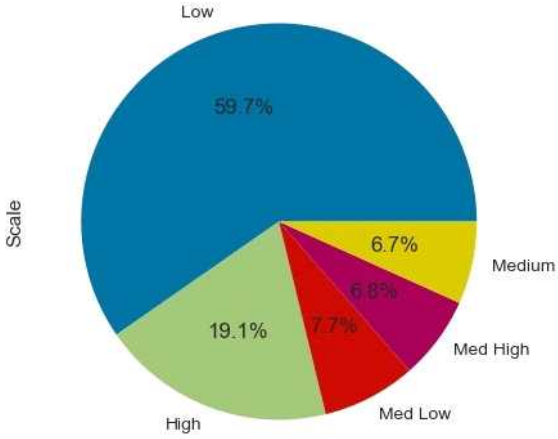
[그림 3-8] 3단계 과제 분포



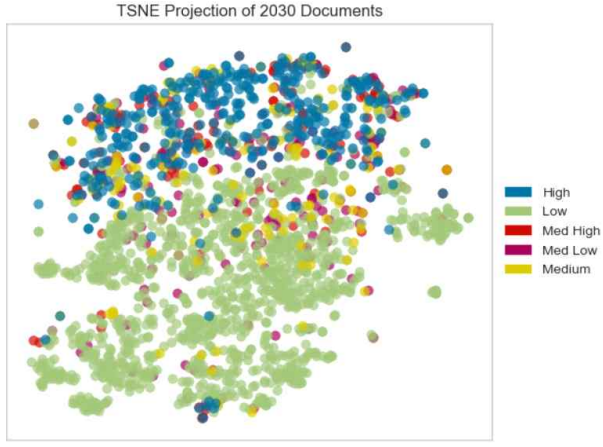
③ 5단계(High, Med High, Medium, Med Low, Low) 구분 및 분포

5단계 구분의 경우, <표 3-1>의 2018년 연구 조사 내용에서 분류한 SW 융합 R&D 구분 기준을 그대로 유지하는 것이다. 이는 SW 융합 R&D 현황을 보다 구체적으로 나누어 단계별 정책 방향을 수립하고자 하는 경우에 유의미하다고 볼 수 있다.

[그림 3-9] 5단계 과제 구분



[그림 3-10] 5단계 과제 분포



제2절 연구 진행 방법

1. 개요

본 연구는 연구 진행 과정은 [그림 3-11]과 같이 총 5단계로 구성된다.

① (1단계) 정량적으로 SW 융합 R&D 현황을 분석하기 위해 기존에 수행된 연구에 대한 정리 및 SW 융합 R&D 기준에 대해 알아보고, 국외 SW 융합 R&D 정책에 대한 현황을 정리한다.

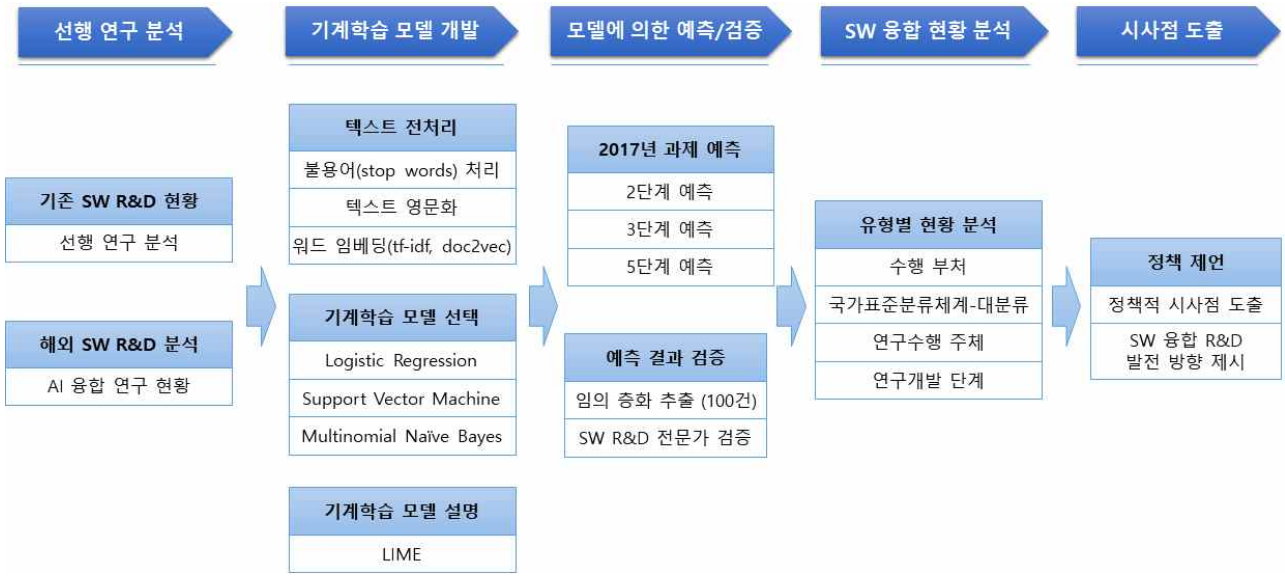
② (2단계) SW 융합 R&D의 자동 분류를 위한 기계학습 모델을 개발하기 위해 학습 데이터인 국가연구개발과제의 연구 요약문 관련 텍스트에 대한 전처리 과정을 거쳐 세 가지의 기계학습 모델에 대한 실험을 통해 모델을 선택하는 단계이다.

③ (3단계) 선택된 분류 모델을 활용하여 2017년 국가연구개발과제를 대상으로 2/3/5단계 형태로 SW 융합 R&D에 대한 예측을 수행한다. 예측 결과에 대한 신뢰도 검증을 위해 100개의 표본 과제를 추출하여 전문가에게 분류를 요청하여 모델의 예측 결과와 전문가의 예측 결과를 비교를 수행한다.

④ (4단계) 2017년 국가연구개발과제의 예측 결과를 토대로 수행부처, 국가표준분류체계(대분류), 수행 주체, 연구개발 단계 등의 유형별 SW 융합 R&D 현황 분석을 위해 시각화 도구를 활용하여 설명한다.

⑤ (5단계) 3단계의 SW 융합 R&D 현황을 토대로 향후 SW 융합 활성화를 위한 정책적 시사점 및 발전 방향을 도출한다.

[그림 3-11] 연구 진행 단계



2. 기계학습을 활용한 분류 모델 개발

2.1 텍스트 전처리

1) 불용어 처리

텍스트 전처리를 위해 전치사나 관사 등 너무 빈번하게 등장하는 단어와 같이 문장이나 문서의 특징을 표현하는데 불필요한 단어인 불용어(stopwords)를 미리 제거하는 작업을 수행하였다. 이를 위해 python 패키지의 stop_words.py[11]을 활용하였으며, 이는 Glasgow Information Retrieval Group[12]에서 제공하는 것으로 알려져 있다.

2) 영문 번역

제1절의 연구 자료에서 설명한 것과 같이 연구 요약문의 연구 목표, 연구 내용, 한글 키워드의 조합을 효율성과 국외 비교 분석을 위해 영문으로 번역하여 활용하기로 결정하였다. 영문 번역은 클라우드 서비스로 제공되는 네이버 파파고

11) Lib\site-packages\sklearn\feature_extraction\stop_words.py
12) 영국 글래스고 대학의 컴퓨터 과학 학교 내에 있는 정보 검색 그룹으로 Information Retrieval(IR) 영역에서 두각을 나타내며, 데이터 스트림을 위한 기계학습과 심층 학습 기술 관련 기술 연구에 앞장서 왔음(https://www.gla.ac.uk/schools/computing/research/researchsections/ida-section/informationretrieval/)

(Papago) 번역 서비스를 사용하였다. 실제로 번역된 데이터는 약 3백만여 문자로 구성된 데이터와 6천만여 문자로 구성된 한글 데이터이고, 번역을 진행하는 데는 약 10시간 정도가 소요되었다. 3백만여 문자로 구성된 데이터는 2016년 국가연구개발과제 중 3,124개를 추출하여 전문가 델파이 조사에 따라 라벨링이 된 과제 데이터이며, 6천 만여 문자로 구성된 데이터는 분류가 되지 않은 2017년 국가연구개발과제의 텍스트이다.

본 연구에서는 분류 결과값을 알고 있는 2016년도 연구 자료를 기반으로 지도학습 방법을 사용하여 SW 융합 R&D 분류 모델을 만들고 개발된 자동 분류 모델을 활용하여 2017년도 국가연구개발과제를 분류한다. 이를 위해 두 데이터를 통합하여 하나의 데이터 집합을 만들어 처리하는 것이 간편하므로 이를 처리하기 위한 프로그램(combine.py)을 별도로 개발하였다. 한편, 번역된 데이터는 무의미한 공백, 구두점, 특수문자, 단/복수나 시제 등의 변화된 단어들을 포함하고 있기 때문에 이 부분을 사전에 처리하는 프로그램(cleanse.py)을 개발하였다.

3) 텍스트 마이닝 방법

① tf-idf

tf-idf는 텍스트 마이닝에서 중요 단어의 추출 시에 대표적으로 활용되는 방법 중에 하나로써 텍스트 내에 특정 단어의 상대적 중요도를 계산하는 통계적인 수치이다. 단어 빈도수인 tf(term frequency)는 문서 내에서 특정 단어가 등장하는 빈도수를 나타낸다. 특정 단어의 빈도수가 높다면 해당 단어가 문서 내에서 중요하다고 생각할 수 있다. idf(inverse document frequency)는 해당 단어의 중요도를 계산하는 것으로 어떠한 단어가 다수의 문서에서 등장한다면 idf 수치는 낮아진다. tf-idf는 tf와 idf를 곱한 수치로 tf-idf를 통해 해당 단어가 갖는 중요도를 알 수 있다(정근하, 2010). 즉, tf-idf 값이 높은 키워드는 문서 분류에서 중요한 역할을 한다고 볼 수 있다.

[수식 2-1] tf-idf

$$tf-idf(t,d,D) = tf(t,d) \times idf(t,D)$$

tf(t,d) : 문서 d 내 특정 단어 t의 빈도수

D : 전체 문서 집합

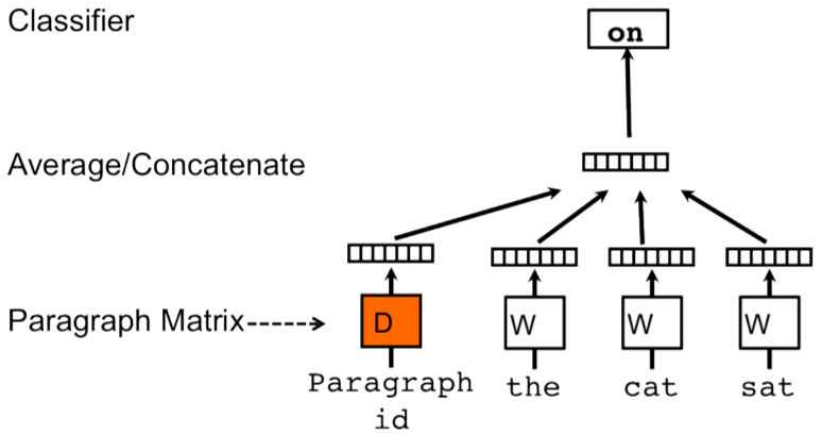
$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

|D| : 전체 문서 집합의 크기

② doc2vec

단어를 벡터로 나타내는 방법론인 word2vec을 응용한 doc2vec은 문서를 벡터로 표현하는 방법론이다. word2vec은 비슷한 의미를 가진 단어를 비슷한 값의 벡터로 임베딩하는 특징을 보인다. word2vec이 문맥 단위의 단어의 연쇄를 이해하는 모델이라면 doc2vec은 하나의 문맥에서 등장하는 단어들의 분포적 특성을 추출하고 이를 여러 문맥을 통해서 특징들을 응축한다고 볼 수 있다(김동성, 2019). doc2vec(Quoc Le, 2014)은 문장, 문단 혹은 문서의 특징을 표현하는 방법으로 아래 모델에서는 세 단어의 맥락을 가진 벡터의 연결(Concatenation) 또는 평균을 사용하여 단락 D를 예측한다. 아래 [그림 3-12]는 PV-DM(Distributed Memory Model of Paragraph Vector) 방식을 보여주고 있다. 여기서 Paragraph id는 현재 문맥에서 누락된 정보로써 현재 문맥에서 빠지거나 단락의 주제를 기억해주는 메모리와 같은 역할을 한다.

[그림 3-12] 단락 벡터의 학습 과정



출처 : Quoc Le. and Tomas Mikolov, Distributed Representations of Sentences and Documents, 2014

2.2 기계학습 모델 선택

텍스트 전처리 과정 이후의 데이터는 분류 모델의 개발을 위해 텍스트 분류 방법을 활용해야 한다. 본 연구에서는 SW 융합 R&D 단계를 구분하기 위해 텍스트 분류에서 많이 활용되고 있는 3가지 모델을 모두 실험하여 그 결과를 비교한다.

1) 로지스틱 회귀(Logistic Regression)

로지스틱 회귀는 종속 변수가 0이나 1인 이진형 변수에서 자주 사용하는 분류 방식으로 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는 방법론이다. 로지스틱 함수의 입력 값은 독립 변수와 계수로 이루어진 회귀식이 사용되며, 확률 값을 계산하여 특정 범주로 분류할 수 있다. 본 연구에서 활용한 방식은 Scikit-learn[13] 라이브러리에서 제공하는 로지스틱 회귀 분류 방식을 사용하였다.

2) Support Vector Machine(SVM)

통계학자인 Vapnik이 개발한 분류기법인 Support Vector Machine(이하 SVM)은 입력 공간과 관련된 비선형 문제를 고차원 공간의 선형문제로 사상(projection)시켜 나타내기 때문에 수학적으로 분석하는 것이 수월하다고 알려져 있다(Cortes and Vapnik, 1995. Hearst et al., 1998). SVM은 지도 학습 방법을 사용하여 정답이 있는 학습 데이터를 통해 클래스를 분류해주는 최적의 분리경계면을 찾아낸다. 최적의 분리경계면이란 분리경계면에 가장 근접한 데이터(support vector)와 분리경계면 사이의 거리(margin)가 가장 큰 분리경계면을 말한다(김승희, 2015). SVM은 학습 데이터를 차원이 높은 공간으로 사상시켜 특징 공간 내에 선형으로 분리 가능한 입력 데이터 집합을 만들기 위해 전이 함수인 커널 함수(kernel function)가 필요하다. Linear, RBF(Radical Basis Function) 등 여러 가지 커널이 존재하며, 일반적으로 실험을 통해 주어진 문제와 학습 데이터에 적합한 함수를 찾아낸다.

3) 다항 나이브 베이즈 (Multinomial Naive Bayes, 이하 Multinomial NB)

13) Python 프로그래밍 언어를 위한 무료 기계 학습 라이브러리(http://scikit-learn.org/)

나이브 베이즈 분류기(Naive Bayes Classifier)는 신뢰도가 높고 효율적인 문서 분류기로 널리 사용되는 기분 문서 분류 방법이다. 나이브 베이즈 정리를 이용하여 개발된 알고리즘으로 텍스트 분류에서 결정 트리(decision-tree) 학습이나 신경망과 유사한 성능을 보이며, 데이터의 양이 많을수록 정확도가 높아진다는 특징이 있다(이상기, 2010). 특히, 다항 나이브 베이즈 방식은 합리적인 예측 성과와 효율성을 제공하며 구현하기도 쉽기 때문에 워드 카운트(단어 빈도) 기반의 문서 분류에서 널리 사용된다(Jiang Su 외, 2011).

<표 3-2> 기계학습 모델에 대한 설명 (요약)

기계학습 모델	설명
로지스틱 회귀 (Logistic Regression)	종속 변수가 0이나 1인 이진형 변수에서 자주 사용하는 분류 방식
Support Vector Machine(SVM)	입력 공간과 관련된 비선형 문제를 고차원 공간의 선형문제로 사상(projection)시켜 나타냄
다항 나이브 베이즈 (Multinomail NB)	신뢰도가 높고 효율적인 문서 분류기로 널리 사용되는 기분 문서 분류 방법

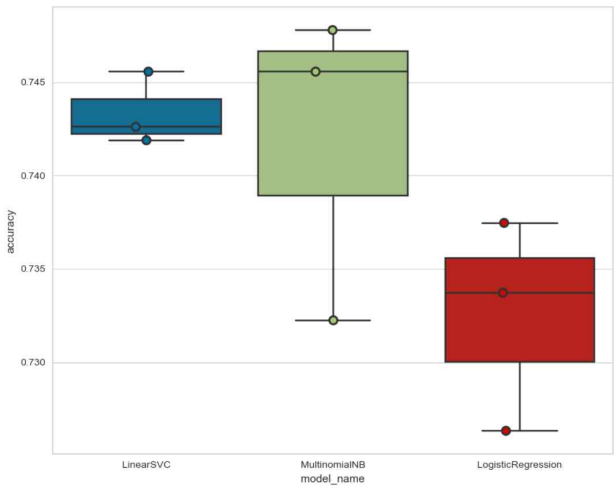
제4장 기계학습 모델 검증 결과

제1절 모델 예측 및 모델별 결과 단계

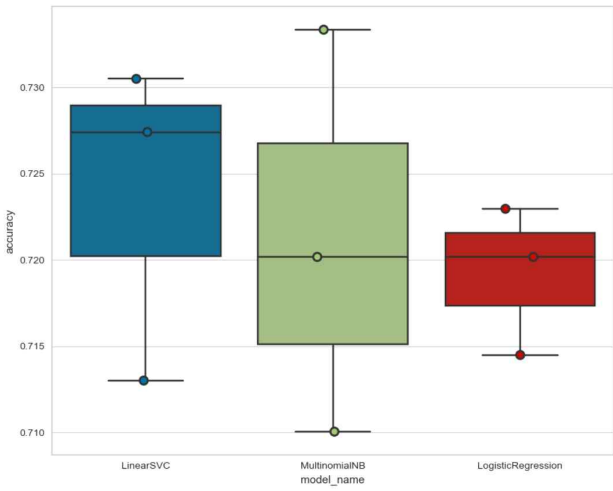
본 연구에서는 SW 융합 R&D 단계의 구분을 위해 2가지의 텍스트 마이닝 기법과 3가지의 모델을 실험한다. 텍스트 마이닝 기법은 tf-idf와 doc2vec이고, 모델은 Logistic Regression, SVM, Multinomial NB로 총 10회의 실험을 수행하여 실험 과정에서 정확도(Accuracy)와 정밀도(Precision)가 큰 변동을 보이지 않는 모델을 선정한다. SVM은 선형함수(Linear function)과 가우시안 RBF 함수(Gaussian radial basis function)를 모두 실험하였으며, 본 연구의 학습 데이터는 가우시안 RBF보다 선형함수가 정확도가 더 높기 때문에 향후 SVM을 Linear SVM으로 표기한다.

학습 데이터에 대해 모델별 성능을 시각적으로 표현하기 위해 seaborn 패키지에서 제공하는 상자 그림(box plot)으로 [그림 4-1]과 [그림 4-2]와 같이 도식화하였다. 상자 그림의 가로선은 중간값(Median)을 나타내며, 네모 박스는 25 ~ 75% 범위의 데이터를 보여준다. 일반적으로 위 범위를 벗어난 데이터는 이상치(outlier)로 볼 수 있다.

[그림 4-1] 모델별 정확도 실험 결과 (1차)



[그림 4-2] 모델별 정확도 실험 결과 (2차)



기계학습 모델 중에 가장 높은 성능을 보이는 모델을 찾는 함수를 구현하여

총 10회의 실험을 한 결과, Multinomial NB 모델이 총 8회 선택되었고 2번은 Linear SVM이 선택되었다. 예측력에 대해 결과적으로 중간값을 기준으로 Linear SVM, Multinomial NB, Logistic Regression이 유사한 성능 특성을 보이고 있다. 그러나 분류 모델의 선택은 성능 특성도 중요하지만, 학습 데이터의 특성에 대한 고려도 필요하다. doc2vec을 벡터화 알고리즘으로 선택할 경우 결과 벡터에는 음수 값이 포함될 수 있으나 Scikit-Learn이 제공하는 Multinomial NB 구현은 음수 값을 처리할 수 없다. 물론 음수를 보정하여 0 이상의 값을 가지도록 조정할 수 있으나 이는 데이터의 왜곡을 불러올 가능성이 있다. 그러므로 위 사항들을 고려하여 본 연구에서는 Linear SVM을 기본 모델로 활용하는 것으로 결정하였다.

정밀한 실험을 위해 제안한 3가지 모델의 성능을 정확도(Accuracy)와 정밀도(Precision), 재현율(Recall)을 활용하여 측정하고 결과를 비교하였다. 정밀도와 재현율의 조화 평균값이 F1 점수(F1-score)이다. 각 수치는 아래의 수식을 통하여 계산되며, 예측 단계별로 산출된다.

[수식 4-1] 정확도, 정밀도, 재현율, F1-score

$$\text{정확도(Accuracy)} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{정밀도(Precision)} = \frac{TP}{TP+FP}$$

$$\text{재현율(Recall)} = \frac{TP}{TP+FN}$$

$$\text{F1-score} = 2\frac{\text{Precision}\times Recall}{\text{Precision}+Recall}$$

“A 단계”에 대한 예측 결과를 평가할 때 <표 4-1>과 같이 TP(True Positive)는 A 단계라고 실제 A 단계인 과제를 A 단계로 예측한 것을 의미한다. TN(True Negative)은 실제 A 단계가 아닌 결과를 A 단계가 아니라고 예측한 것이다. FN(False Negative)은 실제 A 단계인데 예측은 A 단계가 아니라고 예측한 것이고, FP(False Positive)는 실제 A 단계가 아닌데 A 단계로 예측한 것을 의미한다. 분류 모델에 대한 성능 평가는 10회에 걸쳐 반복 무작위 부표본 검증(repeated random sub-sampling validation)을 수행하여 그 평균을 활용하였다.

<표 4-1> A 단계에 대한 예측 결과 평가

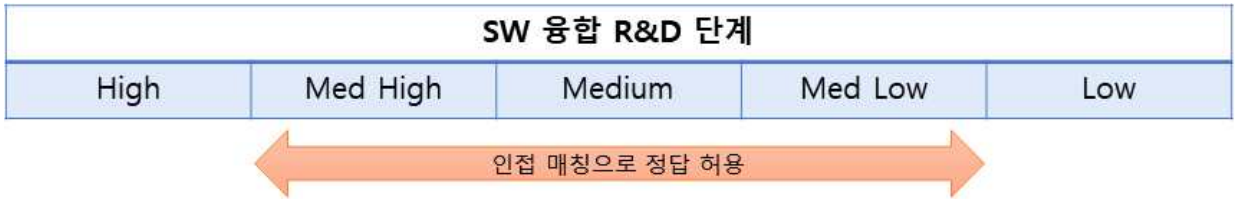
		예측	
		A 단계	**A 단계 X**
실제	**A 단계**	True Positive (TP)	False Negative (FN)
	A 단계 X	False Positive (FP)	True Negative (TN)

정확도는 3가지 방법으로 측정하였다. 가장 먼저 정 매칭(Accurate Matching)은 현재 5단계로 나뉘어 있는 단계(High, Med High, Medium, Med Low, Low)의 데이터를 그대로 활용하고, 결과 값도 5단계로 구분하는 것이다.

두 번째는 결합 매칭(Combined Matching)으로 결과를 총 3가지의 단계(High, Medium, Low)로 구분하는 것이다. High 단계와 Low 단계는 원래의 단계를 그대로 유지하며, Medium 단계는 Med High, Medium, Med Low 단계를 포함한다. 이렇게 결과를 3단계로만 구분하려는 이유는 Med*단계의 데이터가 SW R&D 전문가의 델파이 조사 진행 시, 의견이 일치되지 않은 구간이며, [그림 3-3]과 같이 과제가 뒤섞여 있어서 단계별 구분이 어렵기 때문이다. 즉, Med* 단계를 하나의 구간으로 판단하여 과제에서 SW 융합 연구개발 활동을 일부 포함하는 것으로 파악하고, High 단계(80~100%)는 SW 융합이 매우 높고, Low 단계(0~20%)는 SW 융합이 일어나지 않은 비 SW R&D로 파악하는 것도 정책적 함의를 도출하는데 유의미할 것으로 판단하였다.

세 번째로 [그림 4-3]과 같이 인접 매칭(Adjacent match)은 실제 결과 단계와 한 단계가 근접한 결과라면 맞는 것으로 허용하는 것이다. 예를 들면, 실제 Medium 단계인 과제에 대해 예측한 결과가 Med High 단계이거나 Med Low 단계이더라도 1개 단계만 떨어진 인접한 단계이기 때문에 예측이 맞은 것으로 간주한다.

[그림 4-3] 인접 매칭의 예시 (학습 데이터의 정답이 Medium인 경우)



제2절 텍스트 마이닝 결과

본 연구에서 개발하는 자동 분류 모델의 입력 값은 연구의 정보를 담은 텍스트로써 기계학습을 하기 위해서는 먼저 텍스트의 벡터화가 필요하다. 이를 위해 tf-idf와 doc2vec이라는 알고리즘을 모두 실험하였으며, 본 절에서는 두 알고리즘의 실험 결과를 비교하고자 한다. tf-idf는 여러 문서로 이루어진 문서군이 있을 때, 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 계산하는 통계적인 수치로써 비교적으로 간단하면서 효율적인 방법으로 알려져 있다.

doc2vec의 경우, 문서를 벡터로 변경하는 문서 임베딩 방식으로써 입력 값의 유사도를 고려하여 벡터를 추론할 수 있는 장점을 가지고 있기 때문에 장기적 관점에서 해당 분류 모델을 활용하고자 할 때 유용하다고 볼 수 있다. 그 이유는 향후 지속적으로 국가 전반의 연구개발사업에서 SW 융합 R&D 단계를 파악하고자 하는 경우, 미래 시점에 현재의 학습 데이터인 2016년 기준에서 미처 학습하지 못한 단어가 존재할 수 있기 때문이다. doc2vec은 이와 같은 미등록단어(Out of Vocabulary, OOV) 이슈를 해결하기 위한 대안으로써 실험하여 tf-idf의 결과와 비교 분석을 수행하고자 한다.

<표 4-2> tf-idf/doc2vec의 학습 결과 비교(5단계 & Linear SVM 기준)

구분	단계 (Scale)	정밀도 (Precision)	재현율 (Recall)	f1-score
tf-idf	High	0.59	0.76	0.67
	Med High	0.17	0.06	0.09
	Medium	0.18	0.07	0.09
	Med Low	0.22	0.06	0.09
	Low	0.83	0.95	0.89
	정확도 (Accuracy)	Accurate matches: 72.78% Combined matches: 75.31% Adjacent matches: 84.50%		
doc2vec	High	0.57	0.75	0.64
	Med High	0.21	0.09	0.12
	Medium	0.10	0.05	0.07
	Med Low	0.07	0.01	0.03
	Low	0.82	0.94	0.88
	정확도 (Accuracy)	Accurate matches: 70.11% Combined matches: 73.19% Adjacent matches: 82.57%		

<표 4-3> 단계별 정밀도, 재현율, 정확도 비교 (doc2vec & Linear SVM 기준)

구분	단계 (Scale)	정밀도 (Precision)	재현율 (Recall)	f1-score
2단계	High	0.83	0.82	0.83
	Low	0.89	0.89	0.89
	정확도 (Accuracy)	Accurate matches: 86.17% Combined matches: 86.17% Adjacent matches: 100%		
3단계	High	0.64	0.68	0.66
	Medium	0.52	0.39	0.44
	Low	0.86	0.92	0.89
	정확도 (Accuracy)	Accurate matches: 76.11% Combined matches: 76.11% Adjacent matches: 96.83%		
5단계	High	0.57	0.75	0.64
	Med High	0.21	0.09	0.12
	Medium	0.10	0.05	0.07
	Med Low	0.07	0.01	0.03
	Low	0.82	0.94	0.88
	정확도 (Accuracy)	Accurate matches: 70.11% Combined matches: 73.19% Adjacent matches: 82.57%		

실험의 결과로써, 본 연구가 가진 텍스트 형태의 입력 자료에 대한 벡터화 방식은 doc2vec 알고리즘 보다 tf-idf 알고리즘이 평균적으로 예측력이 높은 것을 확인하였다. doc2vec은 특히 5단계 구분에서 Medium과 Med Low 단계의 정밀도와 재현율이 낮다는 특징을 보였다. 또한, 실험을 수행하는 과정에서 입력 단어에 대한 추론(inference)으로 인해 입력 데이터 전체를 벡터화하기 위해 요구되는 시간이 tf-idf에 비해 현저하게 많이 필요하다는 단점이 존재한다.

doc2vec 알고리즘은 일반적으로 학습 데이터가 많은 경우에 보다 더 좋은 성능을 보이기 때문에 학습 데이터가 상대적으로 적은 현재의 실험에서는 tf-idf가 상대적으로 더 바람직한 알고리즘으로 판단된다. 그러므로 이후의 실험은 tf-idf 알고리즘을 활용하여 결과를 도출하는 것을 기본으로 하고, 2017년 국가연구개발사업에 대한 예측 결과에 대해서는 tf-idf와 doc2vec의 결과를 모두 수행하여 비교 분석한 결과를 도출하기로 하였다.

본 연구에서 개발한 분류 모델은 Low 단계의 데이터가 절대적으로 많은 상황에

서도 Low 단계에 치우치지 않고 5단계 기준으로 약 70% 정도의 정확성을 보여 비교적 신뢰할만한 수준의 분류 결과를 가진다고 판단된다. Med* 데이터는 중심에 수렴하지 않고 분포가 곳곳에 산재되어 있는데 그 분포는 벡터화 방법에 따라 약간의 차이를 보였다.

① tf-idf는 오분류된 다수의 Med*가 Low와 High 단계로 분산 분류되고 있다.

	High		Medium		Low
훈련 데이터 -	19	: 8 :	7	: 7 :	60
실험 데이터 -	23	: 3 :	3	: 3 :	66

② doc2vec의 경우에는 Med*의 다수가 High 보다 Low 단계로 분류되었다.

	High		Medium		Low
훈련 데이터 -	19	: 8 :	7	: 7 :	60
실험 데이터 -	20	: 3 :	4	: 3 :	70

전반적으로 단계 수가 적을수록 정확도가 상대적으로 우수하다고 판단할 수 있다. 예를 들어 2단계 분류의 경우 훈련 데이터 분포인 40.3:59.7에 거의 근접하는 40.2:59.8의 분포를 보였다. 이는 자칫 전문가의 분류에 근접하는 매우 정확한 분류로 판단할 수도 있으나 혼동 매트릭스를 확인해 보면 High(Non-Low) 단계가 Low 단계로 오인된 것과 거의 유사한 비율로 Low 단계가 High 단계로 오인된 결과에 따른 것이다.

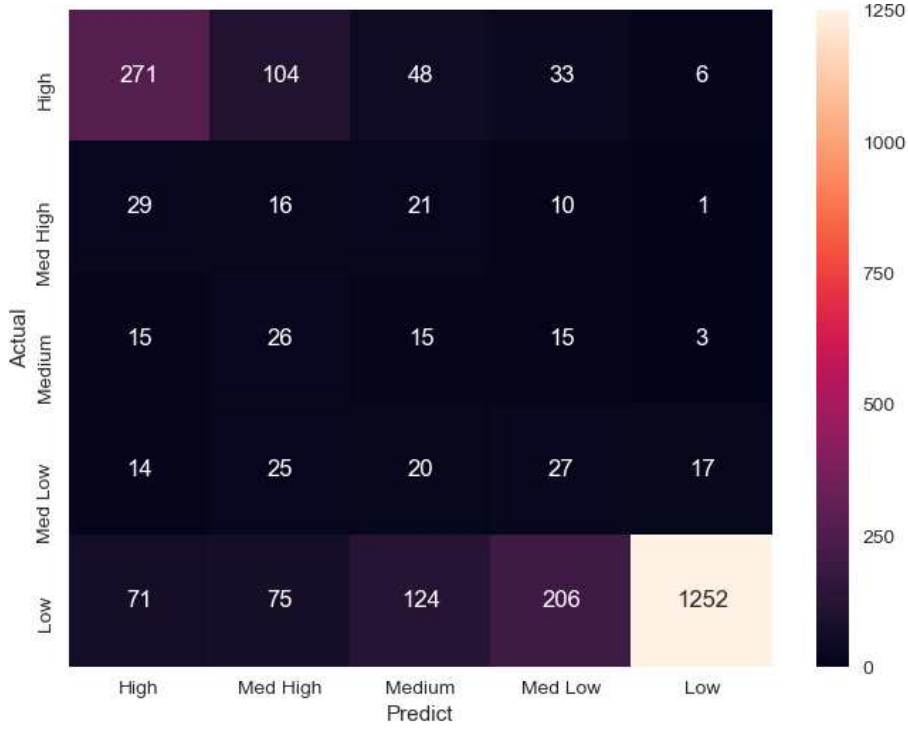
제3절 분류 모델의 예측 결과 비교 (비중 조정)

비중 조정 전과 조정 후의 입력 데이터에 대한 비교를 위해 tf-idf와 Linear SVM 모델을 활용하여 실험한 결과는 [그림 4-4]와 [그림 4-5]와 같다. 모델이 예측한 단계가 X축이고 데이터가 가진 실제 단계는 Y축으로 표기한다. 비중 조정 전은 실제 High 단계를 High로 맞게 예측한 것이 188개, 실제 Low 단계를 Low로 맞게 예측한 것은 589개로 나타났다. 비중 조정 후는 실제 High 단계인 과제를 High로 맞게 예측한 것은 271개이고, 실제 Low 단계인 과제를 Low로 맞게

예측한 경우는 1,252개이다.

[그림 4-4] 비중 조정 전의 실험 결과 (1,094개)

[그림 4-5] 비중 조정 후의 실험 결과 (2,444개)



비중 조정 전후 정확도를 비교해보면, 5단계 정매칭에서 비중 조정 전이 <표 4-4>에서 확인한 것과 같이 정확도가 72.78%, 비중 조정 후는 67.96%이다. 중간 세 단계(Med High, Medium, Med Low)의 정밀도는 비중 조정 전이 더 높고, 재현율은 비중 조정 후가 더 낮은 것으로 나타났다. 즉, 비중을 조정한 후에는 예측 모델이 중간의 3단계로 예측한 과제 중에 제대로 맞춘 경우가 낮아지고, 실제 3단계에 속하는 과제 중에 정답을 맞춘 경우가 많아지는 특징을 보인다고 볼 수 있다. 이를 종합하자면, 비중을 조정하지 않은 것이 중간 단계의 과제들이 상대적으로 차지하는 비중은 적지만, 분류 모델의 예측력 차원에서는 성능이 더 높다는 것을 알 수 있다.

<표 4-4> 비중 조정 전/후의 10회 평균 예측력 비교 (tf-idf & Linear SVM 기준)

구분	단계 (Scale)	정밀도 (Precision)	재현율 (Recall)	f1-score
비중 조정 전	High	0.59	0.76	0.67
	Med High	0.17	0.06	0.09
	Medium	0.18	0.07	0.09
	Med Low	0.22	0.06	0.09
	Low	0.83	0.95	0.89
	정확도 (Accuracy)	Accurate matches: 72.78% Combined matches: 75.31% Adjacent matches: 84.50%		
비중 조정 후	High	0.7	0.545	0.61
	Med High	0.095	0.265	0.145
	Medium	0.09	0.23	0.13
	Med Low	0.09	0.245	0.13
	Low	0.965	0.78	0.86
	정확도 (Accuracy)	Accurate matches: 67.96% Combined matches: 71.81% Adjacent matches: 85.42%		

비중을 조정하지 않은 학습 데이터에 대해 tf-idf를 활용하여 3개 모델의 예측값을 측정한 결과는 <표 4-5>와 같다. 결과 값을 5단계로 매칭한 정매칭(Accurate match)의 경우, Linear SVM은 72.78%, 결과 값을 3단계로의 구분하는 결합 매칭(Combined match)의 경우 75.31%, 인접 매칭(Adjacent match)은 84.50%의 예측률을 보였다.

<표 4-5> 모델별 10회 평균 정확도, 정밀도, 재현율 값 비교

모델	단계 (Scale)	정밀도 (Precision)	재현율 (Recall)	f1-score
Linear SVM	High	0.59	0.76	0.67
	Med High	0.17	0.06	0.09
	Medium	0.18	0.07	0.09
	Med Low	0.22	0.06	0.09
	Low	0.83	0.95	0.89
	정확도 (Accuracy)	Accurate matches: 72.78% Combined matches: 75.31% Adjacent matches: 84.50%		
Logistic Regression	High	0.60	0.78	0.67
	Med High	0.40	**0.00**	0.01
	Medium	0.29	**0.03**	0.05
	Med Low	0.34	**0.01**	0.02
	Low	0.79	0.97	0.87
	정확도 (Accuracy)	Accurate matches: 73.29% Combined matches: 73.29% Adjacent matches: 83.09%		
Multinomial NB	High	0.55	0.88	0.67
	Med High	0.19	0.05	0.08
	Medium	0.23	0.10	0.14
	Med Low	0.23	0.07	0.10
	Low	0.88	0.92	0.90
	정확도 (Accuracy)	Accurate matches: 73.05% Combined matches: 75.79% Adjacent matches: 85.10%		

Logistic Regression은 5단계의 매칭 결과가 가장 좋으나 Med*(Med High, Medium, Med Low) 단계의 재현율(Recall)이 낮은 것으로 나타났다. 재현율이 낮다는 것은 다른 모델에 비해 실제로 중간 세 단계(Med High, Medium, Med Low)에 속하는 과제이지만 예측을 제대로 하지 못한다는 것을 의미한다. Multinomial NB의 정매칭 결과는 73.05%이고, 결합 매칭과 인접 매칭이 각각 75.79%와 85.10%로 가장 높은 예측률을 나타냈다. 이러한 결과를 종합적으로 살펴보았을 때, Multinomial NB를 기준으로 예측을 위한 분류를 수행하고 결합 매칭과 인접 매칭에 좋은 성능을 보이는 Linear SVM도 함께 수행하여 두 모델의 결과를 모두 비교하고자 한다.

제5장 SW 융합 R&D 예측 및 검증

제1절 2017년 국가연구개발과제의 예측

2017년도 국가연구개발사업의 총 과제 수는 61,280건으로 NTIS에서 제공하는 다양한 정보를 취합한 방대한 분량의 데이터 집합이다. 분류를 위한 지표로 사용된 연구 요약문의 “연구 내용”과 “연구 목표”, “한글 키워드”는 한글 데이터이며 이들을 네이버 파파고 번역기를 사용하여 번역한 후 연구 내용과 연구 목표가 누락되거나 번역의 품질이 매우 나쁜 항목 및 연구 요약문의 연구 내용에 보안 과제로 기재된 1,219건의 과제를 제외하였다. 이 과정을 거쳐 최종적으로 55,566건의 2017년도 과제에 대하여 단계가 구분된 2016년도 3,124건의 연구 과제로 학습시킨 분류 모델을 사용하여 나머지 55,566건을 분류하도록 전체 연구 과정을 설정하였다.

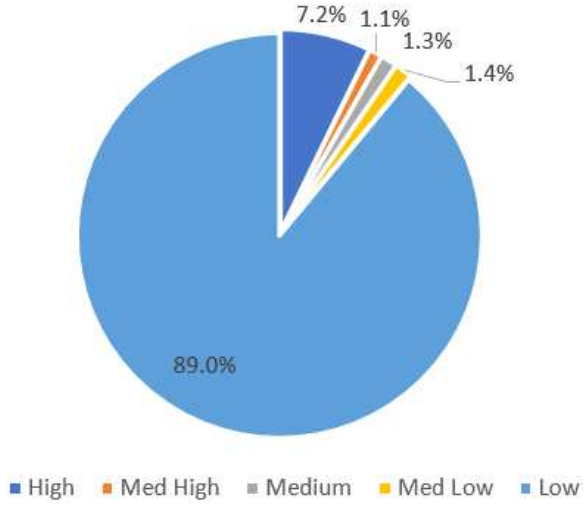
2017년 국가연구개발사업에 대한 예측은 워드 임베딩 방식으로 tf-idf와 doc2vec 알고리즘을 비교하기 위해 Linear SVM 분류 모델을 기준으로 두 경우를 모두 실험하여 비교하고자 한다. 또한, 워드 임베딩 방식으로 tf-idf를 사용한다고 가정하고, 분류 모델에 대해 Linear SVM과 Multinomial NB는 어떻게 예측하는지를 각각 실험하였다. 단계 분류는 기본적으로 5단계 구분과 함께 3단계와 2단계 구분을 기준으로 예측을 수행한다.

1. Linear SVM 모델의 벡터화 알고리즘 및 단계별 예측 결과

① tf-idf & 5단계 분류

다음은 tf-idf를 활용하여 2017년 55,566건의 과제 데이터에 대하여 5단계 형태로 분류한 예측 결과이다.

[그림 5-1] tf-idf 5단계 분류(SVM)



<표 5-1> tf-idf 5단계 과제 수/비중(SVM)

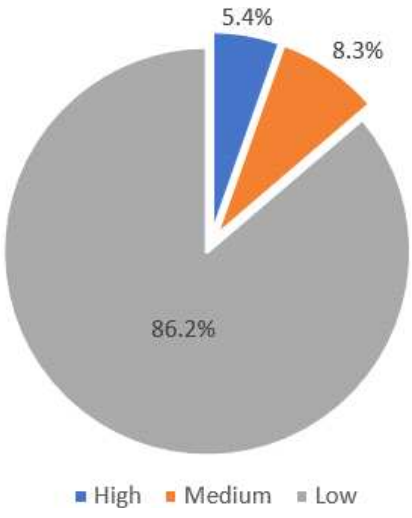
단계	과제 수	비중
High	4,012	7.22%
Med High	587	1.06%
Medium	746	1.34%
Med Low	776	1.40%
Low	49,445	88.98%
Total	55,566	100%

2017년 국가연구개발과제에 대하여 High 단계는 7.2%, Med High 단계는 1.1%, Medium 단계는 1.3%, Med Low 단계는 1.4%, 나머지 89.0%는 Low 단계로 분류되었다.

② tf-idf & 3단계 분류

다음은 3단계로 분류된 학습용 데이터로 훈련한 모델을 사용하여 2017년 55,566건의 국가연구개발과제를 분류한 결과이다. 살펴보면, 5.4%는 High 단계이고, 8.3%는 Medium 단계이며 나머지 86.2%는 Low 단계로 분류한 것을 확인할 수 있다.

[그림 5-2] tf-idf 3단계 분류(SVM)



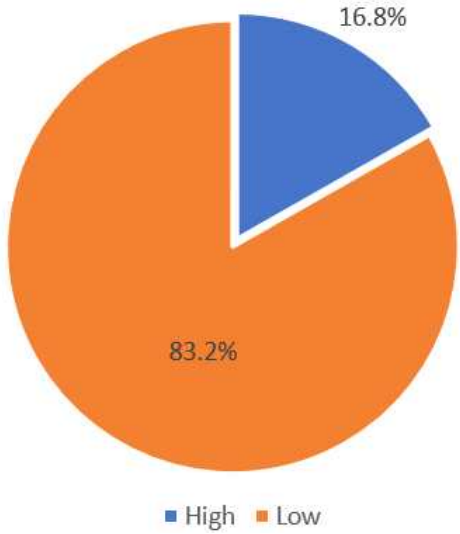
<표 5-2> tf-idf 3단계 과제 수/비중(SVM)

단계	과제 수	비중
High	3,022	5.44%
Medium	4,623	8.32%
Low	47,920	86.24%
Total	55,566	100%

③ tf-idf & 2단계 분류

다음은 2단계로 분류된 학습용 데이터로 훈련한 모델을 사용하여 나머지 55,566건의 데이터를 분류한 결과를 보인 것이다. 결과적으로 2017년 국가연구개발사업에 대하여 16.8%는 High 단계, 나머지 83.2%는 Low 단계로 분류한 것을 확인할 수 있다. 훈련에 사용한 데이터에 SW 융합 R&D 단계(Non-Low 단계)가 많이 증가함에 따라 Low 단계에 치우치는 현상이 어느 정도 해석되어 5단계나 3단계 분류에서의 SW 융합 R&D 단계의 합보다 2단계에서 High 단계가 더 높은 비율로 예측되었으며, 상대적으로 Low 단계의 비율은 감소하였다.

[그림 5-3] tf-idf 2단계 분류(SVM)



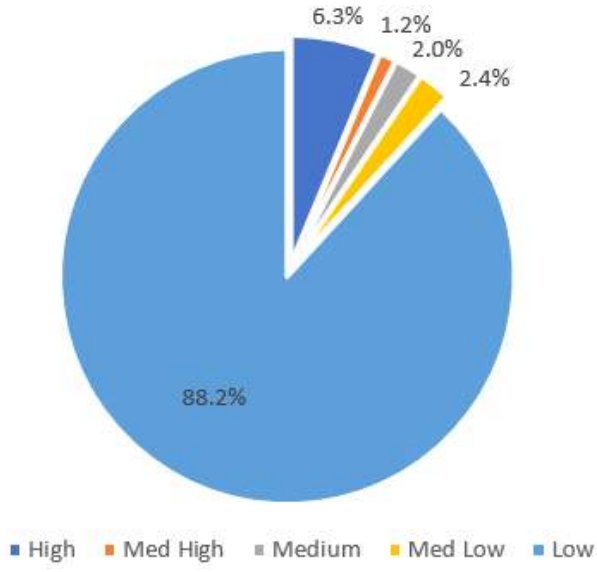
<표 5-3> tf-idf 2단계 과제 수/비중(SVM)

단계	과제 수	비중
High	9,325	16.78%
Low	46,241	83.22%
Total	55,566	100%

④ doc2vec & 5단계 분류

벡터화를 위한 알고리즘으로 doc2vec을 사용한 후 분류한 결과는 다음과 같다. 동일 조건에서 실험한 결과 비교를 위하여 학습 모델은 tf-idf를 사용한 경우와 동일하게 Linear SVM을 사용하였다. 다음은 5단계로 분류된 학습용 데이터로 훈련한 모델을 사용하여 55,566건의 데이터를 분류한 결과를 보인 것이다.

[그림 5-4] doc2vec 5단계 분류(SVM)



<표 5-4> doc2vec 5단계 과제 수/비중(SVM)

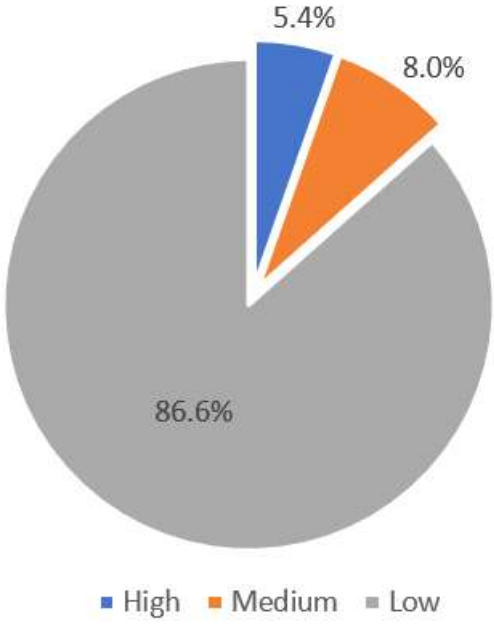
단계	과제 수	비중
High	3,499	6.30%
Med High	668	1.20%
Medium	1,070	1.96%
Med Low	1,313	2.36%
Low	49,017	88.21%
Total	55,566	100%

결과적으로 2017년 국가연구개발과제에 대하여 6.3%는 High 단계, 1.2%는 Med High 단계, 1.93%는 Medium 단계, 1.4%는 Med Low 단계, 나머지 88.21%는 Low 단계로 분류한 것을 확인할 수 있다. 이 결과는 3단계 분류에 비하면 여전히 Med*의 데이터들이 High 또는 Low 단계로 이동한 것이나, tf-idf를 사용한 경우에 비해 상대적으로 적은 양의 데이터가 이동한 것을 확인할 수 있다. 그러므로 Linear SVM 모델의 경우, 5단계 분류의 안정성을 고려하면 doc2vec이 더 안정적인 벡터화 알고리즘으로 볼 수 있다. 그러나 언급한 것과 같이 doc2vec은 실험 데이터 전체를 벡터화하는데 매우 많은 실행 시간이 요구된다는 단점이 있다.

⑤ doc2vec & 3단계 분류

다음은 3단계로 분류된 학습용 데이터로 훈련한 모델을 사용하여 동일한 데이터를 분류한 결과를 보인 것이다. 결과적으로 2017년 국가연구개발과제에 대하여 5.4%는 High 단계, 8.0%는 Med* 단계, 나머지 86.6%는 Low 단계로 분류한 것을 확인할 수 있다. 이 결과는 tf-idf를 사용한 실험 결과와 거의 동일하여 벡터화 알고리즘의 차이로 인한 영향은 미미함을 알 수 있다.

[그림 5-5] doc2vec 3단계 분류(SVM)



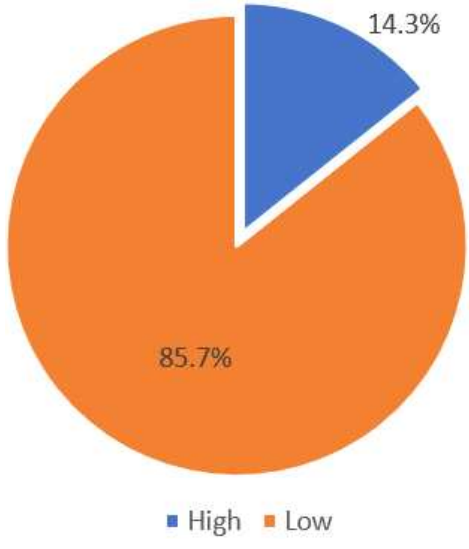
<표 5-5> doc2vec 3단계 과제 수/비중(SVM)

단계	과제 수	비중
High	3,005	5.41%
Medium	4,463	8.03%
Low	48,098	86.56%
Total	55,566	100%

⑥ doc2vec & 2단계 분류

다음은 doc2vec으로 벡터화한 데이터의 2단계 분류 결과를 보인 것이다. 결과는 14.8%는 High 단계, 나머지 85.2%는 Low 단계로 분류한 것을 확인할 수 있다.

[그림 5-6] doc2vec 2단계 분류(SVM)



<표 5-6> doc2vec 2단계 과제 수/비중(SVM)

단계	과제 수	비중
High	8,107	14.28%
Low	48,678	85.72%
Total	55,566	100%

2. Linear SVM 모델의 텍스트 마이닝 분류 방법의 비교

다음 <표 5-7>은 위에서 설명한 tf-idf/doc2vec과 Linear SVM 조합의 실험 결과를 종합적으로 나타낸 것이다.

<표 5-7> 2017년도 국가연구개발과제 분류 결과 요약 (Linear SVM)

	5단계			3단계			2단계		
tf-idf	High	4,012	7.22%	High	3,022	5.44%	High	9,325	16.78%
	Med High	587	1.06%	Medium	4,623	8.32%	Low	46,241	83.22%
	Medium	746	1.34%	Low	47,920	86.24%			
	Med Low	776	1.40%						
	Low	49,445	88.98%						
doc2vec	High	3,499	6.30%	High	3,005	5.41%	High	8,107	14.28%
	Med High	668	1.20%	Medium	4,463	8.03%	Low	48,678	85.72%
	Medium	1,070	1.96%	Low	48,098	86.56%			
	Med Low	1,313	2.36%						
	Low	49,017	88.21%						

tf-idf에서의 결과를 보면 5단계 분류에서는 데이터 개수가 상대적으로 적은 Med* 단계의 데이터가 High 혹은 Low 단계로 일부 이동하였음을 알 수 있다. 이는 학습 모델이 데이터의 양이 많은 분류에 치우친 결과로 기계학습 알고리즘들이 가진 기본적인 약점으로 인식된다. 이 문제에 대한 해결책으로는 Med Low, Medium, Med High 단계에 속하는 데이터를 더 많이 확보하여 치우침 현상을 회피하는 것이다. 그러나 본 연구에서 활용한 2016년도 국가연구개발과제만으로는 치우침 현상을 회피하는 것이 매우 어려우며, 향후 연구에서 2016년과 2017년 국가연구개발과제를 학습 데이터로 사용하여 그 이후 연도를 위한 연구과제를 분석하여 이러한 해석의 타당성을 확인할 수 있을 것이다.

3단계 분류에서는 벡터화 알고리즘에 따라 약간의 편차는 있지만, 안정적으로 High 단계는 약 5.4%, Med* 단계는 약 8.3%, Low 단계는 약 86.6%로 분류된 결과를 보였다. 본 연구의 목적이 국가연구개발사업에서의 SW 융합 정도를 파악하기 위한 것이므로 약 13.5% 정도의 과제들이 SW를 융합한 연구개발을 수행한 것으로 판단할 수 있다.

한편 2단계 분류에서는 16.78%가 High(Non-Low) 단계로, 나머지 전체인

83.22%가 Low 단계로 판별되었다. 학습 데이터에서 High(Non-Low) 단계가 40% 정도를 차지하여 훈련된 모델이 Low에 치우치는 이슈가 상당 부분 해소되어 5단계나 3단계 분류에 비해 Low의 비중이 많이 낮아진 것을 확인하였다.

tf-idf 대신 doc2vec을 사용한 경우에도 유사한 결과를 얻었다. doc2vec의 경우 단계 수의 차이에도 불구하고 치우침 현상이 상대적으로 적게 발생하여 5단계와 2단계 사이의 Low 단계의 비율 차이가 2.75% 밖에 나지 않는 안정적인 모델임을 알 수 있다. tf-idf의 경우는 그 차이가 약 5.76% 정도로 나타났다.

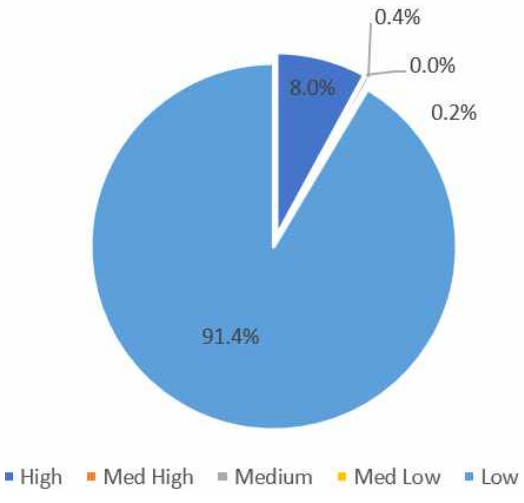
이와 같이 벡터화 알고리즘에 관계없이 분류의 대세 경향은 유사한 패턴을 보이나 doc2vec으로 벡터화한 경우 치우침 현상이 덜 발생하여 단계 수의 차이에도 불구하고 안정적인 예측을 할 수 있는 특징을 보였다. 또한, 단계 수가 많을수록 Med* 단계의 데이터가 High 혹은 Low 단계로 치우침이 발생하여 결과가 왜곡되는 것을 알 수 있다. Linear SVM 모델에서는 tf-idf 방식이 학습 데이터가 가진 특성을 반영하고 있으며, doc2vec 방식으로 벡터화한 데이터에 대하여 3단계 혹은 2단계로 예측한 결과는 보다 안정적인 방식이라고 판단할 수 있다.

3. Multinomial NB 모델의 벡터화 알고리즘 및 단계별 예측 결과

① tf-idf & 5단계 분류

다음은 tf-idf를 활용하여 2017년 55,566건의 과제 데이터에 대하여 5단계 형태로 분류한 예측 결과이다.

[그림 5-7] tf-idf 5단계 구분(MNB)



<표 5-8> tf-idf 5단계 과제 수/비중(MNB)

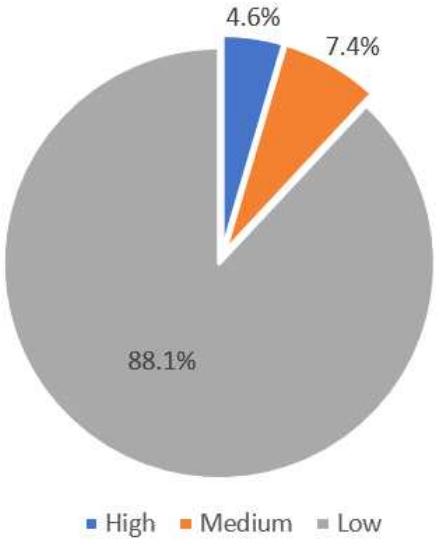
단계	과제 수	비중
High	4,513	7.95%
Med High	9	0.02%
Medium	241	0.42%
Med Low	109	0.19%
Low	51,913	91.42%
Total	55,566	100%

2017년 국가연구개발과제에 대하여 High 단계는 약 7.95%, Med High 단계는 0.02%, Medium 단계는 0.42%, Med Low 단계는 0.19%, 나머지 91.42%는 Low 단계로 분류되었다.

② tf-idf & 3단계 분류

다음은 3단계로 분류된 학습용 데이터로 훈련한 모델을 사용하여 2017년 55,566건의 국가연구개발과제를 분류한 결과이다. 결과적으로 4.56%는 High 단계이고, 7.37%는 Med* 단계이며 나머지 88.07%는 Low 단계로 분류한 것을 확인할 수 있다.

[그림 5-8] tf-idf 3단계 구분(MNB)



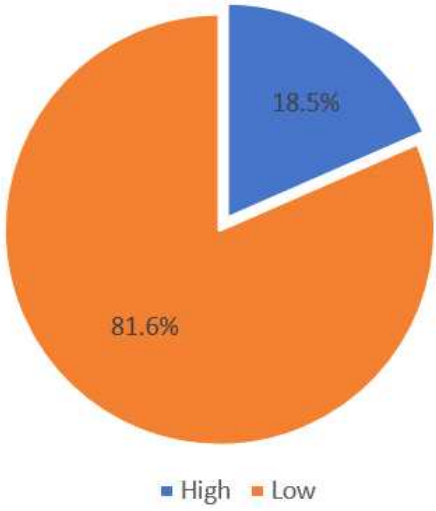
<표 5-9> tf-idf 3단계 과제 수/비중(MNB)

단계	과제 수	비중
High	2,590	4.56%
Medium	4,185	7.37%
Low	50,010	88.07%
Total	55,566	100%

③ tf-idf & 2단계 분류

다음은 2단계로 분류된 학습용 데이터로 훈련한 모델을 사용하여 나머지 55,566건의 데이터를 분류한 결과를 보인 것이다. 결과적으로 2017년 국가연구개발사업에 대하여 18.45%는 High 단계, 나머지 81.55%는 Low 단계로 분류한 것을 확인할 수 있다.

[그림 5-9] tf-idf 2단계 구분(MNB)



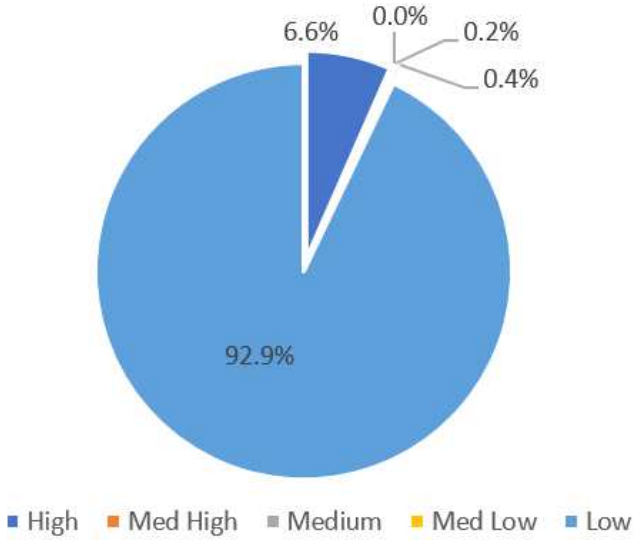
<표 5-10> tf-idf 2단계 과제 수/비중(MNB)

단계	과제 수	비중
High	10,477	18.45%
Low	46,308	81.55%
Total	55,566	100%

④ doc2vec & 5단계 분류

동일 조건에서 실험한 결과 비교를 위하여 학습 모델은 Multinomial NB 모델을 사용하고, 벡터화를 위한 알고리즘으로 doc2vec을 사용한 후 분류한 결과는 다음과 같다. 결과적으로 2017년 국가연구개발과제에 대하여 6.58%는 High 단계, Med High 단계는 0%, Medium 단계는 0.16%, Med Low 단계는 0.35%, 나머지 92.91%는 Low 단계로 분류한 것을 확인할 수 있다.

[그림 5-10] doc2vec 5단계 구분(MNB)



<표 5-11> doc2vec 5단계 과제 수/비중(MNB)

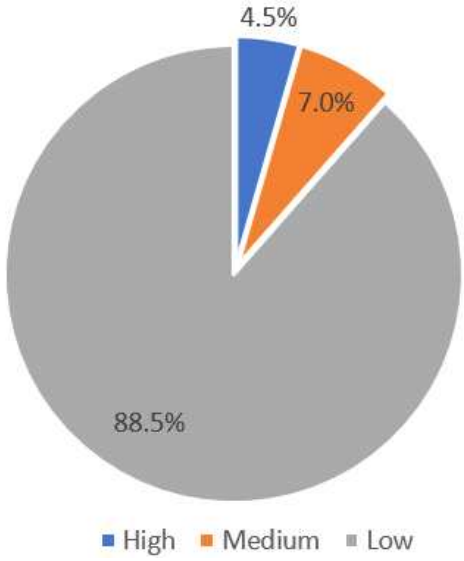
단계	과제 수	비중
High	3,736	6.58%
Med High	0	0.00%
Medium	89	0.16%
Med Low	200	0.35%
Low	52,760	92.91%
Total	55,566	100%

⑤ doc2vec & 3단계 분류

다음은 3단계로 분류된 학습용 데이터로 훈련한 모델을 사용하여 동일한 데이터를 분류한 결과를 보인 것이다.

결과적으로 2017년 국가연구개발과제에 대하여 4.46%는 High 단계, 7.0%는 Med* 단계, 나머지 88.54%는 Low 단계로 분류한 것을 확인할 수 있다. 이 결과는 tf-idf를 사용한 실험 결과와 거의 동일하여 벡터화 알고리즘의 차이로 인한 영향은 미미함을 알 수 있다.

[그림 5-11] doc2vec 3단계 구분(MNB)



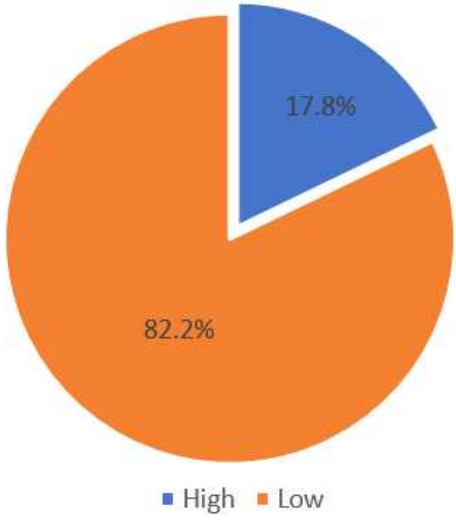
<표 5-12> doc2vec 3단계 과제 수/비중(MNB)

단계	과제 수	비중
High	2,534	4.46%
Medium	3,976	7.00%
Low	50,275	88.54%
Total	55,566	100%

⑥ doc2vec & 2단계 분류

다음은 doc2vec으로 벡터화한 데이터의 2단계 분류 결과를 보인 것이다. 결과는 17.84%는 High 단계, 나머지 82.16%는 Low 단계로 분류한 것을 확인할 수 있다.

[그림 5-12] doc2vec 2단계 구분(MNB)



<표 5-13> doc2vec 2단계 과제 수/비중(MNB)

단계	과제 수	비중
High	10,129	17.84%
Low	46,656	82.16%
Total	55,566	100%

4. Multinomial NB 모델의 텍스트 마이닝 분류 방법의 비교

다음 <표 5-14>는 위에서 설명한 tf-idf/doc2vec 알고리즘과 Multinomial NB 모델 조합의 실험 결과를 종합적으로 요약한 것이다.

<표 5-14> 2017년도 국가연구개발과제 분류 결과 요약(Multinomial NB)

	5단계			3단계			2단계		
tf-idf	High	4,513	7.95%	High	2,590	4.56%	High	10,477	18.45%
	Med High	9	0.02%	Medium	4,185	7.37%	Low	46,308	81.55%
	Medium	241	0.42%	Low	50,010	88.07%			
	Med Low	109	0.19%						
	Low	51,913	91.42%						
doc2vec	High	3,736	6.58%	High	2,534	4.46%	High	10,129	17.84%
	Med High	0	0.00%	Medium	3,976	7.00%	Low	46,656	82.16%
	Medium	89	0.16%	Low	50,275	88.54%			
	Med Low	200	0.35%						
	Low	52,760	92.91%						

Multinomial NB 모델은 5단계 예측에서 tf-idf와 doc2vec 모두 Med* 단계의 예측을 거의 하지 못하는 것으로 나타났다. 특히 Med High 단계의 과제를 거의

분류하지 못한 것으로 나타났다. 또한, 두 알고리즘에서 모두 5단계와 3단계 구분에서 Low 단계로 구분한 과제가 약 90% 초반과 약 88% 정도의 수치를 보이고 있어 2016년 학습 데이터에 비해 Low 단계로의 예측에 대한 쏠림 현상이 있는 것으로 조사되었다.

tf-idf의 결과에서 2단계와 5단계의 Low 단계에 대한 비율 차이는 9.87%이고, doc2vec은 10.75%의 차이를 보였고, 5단계에서도 doc2vec은 tf-idf에 비해 Med* 단계에 대한 예측을 하지 못하는 것으로 나타났다. 그러므로 기계학습 모델(Multinomial NB/Linear SVM) 및 벡터화 알고리즘(tf-idf/doc2vec)의 교차 검증 결과, 기계학습 모델로는 Linear SVM과 tf-idf 방식으로 벡터화한 데이터에 대하여 2단계로 예측한 것이 2016년 국가연구개발사업에서 표본 추출한 학습 데이터와 가장 유사한 결과를 도출[14]한다고 판단할 수 있다.

제2절 예측에 대한 전문가 검증

2016년 국가연구개발과제 중 3,214건의 과제 정보를 입력 데이터로 학습한 기계학습 모델을 2017년 국가연구개발과제를 대상으로 적용하여 예측 결과를 도출하였다. 이번 절에서는 본 연구에서 개발한 분류 모델의 예측 결과를 검증하기 위해 텍스트 마이닝 방법으로 tf-idf를 활용하고 분류 모델은 Linear SVM을 활용한 예측 결과 중, 100건의 표본을 추출한 뒤 SW 융합 R&D 전문가의 검증을 수행하여 모델의 예측 결과와 비교하고자 한다.

일반적으로 표본 추출을 위한 가장 기본적인 방법은 단순 무작위 추출법(Simple random sampling)이다. 그러나 본 연구에서는 추출된 표본이 5가지의 SW 융합 R&D 단계의 특징을 모두 반영할 필요가 있기 때문에 층화 추출법(Stratified sampling)을 활용하고자 한다. 층화 추출법은 모집단을 구성하는 층에 대한 정보를 표본 설계에 반영하여 추정량의 분산을 낮추기 위한 표본 추출 방법으로 알려져 있다(이인규, 2015). 층화 추출을 위해 2017년 국가연구개발과제에서 연구요약문 중 연구 목표 및 연구 내용이 존재하지 않은 과제를 제외한 총 55,566건에 대해 모델에 의해 예측된 SW 융합 R&D 단계를 기준으로 <표 5-15>

14) 2018년의 선행 연구 결과, 과제수 기준으로 15.9%, 총 연구비 기준으로 16.3%가 SW 융합 R&D 과제인 것으로 조사됨(국가연구개발사업에서의 SW융합에 관한 연구, SPRi, 2018)

과 같이 5개의 층을 두어 표본을 추출하였다.

<표 5-15> 검증을 위한 표본 추출 방법

표본 추출 단위	층화 변수	표본 추출 방식	표본의 크기
과제 수	SW 융합 R&D 단계 (총 5개)	층화추출법	100건

표본 배분은 비례 배정 방식을 활용하였고, 아래는 그 할당 식을 나타낸다.

$$n_h = n \times \frac{N_h}{\sum_{h=1}^{L} N_h}$$

n_h : h층에 할당된 표본 수

N : 전체 과제 수 (55,566건)

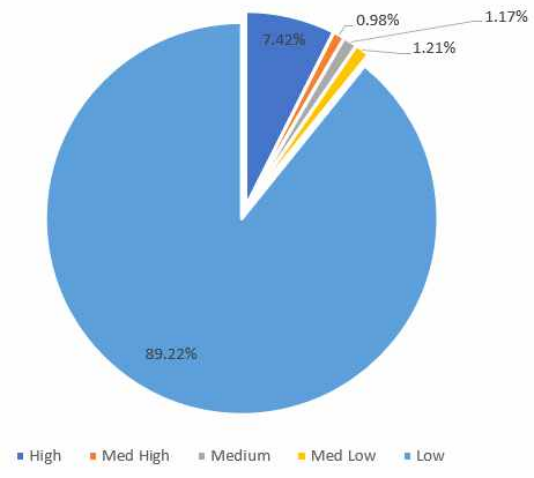
n : 전체 표본 수 (100건)

N_h : h층에 있는 전체 과제 수

L : SW 융합 R&D 단계 층 (총 5개)

SW 융합 R&D 단계는 [그림 5-13]과 같이 High가 7.42%, Low가 89.22% 등의 층별 분포를 보이고 있다.

[그림 5-13] 5단계 과제 비중(SVM 기준)

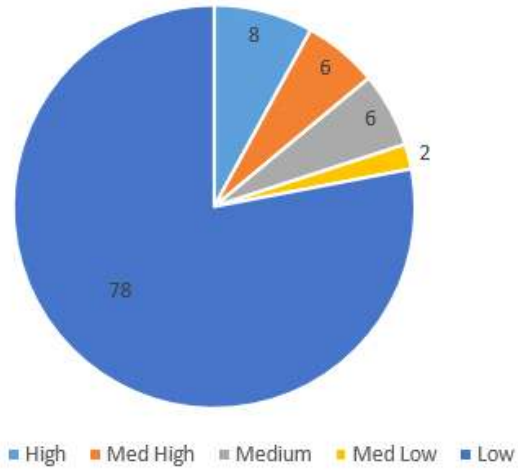


<표 5-16> 검증용 표본 도출

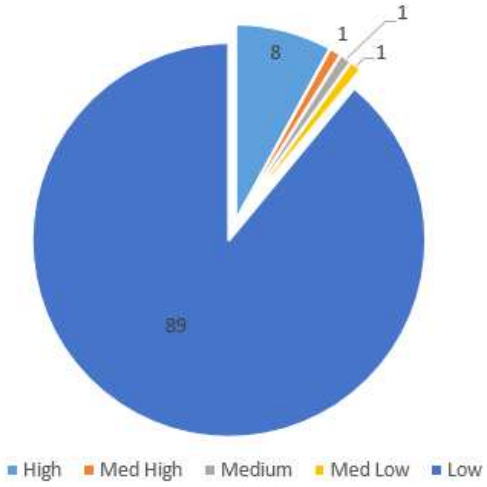
SW 융합 R&D 단계	표본 수
High	8
Med High	1
Medium	1
Med Low	1
Low	89
Total	100

100건의 표본 과제에 대한 전문가의 검증 결과는 [그림 5-14]와 같고 예측 모델의 결과는 [그림 5-15]의 분포를 보였다. High 단계는 전문가와 분류 모델의 예측이 모두 8개로 동일하다. 그러나 전문가가 Med High, Medium 단계로 예측한 과제가 개발된 분류 모델은 1개씩만 해당 단계로 예측하고 나머지는 Low 단계인 것으로 예측하여 전문가가 예상한 Low 단계가 분류 모델에 비해 12.4% 정도 적은 것을 알 수 있다.

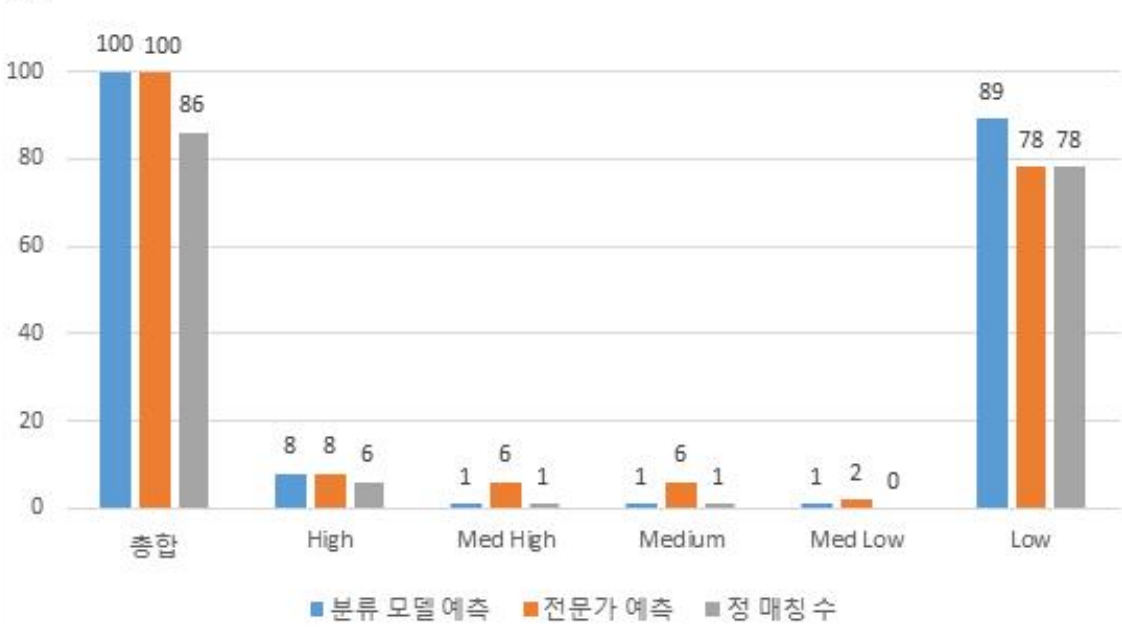
[그림 5-14] 전문가의 예측 분포



[그림 5-15] 모델의 예측 분포



[그림 5-16] 분류 모델 및 전문가 판단 매칭 결과



보다 자세한 검증을 위해 5가지 단계로 구분한 전문가의 판단과 분류 모델이 예측한 결과를 정확하게 맞춘 것에 해당하는 정 매칭, 근접한 1개 단계까지 맞춘 것으로 허용하는 인접 매칭, 2단계 이상의 차이는 맞추지 못한 것으로 보는 비 매칭으로 구분하였다. 전문가 판단 결과를 정답으로 보고, 개발된 분류 모델의 매칭 결과를 대조해 보면 <표 5-17>과 같이 분류 모델이 Low 단계로 판단한 것은 전문가도 모두 Low 단계로 판단하였다. High 단계는 전문가의 판단과 분류 모델이 75% 일치하였고, Med High와 Medium 단계는 16.7%가 일치한 것을 알 수 있다.

<표 5-17> 전문가 예측 기준 정 매칭, 인접 매칭, 비 매칭 수와 비중

전문가 예측	개수	정 매칭		인접 매칭		비 매칭	
		개수	비중	개수	비중	개수	비중
High	8	6	75.0%	0	0.0%	2	25.0%
Med High	6	1	16.7%	1	16.7%	4	66.7%
Medium	6	1	16.7%	0	0.0%	5	83.3%
Med Low	2	0	0.0%	2	100.0%	0	0.0%
Low	78	78	100.0%	0	0.0%	0	0.0%
합계	100	86	86.0%	3	3.0%	11	11.0%

전문가가 예측한 결과를 기준으로 분류 모델의 예측 결과에 대한 5단계 정 매칭의 정밀도와 재현율은 <표 5-18>과 같다. 재현율은 전문가가 A 단계로 판단한 것 중에 분류 모델 역시 A 단계로 예측한 것을 나타내며, Med High와 Medium, Med Low에서 낮은 수치를 보였다. 정밀도는 분류 모델이 A 단계로 예측한 것 중 전문가 결과도 A 단계로 판단한 것에 해당하며 Med High와 Medium 단계는 전문가와 분류 모델의 결과가 일치하는 것으로 나타났으나 Med Low 단계는 정밀도가 0인 것을 알 수 있다. 이를 통해 본 연구에서 개발한 분류 모델이 Med High, Medium, Med Low 단계에 해당하는 과제를 상대적으로 과소하게 예측하는 경향을 보인다고 판단할 수 있다.

<표 5-18> 단계별 재현율, 정밀도 검증 (전문가 판단 기준)

예측 단계	재현율(Recall)	정밀도(Precision)
High	0.75	0.75
Med High	0.17	1.00
Medium	0.17	1.00
Med Low	0.00	0.00
Low	1.00	0.88
합계	0.86	0.86

전문가 검증 결과를 종합하면, SW R&D 전문가가 분류한 결과와 자동 분류 모델이 예측한 결과는 86%가 일치하는 것을 알 수 있다. 하지만 본 연구의 학습 데이터가 가지는 특성 때문에 Med High, Medium과 Med Low 단계는 전문가와 분류 모델이 일치하는 결과가 적었다. 이를 통해 개발한 분류 예측 모형은 High와 Low 단계를 구분하는 것에 보다 더 특화되었다고 판단할 수 있다.

제3절 모델 예측에 대한 설명

본 연구에서 사용한 지도학습 방법은 학습에 사용되는 데이터가 충분히 누적되어 있고 과거 데이터가 미래 데이터 예측을 위한 충분한 근거가 될 때 더욱 유의미한 결과를 산출한다. 그러나 실제 연구에서는 델파이 조사에서 선별한 3,124건의 2016년도 연구 자료를 훈련 데이터로 사용하여 학습한 모델을 사용하여 약 20여 배에 달하는 2017년도 연구 자료에 대한 예측을 시도하였다. 그 결과, 2016년도 데이터에 준하는 수준의 신뢰도를 2017년도 데이터에 대해서 기대하는 것은 무리가 있다. 물론 분류 결과가 비교적으로 안정적인 결과를 산출하여 대세 경향을 파악하는 목적으로는 활용 가능하지만, 개별 데이터에 대하여 이해하기 힘든 분류 결과를 산출하였을 때 그와 같은 분류의 근거가 되는 설명을 요구할 수 있다면 개발된 분류기의 활용도가 한층 더 높아질 수 있을 것이다.

그러므로 본 연구에서는 Lime 알고리즘을 활용하여 설명 기능을 구현하였다. Lime은 입력 데이터를 변경시켜 가면서 예측이 어떻게 변하는지를 확인함으로써 특정 예에 대한 분류기의 결정에 대한 설명을 시도하는 블랙박스에 대한 설명자

이다. Lime을 사용하기 위해서는 먼저 다음과 같은 패키지를 사용하여야 한다.

```
from lime import lime_text
from lime.lime_text import LimeTextExplainer
```

한편 Lime의 적용을 위해서는 분류기가 predict_proba 함수를 지원해야 하나, 본 연구에서 선택한 Linear SVM은 이 함수를 지원하지 않는다. 이를 위한 해결책으로 임의의 분류기를 사용하여 교차 검증(cross validation)을 수행할 수 있는 분류기로 변환하는 CalibratedClassifierCV 클래스로 감싸는 방법을 사용하였다. 변환된 클래스의 객체는 predict_proba 함수를 지원하기 때문이다. 그러므로 벡터화를 실행한 후 교차검증기로 감싼 분류기를 실행하는 파이프라인을 실행하고 그 결과에 대하여 Lime을 적용하여 설명하도록 구현하였다.

설명 기능은 분류기 모델이 구축된 이후에 적용 가능하므로 한 번의 실행에 하나의 데이터에 대한 설명만 시도하는 것은 매우 비실용적이다. 이에 본 연구에서는 데이터 집합 원소들의 유효한 인덱스를 입력하여 그 인덱스의 데이터에 대한 설명을 계속하는 인터페이스를 구현하여 이 문제를 해결하였다. 인덱스 대신 ‘q’를 입력하면 설명의 종료를 수행하고, ENTER를 입력하면 가장 최근에 설명한 데이터 바로 다음 데이터에 대한 예측을 설명한다. 다음은 임의로 선택한 몇 건의 데이터에 대한 3단계 예측 결과와 설명을 보인 것이다.

1) High 단계

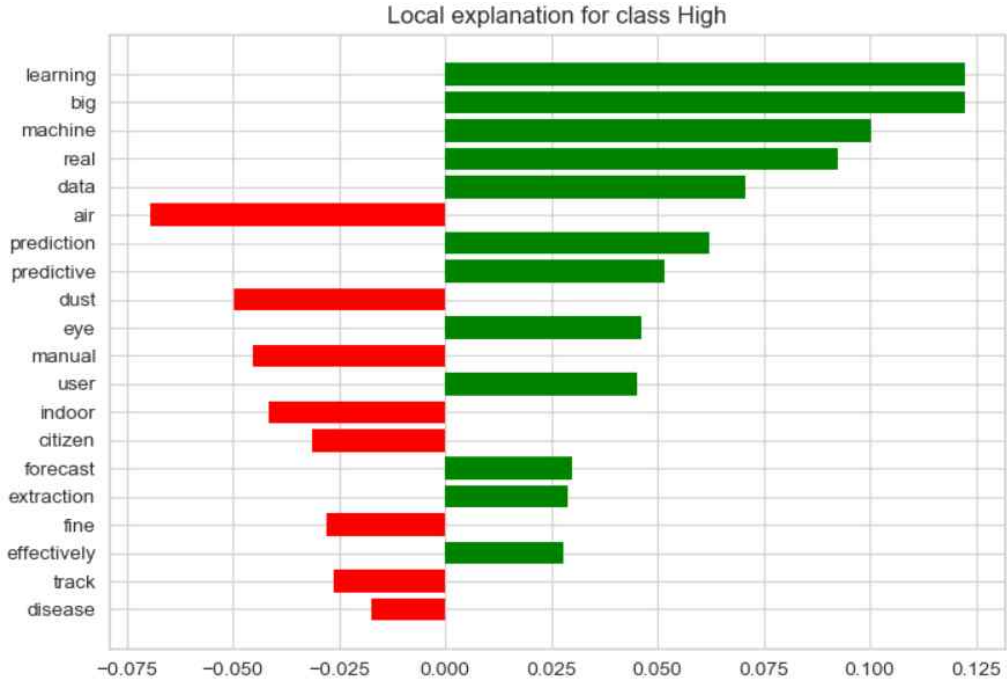
다음은 High로 분류된 어떤 데이터에 대하여 Lime을 이용하여 분류의 근거를 설명한 것이다. 예측 설명에서 보면 <표 5-19>와 같이 “learning”, “big”, “machine”, “real” 등과 같은 단어들이 주어진 데이터를 High로 분류되게 하는 데 긍정적으로 작용한 것을 확인할 수 있다. 반면 “air”와 “dust”라는 단어는 High 단계가 아닌 것으로 영향을 주었으나, 긍정적으로 작용한 단어들의 기여도가 부정적으로 기여한 단어들의 기여도보다 월등히 높아서 High로 선택되었음을 확인할 수 있다. 이 사례는 수작업으로 SW와 유관한 과제를 선택하는 경우와도 직관적으로 일치한다.

<표 5-19> High 단계 판단에 대한 각 단어의 가중치

키워드	수치	키워드	수치
learning	0.12228990403	manual	-0.0453623161
big	0.12225027321	user	0.0451530031
machine	0.10009845616	indoor	-0.041682817
real	0.09246689101	citizen	-0.031306297
data	0.07081077260	forecast	0.0300156063
air	-0.0693695643	extraction	0.0288781918
prediction	0.06220993747	fine	-0.028046126
predictive	0.05156741569	effectively	0.0279486047
dust	-0.0498424672	track	-0.026384558
eye	0.04628498173	disease	-0.017569984
Sum of positives	0.799974037	Sum of negatives	-0.309564131

실제 구현에서는 절대값을 기준으로 중요도가 가장 높은 20건의 단어를 기준으로 텍스트와 그래프로 설명을 보여주도록 하였다. 아래 [그림 5-17]은 앞에서 보인 텍스트 출력 결과를 이해하기 쉽게 기여도의 절대값을 기준으로 좌우로 배치하여 그래프로 재구성한 것이다.

[그림 5-17] High 단계 판단에 대한 각 단어의 가중치 그래프



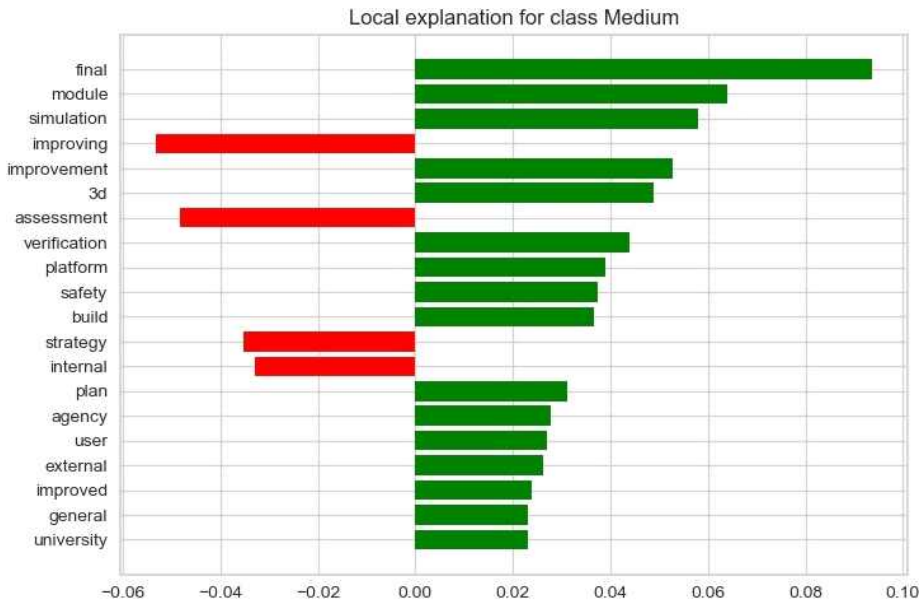
2) Medium 단계

다음은 Medium 단계로 분류된 어떤 데이터에 대하여 Lime을 이용하여 분류의 근거를 설명한 것이다. 예측 설명에서 보면 “module”, “3d”, “verification”, “platform” 등과 같은 단어들이 주어진 데이터를 Medium 단계로 분류되게 하는 데 긍정적으로 작용한 것을 확인할 수 있으며, 긍정적으로 작용한 단어들의 기여도가 부정적으로 기여한 단어들의 기여도보다 높아서 Medium 단계로 선택되었음을 확인할 수 있다.

<표 5-20> Medium 단계 판단에 대한 각 단어의 가중치

키워드	수치	키워드	수치
final	0.0935922364	build	0.0366459599
module	0.0639409427	strategy	-0.0351667732
simulation	0.0580278540	internal	-0.0329895472
improving	-0.0531898945	plan	0.0312001756
improvement	0.0528218853	agency	0.0278509482
3d	0.0487457561	user	0.0270339848
assessment	-0.0482455391	external	0.0262400789
verification	0.0439077128	improved	0.0239093026
platform	0.0390640581	general	0.0229560725
safety	0.0373600040	university	0.0229497142
Sum of positives	0.65624668613	Sum of negatives	-0.1695917540

[그림 5-18] Medium 단계 예측 데이터에 대한 설명



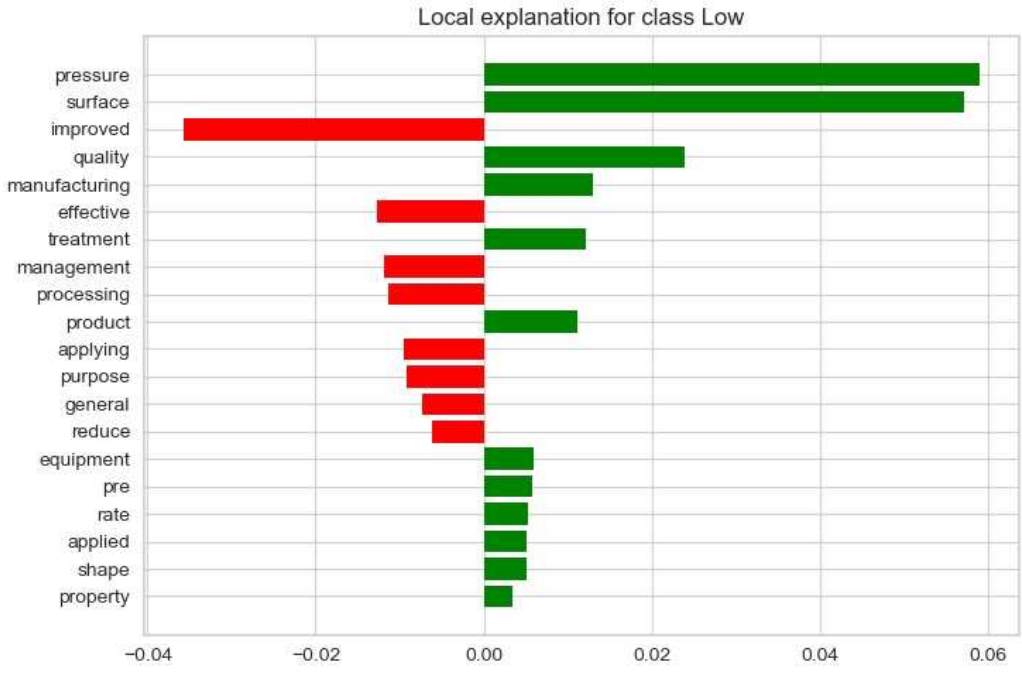
3) Low 단계

다음은 Low 단계로 분류된 어떤 데이터에 대하여 Lime 알고리즘을 이용하여 분류의 근거를 설명한 것이다. 예측 설명에서 보면 “pressure”나 “surface”와 같은 단어가 Low 단계로 분류되는데 긍정적 역할을 하고 있고, “improved” 등을 포함한 단어가 부정적 역할을 하고 있음을 알 수 있다.

<표 5-21> Low 단계 판단에 대한 각 단어의 가중치

키워드	수치	키워드	수치
pressure	0.0590075800	applying	-0.0094257811
surface	0.0571764849	purpose	-0.0091321177
improved	-0.0356741240	general	-0.0072524397
quality	0.0239722605	reduce	-0.0061681157
manufacturing	0.0130632380	equipment	0.0059466727
effective	-0.0127704719	pre	0.0058260425
treatment	0.0121673984	rate	0.0053487635
management	-0.0117756953	applied	0.0051853217
processing	-0.0113881923	shape	0.0051742794
product	0.0111785289	property	0.0033539172
Sum of positives	0.2074004876	Sum of negatives	-0.1035869377

[그림 5-19] Low 단계 예측 데이터에 대한 설명



Low 단계에 속하는 데이터는 Med*나 High 단계와 달리 어떤 특징으로 인해 Low 단계로 분류되기보다는 긍정적 요인과 부정적 요인이 큰 차이 없이 골고루 포함되는 경우가 대부분이다. 즉, Med*나 High 단계로 분류되기 어려운 데이터가 결국은 Low 단계로 분류된다는 의미로 해석할 수 있으며, 이는 본 연구의 분류 목적과도 일치한다고 볼 수 있다.

제6장 SW 융합 R&D의 유형별 분석 및 시사점

제1절 개요

본 장에서는 tf-idf와 Linear SVM 모델을 활용한 자동 분류기를 활용하여 2017년 국가연구개발과제 중 예측 결과가 존재하는 총 55,566건의 과제의 예측 결과를 유형별로 나누어 현황을 분석하고자 한다. 유형별 분석을 위해 Business Intelligence 툴[15]을 활용하여 시각화한다. 살펴보고자 하는 유형별 현황은 SW 융합 R&D 단계별(3단계 혹은 5단계로 구분하여 분석)로 수행 부처, 국가과학기술표준분류별, 연구개발의 단계, 연구개발의 수행 주체, 연구책임자의 전공 등의 구분에 따른 분석을 진행하여 각 특징을 살펴보고자 한다. 기본적으로 유형별로 전체 국가연구개발사업의 특징과의 비교 분석을 위해 과제 수와 총 연구비 합계, 평균의 통계 등을 제시한 뒤, 과제 수와 총 연구비 측면에서 SW 융합 R&D 단계의 특징을 보여준다. 이러한 현황 분석 자료를 활용하여 SW 융합 촉진을 위한 정부 차원의 SW 융합 R&D의 방향성을 제시할 수 있다.

제2절 주요 유형별 분석 내용

1. 수행 부처별

2017년에 진행된 총 55,556건의 국가연구개발사업과제는 23개의 부처에서 수행되었다. 각 부처에서 수행된 과제의 개수 및 총 연구비 합계 및 평균은 아래 <표 6-1>과 같다. 과제 전반에 대한 일반적인 상황을 살펴보면, 과학기술정보통신부(이하 과기정통부)의 과제 수와 총 연구비가 가장 많으며, 과제 수는 15,505건이고 총 연구비의 합계는 6조 3,596억 원이다. 산업통상자원부(이하 산업부)의 경우, 과제 수 기준으로 4,564건이나 총 연구비가 4조 6,320억 원으로 과제의 평균 총 연구비가 약 10억 원인 것을 알 수 있다. 국가연구개발과제의 총 연구비 평균은 약 3억 원으로 산업부의 과제당 평균 총 연구비가 이보다 약 3배 이상 높은 편이다.

15) Microsoft Power BI Desktop 프로그램을 활용

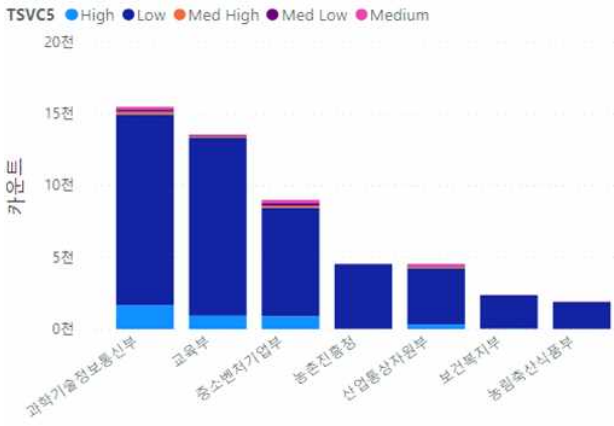
<표 6-1> 부처별 총 연구비 합계 및 과제 수, 평균 금액

부처명	총연구비 합계(원)	과제 수(건)	과제 평균 금액(원)
과학기술정보통신부	6,359,607,124,981	15,505	410,164,923
산업통상자원부	4,631,998,522,672	4,564	1,014,898,888
중소벤처기업부	1,714,635,946,815	9,015	190,198,108
교육부	937,545,632,695	13,591	68,982,829
국토교통부	575,778,889,978	662	869,756,631
보건복지부	529,462,938,381	2,396	220,977,854
해양수산부	529,081,141,333	1,048	504,848,417
농촌진흥청	333,000,550,000	4,569	72,882,589
환경부	272,326,116,199	664	410,129,693
농림축산식품부	232,391,142,570	1,944	119,542,769
기상청	118,080,716,000	188	628,088,915
문화체육관광부	87,687,497,000	119	736,869,723
산림청	84,981,615,000	323	263,100,975
식품의약품안전처	81,600,616,266	588	138,776,558
원자력안전위원회	50,407,038,000	108	466,731,833
행정안전부	34,325,413,571	100	343,254,136
소방청	20,641,521,420	63	327,643,197
문화재청	15,929,544,000	48	331,865,500
해양경찰청	10,587,967,000	29	365,102,310
경찰청	7,793,756,000	36	216,493,222
기획재정부	3,205,000,000	2	1,602,500,000
특허청	2,160,000,000	1	2,160,000,000
국무조정실	1,876,000,000	3	625,333,333
합계	**16,635,104,689,881**	**55,566**	**299,375,602**

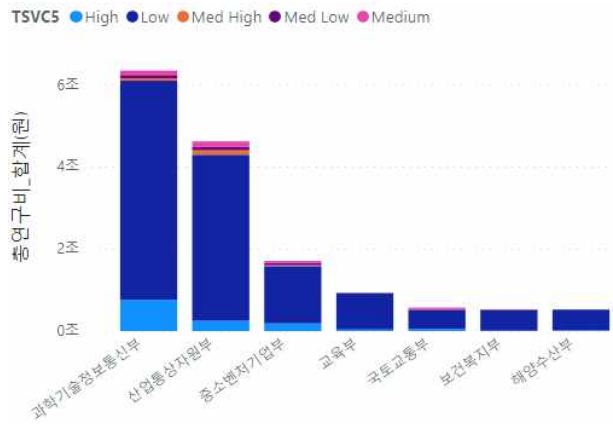
각 부처의 5단계 구분에 의한 분류 결과에서 과제 수 기준의 상위 7개 부처에 대한 내용은 [그림 6-1]과 같고, 총 연구비[16] 합계 기준은 [그림 6-2]에서 확인할 수 있다.

16) 총 연구비는 정부연구비와 매칭펀드 금액을 합한 금액으로, 항목별로는 현금과 현물에 해당하는 인건비, 직접비, 간접비, 위탁연구비를 포함함

[그림 6-1] 부처별 5단계 구분 결과(과제 수 기준, 상위 7개)



[그림 6-2] 부처별 5단계 구분 결과(총 연구비 기준, 상위 7개)



부처별 과제 수 기준과 총 연구비 기준으로 SW 융합 R&D 현황을 살펴보면 [그림 6-3]과 [그림 6-5], [그림 6-11]과 같이 과제 수 기준으로 High와 Med High 그리고 Low 단계는 과기정통부에서 수행한 과제가 가장 많다. 총 연구비 기준은 과기정통부가 High와 Med Low, 그리고 Low 단계에서 비중이 가장 높은 것으로 나타났다. 특징적인 것은 과기정통부가 High 단계에서 과제 수 기준으로 2번째 많은 비중을 차지하는 교육부에 비해 약 1.8배 정도 많은 과제를 진행하고 있고, 총 연구비 측면에서는 2번째로 연구비가 많은 산업부에 비해 약 3배 정도 큰 금액을 차지하고 있다는 것이다. 이를 통해 과기정통부가 타 부처에 비해 SW 융합 R&D 활동을 80% 이상 포함하는 단계인 High 단계의 과제를 가장 많이 활발하게 수행하는 주체라는 것을 나타낸다고 판단할 수 있다.

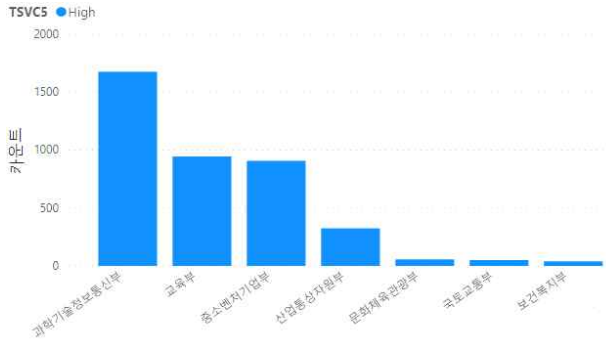
중소기업벤처부(이하 중기부)의 경우, Medium과 Med Low 단계에서 수행하는 과제 수가 가장 많은 비중을 차지하고 있다는 특징을 보였다. 총 연구비 기준에서는 5가지 모든 단계에서 3번째로 높은 비중을 차지하였다. 과제 수는 가장 많으나 총 연구비 측면에서 차지하는 비중이 낮은 이유는 중기부의 과제당 평균 연구비는 약 1억 9천만 원으로 평균 연구비 규모가 작기 때문에 발생하는 현상으로 판단된다. 한편, 교육부가 수행하는 과제 수는 전체 국가연구개발사업에서 두 번째로 많은 과제를 수행하고 있으며, SW 융합 R&D 단계에서도 과제 수 기준으로 High와 Low 단계에서 두 번째로 큰 비중을 차지하는 부처로 나타났다. 총 연구비 기준으로는 단계별로 4번째 혹은 6, 7번째를 차지하고 있다. 즉, 교육부는 수행하는 SW 융합 R&D 과제 수는 많으나 과제당 총 연구비가 적기 때문

에 연구비 기준으로는 큰 비중을 차지하지 않는 것으로 나타났다.

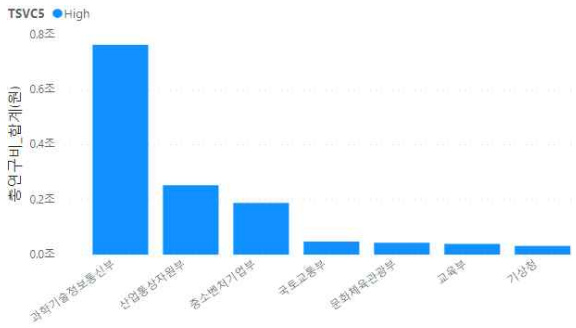
산업부는 Med High, Medium과 Med Low 단계에서 과제 수가 3번째로 높은 것으로 나타났으나 총 연구비 기준으로는 Med High와 Medium 단계에서 가장 높은 비중을 차지한 것으로 조사되었다. High, Med Low, Low 단계에서도 총 연구비 기준으로 산업부가 두 번째로 큰 비중을 차지하고 있다. 이 부분은 산업부 과제의 평균 연구비가 10억 원으로 전체의 평균 금액과 약 3배 이상의 차이를 보이기 때문인 것으로 추측된다.

정리하면, 과기정통부는 SW 융합 R&D 활동이 80% 이상인 High 단계 과제를 가장 많이 수행하고 있는 부처로 나타났고, 산업부는 총 연구비 기준으로 SW 융합 R&D 활동이 40~80% 사이의 과제를 가장 활발히 수행하는 부처로 조사되었다. 또한, 중기부는 과제 수 기준으로 SW 융합 R&D 활동이 20~60% 정도인 과제를 가장 많이 수행하는 부처로 나타났다.

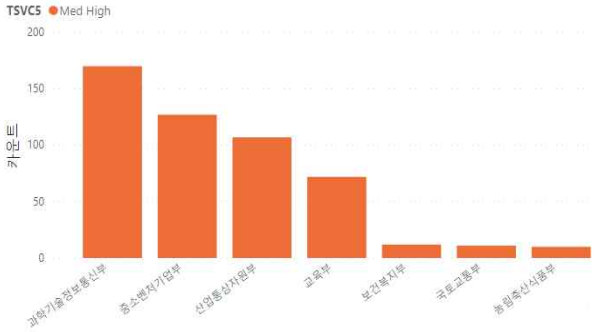
[그림 6-3] High-부처(과제 수)



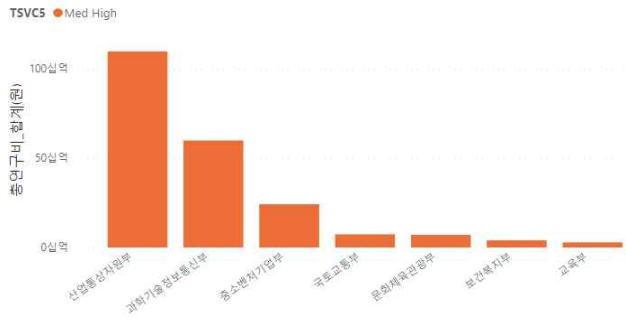
[그림 6-4] High-부처(연구비)



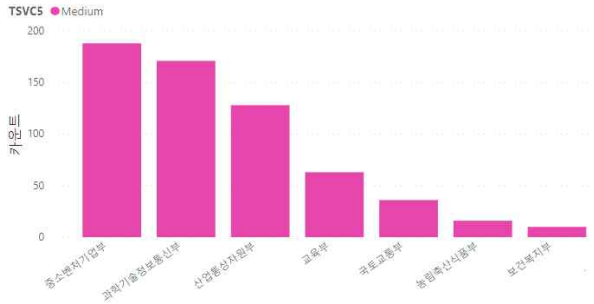
[그림 6-5] Med High-부처(과제 수)



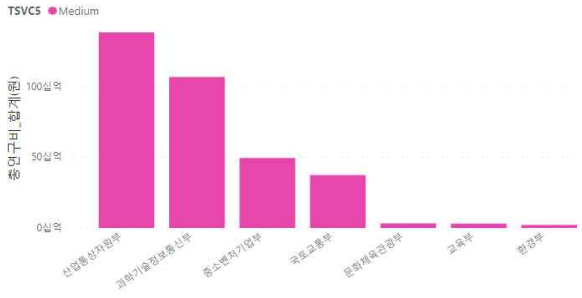
[그림 6-6] Med High-부처(연구비)



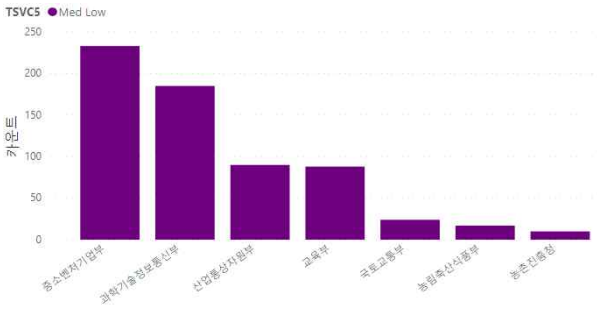
[그림 6-7] Medium-부처(과제 수)



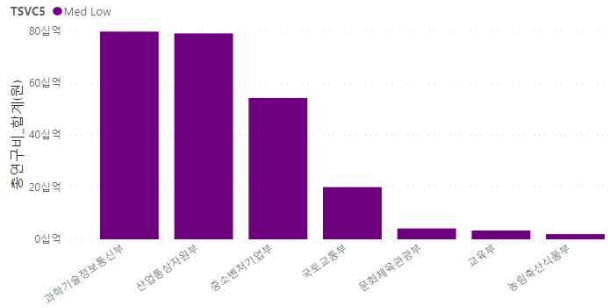
[그림 6-8] Medium-부처(연구비)



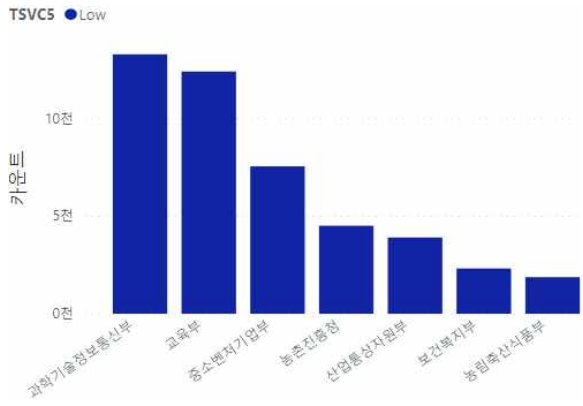
[그림 6-9] Med Low-부처(과제 수)



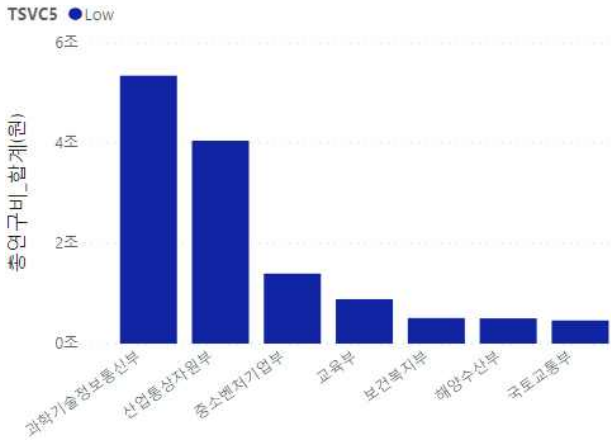
[그림 6-10] Med Low-부처(연구비)



[그림 6-11] Low-부처(과제 수)

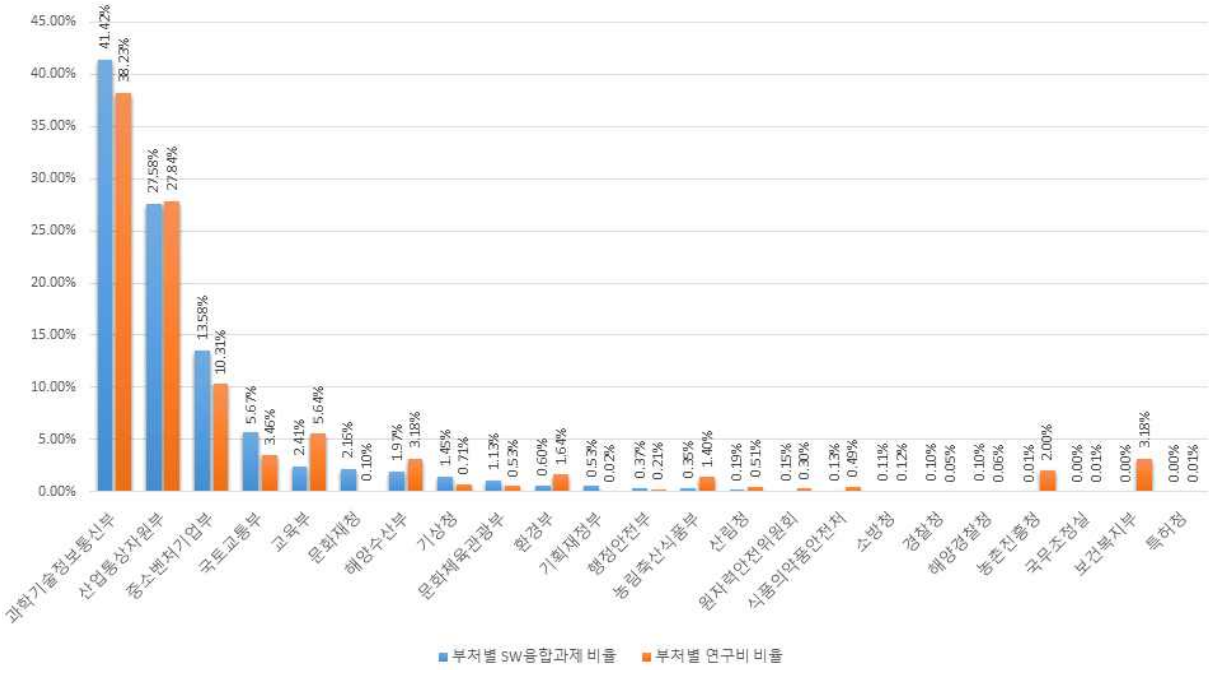


[그림 6-12] Low-부처(연구비)



국가연구개발과제를 수행한 23개 부처에 대해서 부처별 SW 융합 과제 비율(SW 융합 R&D 과제의 총 연구비 중 해당 부처의 총 연구비 비율)과 부처별 총 연구비 비율(국가연구개발사업의 총 연구비 중 해당 부처의 총 연구비 비율) 현황을 분석한 결과는 [그림 6-13]와 같다.

[그림 6-13] 부처별 연구비 및 SW 융합 과제 현황



SW 융합 R&D의 총 연구비 중 부처별 비중은 과기정통부, 산업부, 중기부 순으로 높은 것으로 나타나고 있으나 그 외 20개 부처의 경우, SW 융합 R&D 비율이 10% 미만으로 매우 저조한 것으로 나타나고 있다. 특히, 연구비 비중이 국가 전체의 총 연구비의 5%를 넘는 교육부의 경우, SW 융합 비중이 2.4% 정도로 상당히 저조한 것으로 나타났다. 농촌진흥청, 농림축산식품부와 보건복지부 역시 국가연구개발사업에서 차지하는 비중에 비해 SW 융합 R&D에서 차지하는 비중이 적은 것을 확인하였다. 또한, 과기정통부, 산업부, 중기부 3개 부처의 총 예산이 전체 예산의 76.38%를 차지하고 있고, SW 융합 비중에서는 82.58%로 나타났다. 위 3개 부처의 SW 융합 R&D 비중의 평균은 27.53%로 조사되었다. 그 외 23.62%의 연구비를 사용하는 나머지 모든 부처의 SW 융합 정도는 11.76%에 불과한 것으로 나타나 SW 융합 R&D 과제가 과기정통부, 산업부와 중기부 3개 부처에 쏠림 현상이 존재한다는 것을 알 수 있다.

2. 연구개발 주체별

2017년에 수행된 국가연구개발과제의 수행 주체에 따른 과제 수와 총 연구비 구성은 아래 <표 6-2>과 같다. 출연연구소가 총 연구비 기준으로 가장 높은

비중을 차지하고 있고, 과제 수 기준으로는 대학이 수행한 연구가 가장 많은 것으로 나타났다. 출연연구소에서 수행된 과제의 평균 금액은 약 10억 6천만 원이고, 대학은 약 1억 3천만 원으로 나타났다.

<표 6-2> 연구개발 수행 주체별 총 연구비 합계 및 과제 수, 평균 금액

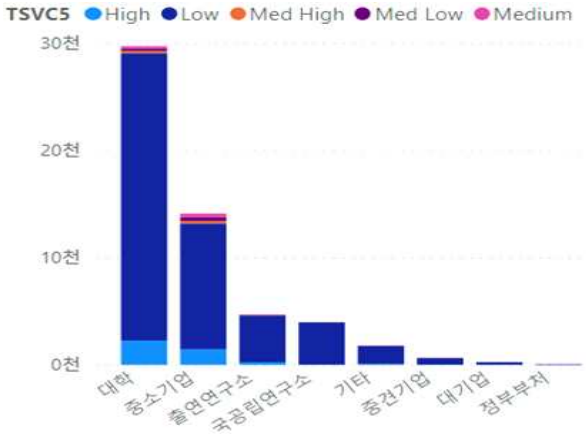
연구개발 주체	총 연구비 합계(원)	과제 수(건)	과제 평균 금액(원)	비중(%)[17]
출연연구소	5,032,131,429,720	4,740	1,061,631,103	30.3
중소기업	4,500,096,550,905	14,159	317,825,874	27.1
대학	3,923,731,027,213	29,811	131,620,242	23.6
기타	1,259,316,680,527	1,823	690,793,571	7.6
중견기업	786,669,884,784	685	1,148,423,189	4.7
대기업	523,231,324,711	269	1,945,097,861	3.1
국공립연구소	465,348,292,021	4,004	116,220,852	2.8
정부부처	144,579,500,000	75	1,927,726,667	0.9
전체	16,635,104,689,881	55,566	299,375,602	100

중소기업의 경우, 총 연구비 기준에서 Low 단계를 제외한 모든 SW 융합 R&D 단계(High, Med High, Medium, Med Low)에서 과제를 가장 많이 수행하고 있는 주체로 나타났다[18]. 또한, 중소기업은 과제 수 기준에서 Medium과 Med Low 단계의 과제를 가장 많이 수행하는 주체로 조사되었고, Med High 단계는 대학과 비슷한 수준으로 SW 융합 R&D 과제를 수행하고 있는 것으로 나타났다.

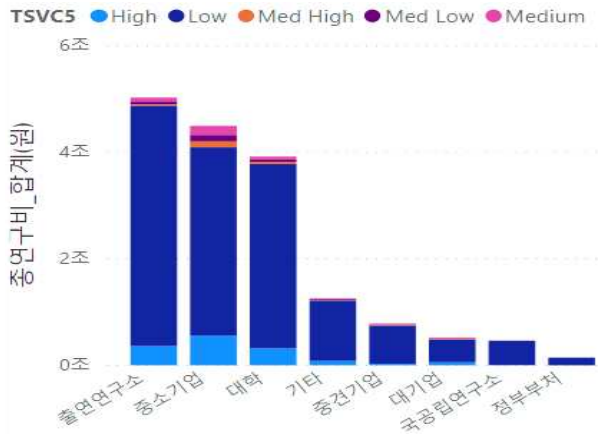
대학은 과제 평균 금액이 적으므로 총 연구비 측면에서 SW 융합 R&D 과제의 수행 비중이 높은 것은 아니지만, 과제 수 기준으로는 High와 Med High 단계에서 SW 융합 R&D 과제를 가장 많이 수행하는 주체로 나타났다. 한편, 출연연구소는 총 연구비 기준으로 전체 국가연구개발과제 중 30% 정도를 수행하고 있는 것으로 나타났으나 이에 비해 SW 융합 R&D 과제를 적극적으로 수행하는 주체는 아닌 것으로 나타났다.

17) 총 연구비 기준 비중
18) 5개 단계별 상세한 현황은 부록의 연구개발(R&D) 단계 - SW 융합 R&D 5단계 구분 현황 참조

[그림 6-14] 연구수행 주체-5가지 단계(과제 수)



[그림 6-15] 연구수행 주체-5가지 단계(연구비)

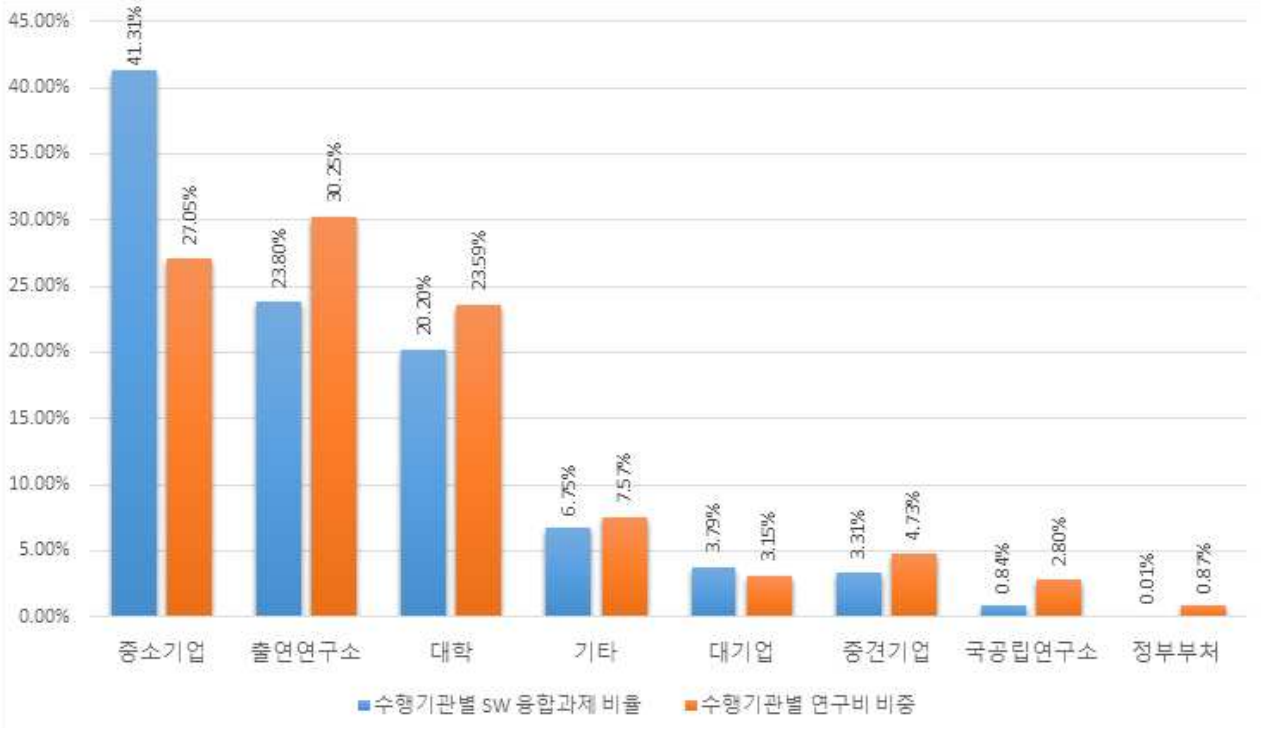


국가연구개발과제의 수행 주체에 대해서 주체별 SW 융합 과제 비율(SW 융합 R&D 과제의 전체 연구비 중 해당 주체의 총 연구비 비율)과 기관별 연구비 비율(국가 전체의 총 연구비 중 해당 주체의 총 연구비 비율) 현황을 비교한 결과는 아래 [그림 6-16]와 같이 나타났다.

SW 융합 과제를 주로 수행하는 주체는 중소기업, 출연연구소, 대학 순으로 나타나고 있으며, 대기업, 중견기업, 국공립연구소 등은 SW 융합 과제에서 차지하는 비중이 5% 미만으로 매우 저조한 것으로 나타났다. 국가 전체 연구개발과제에 대한 총 연구비 비중은 출연연구소, 대학, 중소기업 순으로 많이 사용하고 있으나 SW 융합 과제는 중소기업이 수행하는 연구비가 가장 높고, 출연연구소, 대학의 순서를 보여 국가연구개발사업과 차이를 보이고 있다. 이러한 현상은 출연연구소와 대학에서 나온 연구 성과 중 중소기업을 통해 사업화로 이어지는 데 있어 SW 융합 비중이 적은 것으로도 볼 수 있다.

또한, 국가 연구개발과제에 비해 SW 융합 R&D 과제의 수행 주체가 핵심 3개 주체에 치중되어 있음을 알 수 있다. 중소기업과 출연연구소, 대학이 수행하는 총 연구비는 국가 전체 예산의 80.89%에 달하나 세 주체가 수행하는 SW 융합 R&D의 총 연구비 중 85.31%를 차지하고 있기 때문이다. 그 외 19.12%의 연구비를 사용하는 나머지 모든 기관들의 SW 융합 R&D의 총 연구비는 전체의 14.70%를 비중에 불과한 것으로 조사되었다.

[그림 6-16] 연구수행 주체별 연구비 및 SW 융합 과제 현황



3. 연구개발(R&D) 단계

2017년 국가연구개발과제의 연구개발 단계별 과제 수와 총 연구비의 구성은 <표 6-3>에서 확인할 수 있다. 개발 단계의 연구는 총 연구비 기준으로 41.8% 정도로 가장 높은 비중을 차지하고 있고, 과제 수 기준은 기초 단계의 연구가 전체 연구개발과제의 44.6%를 차지하고 있어 가장 비중이 높은 것을 알 수 있다.

<표 6-3> 연구개발 단계별 총 연구비 합계 및 과제 수, 평균 금액

연구개발 단계	총 연구비 합계(원)	과제 수(건)	과제 평균 금액(원)	비중(%)[19]
개발연구	6,957,239,332,238	17,767	391,582,109	41.8
기초연구	4,648,375,961,989	24,796	187,464,751	27.9
기타	2,593,628,567,712	7,110	364,786,015	15.6
응용연구	2,435,860,827,942	5,893	413,348,181	14.6
합계	16,635,104,689,881	55,566	299,375,602	100

19) 총 연구비 기준 비중

연구개발 단계별 SW 융합 R&D 현황은 아래 [그림 6-17]와 [그림 6-18]에서 알 수 있듯이, 과제 수 기준으로 기초연구는 SW 융합 R&D 연구 단계 중 High와 Low 단계에서 가장 큰 비중을 차지하고 있고, 총 연구비 기준에서는 개발연구가 5개의 모든 단계에서 가장 큰 비중을 차지하고 있다는 것을 알 수 있다. 중간의 세 단계(Med High, Medium, Med Low)에서는 개발 단계의 연구가 과제 수와 총 연구비 기준 모두 가장 큰 비중을 차지하였다[20]. 중간 단계는 총 연구비 기준으로 개발 단계와 기초 단계의 비중 차이가 약 3배 정도 크게 차이나는 것으로 나타났고, High 단계 역시 평균적인 개발연구와 기초연구의 차이보다 더 큰 차이를 보이는 것을 알 수 있다.

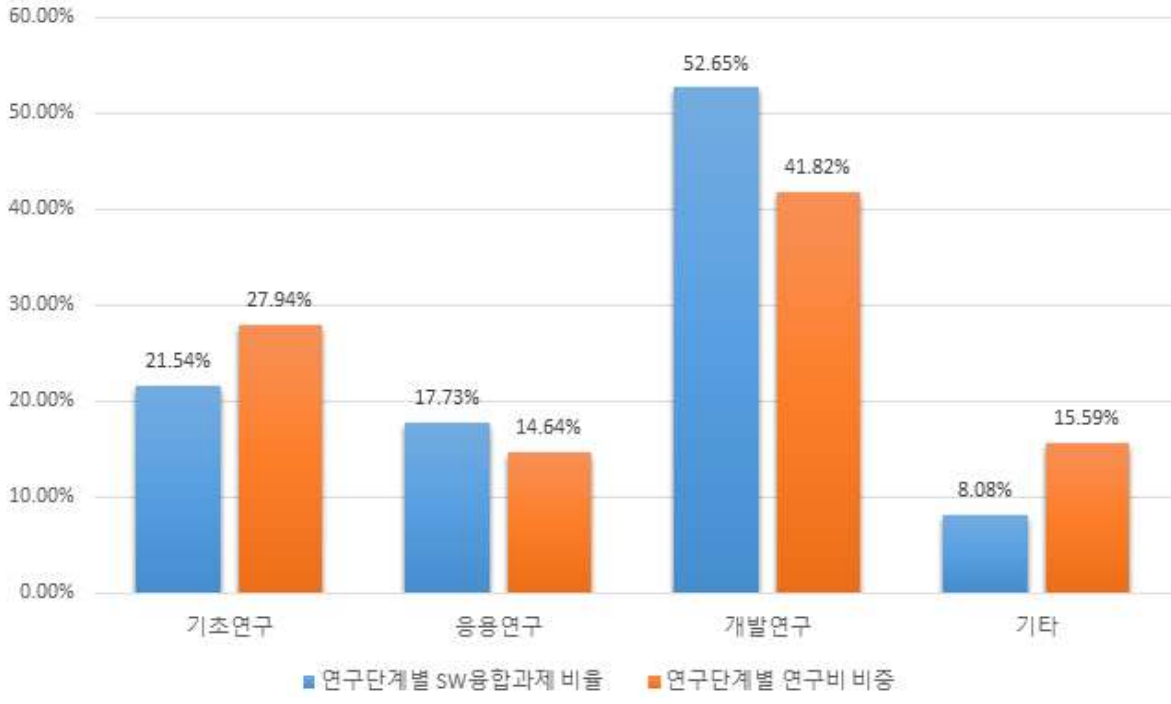
[그림 6-17] 5가지 단계-R&D 단계 (과제 수)

[그림 6-18] 5가지 단계-R&D 단계 (연구비)

국가연구개발과제를 연구단계별 SW 융합 과제 비율(SW 융합 과제의 전체 총 연구비 중 해당 단계의 총 연구비 비율)과 연구단계별 연구비 비율(국가 전체의 총 연구비 중 해당 단계의 총 연구비 비율) 현황을 분석한 결과 [그림 6-19]과 같이 나타난다.

20) 5개 단계별 상세한 현황은 부록의 연구개발(R&D) 단계-SW 융합 R&D 5단계 구분 현황 참조

[그림 6-19] 연구개발 단계별 연구비 및 SW 융합 과제 현황



국가연구개발의 총 연구비 중 단계별 비중은 개발연구, 기초연구, 응용연구 순으로 높게 나타났으며, SW 융합 R&D의 총 연구비에 대한 단계별 비중 또한 동일한 순서로 나타났다. SW 융합 R&D의 경우, 개발연구가 52% 이상을 차지하고 있으며, 이는 SW 연구개발 특성과 일치하는 현상인 것으로 판단된다. 이는 대부분의 SW R&D 과제가 기초 단계에 해당하는 새로운 사실의 발견보다는 제품이나 서비스의 창출을 목표로 하며, 목표 시스템의 개발에 역점을 두고 있기 때문[21]인 것으로 보인다(김진형, 2011).

4. 과학기술표준분류체계 (대분류 기준)

과학기술표준분류체계 대분류를 기준으로 살펴보면, 2017년의 국가연구개발 과제(총 55,566건)는 아래 <표 6-4>와 같이 기계 분야의 총 연구비 합계가 가장 높고, 보건의료 분야가 과제 수 기준으로 가장 많은 과제를 수행하는 것으로 조사되었다.

21) SW R&D체계 개편방안 연구, 한국과학기술원, 2011.

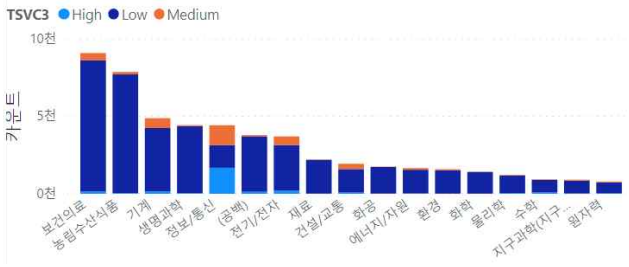
<표 6-4> 과학기술표준분류별 총 연구비 합계 및 과제 수, 평균 금액

과학기술표준분류1-대	총연구비 합계(원)	과제 수(건)	과제 당 평균 연구비(원)
기계	2,866,989,114,575	4,888	586,536,235
정보/통신	1,850,143,498,323	4,433	417,356,981
보건의료	1,778,963,934,898	9,096	195,576,510
전기/전자	1,221,804,562,988	3,712	329,149,936
에너지/자원	1,040,169,818,986	1,662	625,854,283
건설/교통	927,920,284,890	1,940	478,309,425
생명과학	911,933,562,572	4,439	205,436,712
농림수산식품	884,175,773,852	7,878	112,233,533
재료	833,850,045,267	2,221	375,439,012
지구과학(지구/대기/해양/천문)	592,006,039,200	915	647,001,136
화공	551,796,440,721	1,758	313,877,384
원자력	545,007,634,091	797	683,823,882
환경	474,057,640,742	1,592	297,774,900
과학기술과 인문사회	470,521,475,013	491	958,292,210
물리학	415,092,243,990	1,222	339,682,687
화학	336,198,214,892	1,422	236,426,311
문화/예술/체육	221,387,101,853	549	403,255,195
공백	127,318,926,000	3,794	33,557,967
뇌과학	124,608,667,302	489	254,823,451
경제/경영	117,040,372,131	327	357,921,627
교육	88,732,811,055	202	439,271,342
수학	75,208,950,862	939	80,094,729
미디어/커뮤니케이션/문헌정보	72,643,994,254	155	468,670,931
정치/행정	28,072,547,994	67	418,993,254
생활	21,599,417,974	131	164,881,053
사회/인류/복지/여성	16,134,258,500	88	183,343,847
인지/감성과학	13,001,179,701	97	134,032,780
지리/지역/관광	11,404,027,240	75	152,053,697
법	4,626,110,767	29	159,521,061
언어	3,601,910,000	37	97,348,919
문학	3,159,838,500	38	83,153,645
심리	3,106,293,248	34	91,361,566
철학/종교	1,524,262,000	30	50,808,733
역사/고고학	1,303,735,500	19	68,617,658
합계	16,635,104,689,881	55,566	299,375,602

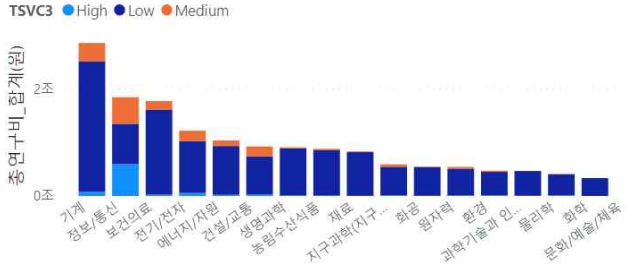
과학기술표준분류체계의 경우, SW 융합 R&D 구분을 5단계와 3단계(High,

Medium, Low)로 구분한 결과를 모두 살펴보았을 때, Med High와 Medium, Med Low 단계에서 분석상의 큰 차이가 없으므로 총 3단계 구분으로 분석한다. 확인 결과, 정보/통신 분야에서는 과제 수와 총 연구비 기준 모두 High와 Medium 단계의 과제를 가장 많이 수행하는 것으로 나타났다. 특징적인 부분은 High 단계(80% 이상의 SW R&D 활동을 포함)에서 두 번째로 큰 비중을 차지하는 기계 분야에 비해 정보/통신 분야 과제가 약 7.8배 정도의 큰 비중을 차지한다는 것이다. Medium 단계의 경우에도 정보/통신 분야의 과제 수가 가장 많으나 기계 분야에 비해 약 1.4배 정도가 많은 것으로 나타나 상대적으로 큰 차이를 보이지 않는 것을 알 수 있다. 그리고 Low 단계에서는 과제 수 기준으로 보건의료 분야에서 가장 많은 과제를 수행하고 있고, 총 연구비 기준으로는 기계 분야가 가장 큰 비중을 차지하였다. 한편, 생명과학(0.7%)과 농림수산식품(0.7%), 원자력(1.4%), 보건의료(1.4%), 환경(2.0%)과 에너지/자원(2.1%) 등의 분야의 경우, 수행하고 있는 과제의 총 연구비에 비해 SW 융합도가 높은 High 단계의 연구개발 금액의 비중이 비교적 적다는 특성을 보였다.

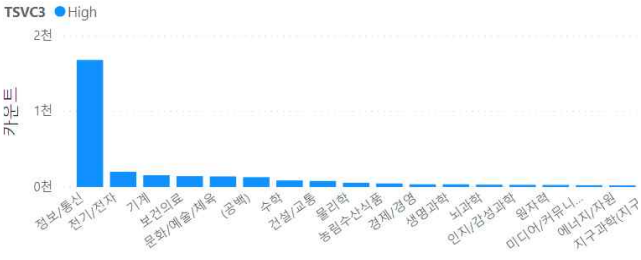
[그림 6-20] 과학기술표준분류-3단계(과제 수)



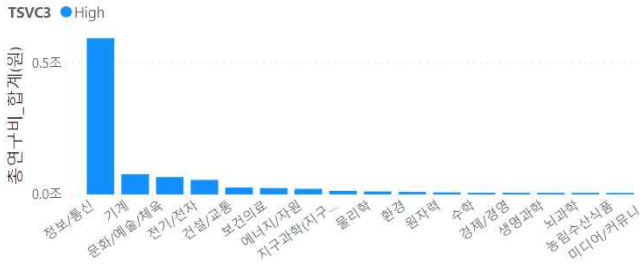
[그림 6-21] 과학기술표준분류-3단계(연구비)



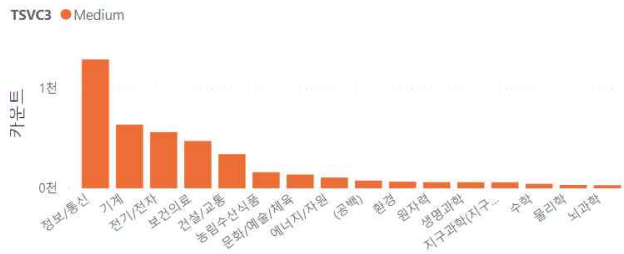
[그림 6-22] 과학기술표준분류-High(과제 수)



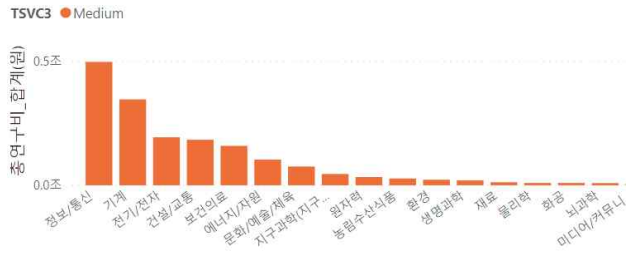
[그림 6-23] 과학기술표준분류-High(연구비)



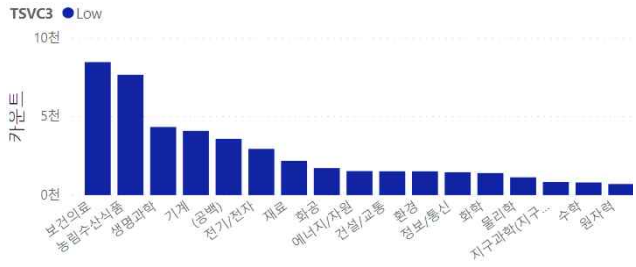
[그림 6-24] 과학기술표준분류-Medium(과제 수)



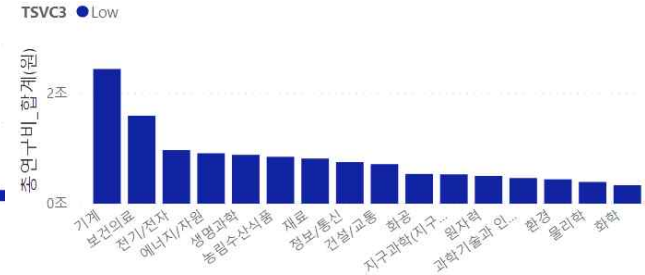
[그림 6-25] 과학기술표준분류-Medium(연구비)



[그림 6-26] 과학기술표준분류-Low(과제 수)

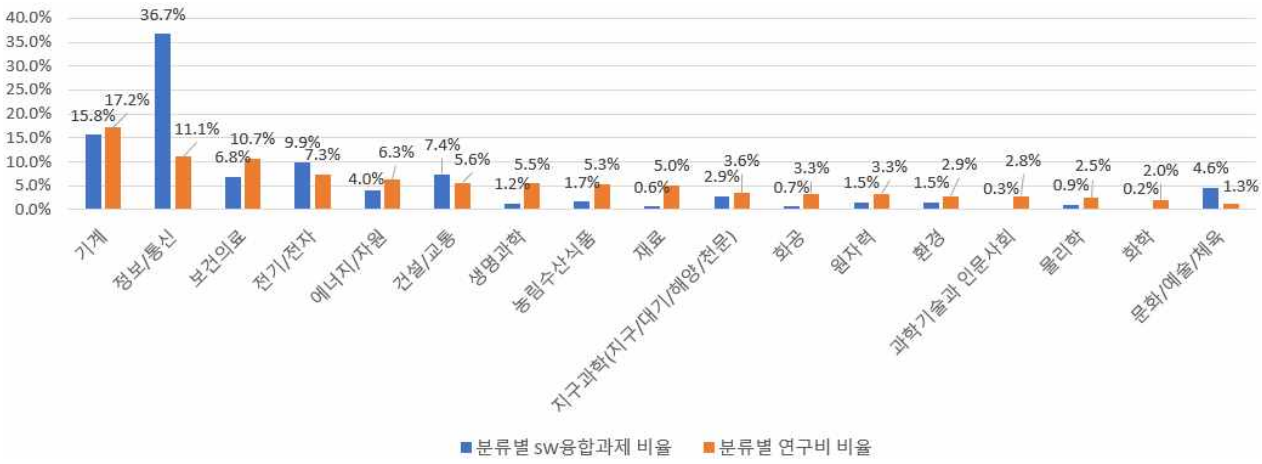


[그림 6-27] 과학기술표준분류-Low(연구비)



과학기술표준대분류별 SW 융합 과제 비율(SW 융합 과제 총 연구비 중 분류별 비율)과 연구비 비율(국가 총 연구비 중 해당 분류의 연구비 비율) 현황을 분석한 결과 [그림 6-28]과 같다.

[그림 6-28] 과학기술표준분류별 연구비 및 SW 융합 과제 현황



정보/통신 분야는 국가연구개발사업에서 차지하는 비중(11.1%)에 비해 SW 융합 R&D 과제 전체의 연구비에서 차지하는 비중이 3배 이상 높은 것(36.7%)으로 가장 큰 차이를 보였다. 또한, 전기/전자와 건설/교통, 문화/예술/체육 분야 역시

국가연구개발사업에서 차지하는 분야별 비율에 비해 SW 융합 R&D 비율이 높은 것으로 나타나 SW 융합이 잘 이루어지고 있는 분야라고 판단할 수 있다. 한편, 생명과학과 농림수산식품, 재료, 화공 등의 분야는 국가연구개발의 총 연구비에서 차지하는 분야의 비중에 비해 SW 융합 정도가 현저히 떨어지는 것으로 조사되었다.

제3절 SW 융합 현상에 따른 시사점

1. SW 융합 현상 분석에 대한 시사점

본 장에서 분석한 국가연구개발과제의 부처별, 기관별 총 연구비 비중과 SW 융합 R&D의 수준을 살펴보면 다음과 같은 시사점을 얻을 수 있다.

국가개발연구과제의 사업비 중 76.4% 정도를 사용하는 주요 3개 부처(과학기술정보통신부, 산업통상자원부, 중소벤처기업부)가 SW 융합 R&D의 약 83%를 수행하는 것으로 조사되었다. 또한, 주요 3개 부처는 평균적으로 27.5% 정도의 SW 융합 수준을 보였다.

그러나 제4차 산업혁명 시대에 대응하여 전 부처의 SW 융합이 강조되어야 함에도 불구하고 주요 부처를 제외한 나머지 정부 부처의 연구개발사업의 SW 융합 정도는 1% 정도 수준으로 나타나, 예산 규모가 있고 SW 융합이 중요할 것으로 보이는 방위사업청, 교육부, 해양수산부, 농촌진흥청, 보건복지부의 SW 융합 정도를 상향하기 위한 시범 사업이나 협업 프로젝트 추진 등의 정책적 판단이 필요해 보인다. 예를 들면, 구글의 모기업인 알파벳에서 진행하는 것과 같이 드론과 AI 기술을 활용하여 기후 변화나 주변 환경을 분석한 뒤 해충이나 재해로부터 식량 생산 관리를 자동으로 하는 SW 융합 R&D 사업 등을 고려할 수 있다.

국가연구개발과제의 총 연구비는 출연연구소, 대학, 중소기업에 약 81% 정도가 할당되어 있으나 위 연구 주체들의 SW 융합의 총 연구비 중 약 85%에 달해, 대기업이나 중견기업, 국공립연구소 등에서 수행하는 SW 융합 과제가 상대적으로 매우 적은 것을 알 수 있다.

특히, 국가연구개발과제의 30% 이상의 연구비를 사용하는 출연연구소의 SW 융합 정도가 23% 수준에 머무는 반면, 중소기업의 SW 융합 R&D 비중은 40%를 넘고 있어 중소기업 위주의 SW 융합 R&D 과제가 주를 이루는 것을 알 수 있다. 그러므로 주로 대형·중장기 연구를 수행하고 있는 출연연구소의 SW 융합 R&D 과제의 성과를 중소기업의 사업화로 연결하기 위한 노력이 필요할 것으로 보인다.

2. SW 융합 현상 분석에 대한 정책적 함의

2017년 국가연구개발사업의 예측 및 현황 분석을 통해 파악한 시사점을 바탕으로 국가연구개발사업과 SW 융합 정도의 향상을 위한 정책 수립에 활용할 수 있는 제언을 다음과 같이 도출하였다.

전반적으로 전 산업 분야에 있어 AI·SW 중심의 융합이 가속화되고 있는 시점에서 현재의 SW 융합 정도는 14.67%로 조사되었다. 특히, 주요 3개 부처 외의 SW 융합 정도는 1% 수준으로 각 분야의 기술경쟁력 약화를 초래할 것으로 보인다. 따라서 국가연구개발과제 전체의 SW 융합 정도를 제고할 수 있는 정책 수립이 필요하며, 1% 대에 머물고 있는 정부 부처의 SW 융합 정도 향상을 위해서는 전체 SW 융합 R&D 방향을 조절할 수 있는 컨트롤타워 역할 수립이 필요한 것으로 보인다.

또한, 출연연구소의 중요한 역할 중 하나는 미래 유망 기술을 선도적으로 연구 개발하여 사업화로 이어질 수 있도록 하는 것인데 SW 융합 분야에 있어 중소기업의 SW 융합 정도의 절반 수준에 머물러 있어 SW 융합 연구개발 생태계에 불균형을 초래하고 있는 것으로 판단된다. 그러므로 출연연구소의 연구개발 정책 수립에 산업체의 SW 융합도 수요를 적극 반영하여 균형적인 연구개발 생태계 수립이 시급한 것으로 보인다. 또한, 출연연구소의 연구 성과가 SW 융합 과제의 핵심 주체인 중소기업을 통해 사업화로 연결되기 위한 제도적 지원이 필요할 것으로 판단된다.

제4절 SW 융합 R&D 정책 방향 제시

1. 개요

제4차 산업혁명 시대에는 산업의 고부가가치와 신산업 창출을 위한 핵심 SW 기술의 확보가 경쟁력이다. 제4차 산업혁명을 위해 독일은 인더스트리 4.0, 미국은 산업 인터넷, 일본은 로봇 신전략, 중국은 제조 2025 등 산업의 SW 융합 기반 국가 전략을 수립하였다. 미국의 전통제조업체인 GE는 SW 융합 기업으로 변신을 통해 산업의 고부가가치화를 추구하고, 선진 ICT 기업인 애플과 구글은 축적된 SW 융합 기술로 자동차 산업에 뛰어들어 산업간 융합을 선도하고 있다. 이러한 현상으로 살펴보면, 그 간의 산업은 대기업 주도로 고도성장을 위한 신속한 기술의 확보에 초점이 맞춰졌으나, 산업별 핵심 SW에 대한 외산 의존도와 국산 SW 비중은 개선되지 않아 국가 경제 성장을 견인할 부가가치와 신산업 창출에 한계를 보였다.

산업별 핵심 SW 기술은 SW 융합화를 위한 핵심 엔진이며 이에 대한 육성을 통해 부가가치와 신산업을 창출할 국가 차원의 전략이 필요하다. SW 융합화를 위한 핵심 SW 기술은 산업별 부가가치 경쟁력을 제공하는 기술로 SW 활용 비율 향상, 외산 SW 대체, SW 접목에 의한 산업 확장을 가능하게 하는 SW 기술을 말한다. 국내 중소기업들이 제4차 산업혁명 시대를 위한 경쟁력을 갖추어 세계시장으로 진출하고 신산업 창출이 가능하도록 지원할 필요가 있다. 그러나 대기업 의존적인 영세 중소기업의 산업별 핵심 SW 기술 확보는 높은 기술 장벽과 R&D 투자 비용을 극복해야 하므로 불가능한 현실에 처해있는 상황이다. 현시점에서 전 세계적인 제4차 산업혁명 파도의 핵심적인 역할을 하는 산업별 핵심 SW 기술을 조속히 확보하지 않으면 선진국과 후발국의 도전 속에 자동차, 조선 등 주력 산업도 경쟁력 유지가 어려울 것으로 보인다.

2. 주요 산업 분야에서의 SW 현황

국내 SW 산업은 전반에 걸쳐 선진국과 후발국 사이에 끼어 격차가 좁혀지지 않고, 최근에는 후발국인 중국, 인도 등은 빠르게 국내 산업과 기술격차를 좁혀가고 있으며, 일부 산업에서는 가격 경쟁력을 앞세워 시장 점유율을 빼앗기고

있는 상황이다. 소수 대기업 위주의 산업구조로 일부 대기업의 매출 성장에 따른 산업 경쟁력 향상의 착시현상을 불러일으키기도 하나, 산업의 다수를 차지하는 중소기업의 경쟁력은 하락할 수밖에 없다. 국내 중소기업은 자체 고부가가치 핵심 기술이 부족하고, 대기업 의존적으로 낮은 영업이익률과 지속가능성을 보이는 것으로 나타나고 있다.

기술 장벽이 높은 핵심 SW는 주로 외산을 도입하여 매출이 늘어도 실제 영업 이익률 향상으로 연결되지 않고 있다. 산업별 중소기업은 시스템 통합, 응용 기술보다 기술 장벽이 높은 제어 SW, 알고리즘, OS 등 핵심 SW를 확보하여 자체 경쟁력과 지속적인 발전 가능성을 보유하는 것이 필요하고, 기술 장벽이 높아 확보에 위험성이 있는 기술보다 고도성장을 위해 신속하게 확보할 수 있는 기술에 무게중심을 둔 것이 현실이며, 글로벌 경쟁 시대에 산업의 지속가능성을 위해 산업별 강점을 살릴 수 있고, 외산 대체 가능한 핵심 SW 기술의 확보가 필요한 시점이다.

기술발전의 방향을 살펴보면, 산업별 핵심 SW는 지능/무인화, 고신뢰/고안전, 고성능/저전력, 확장/맞춤형 기술로 변화하고 있다. SW로 사람과 같은 효율적인 인지 및 판단을 지원한다. 최종적으로는 자율주행 무인차, 자율주행 무인기와 같이 사람의 개입 없이 시스템을 운영하는 기술, 오류/공격에 의한 사고 발생을 방지하고, 방어할 수 있는 기술, 작은 기기에서도 고성능을 지원하고, 적은 전력으로도 동작을 가능하게 하는 기술, 개인 및 시스템의 다양한 요구사항을 맞춤형으로 빠르게 개발/적용할 수 있는 기술 등이 새롭게 출현하고 있다. 따라서 외산 SW에 의존적인 고부가가치 산업과 국산 SW 비중이 낮은 산업 분야에서 독자적인 핵심 SW 기술을 확보하여 중소·중견기업의 자생적 역량과 신산업 창출을 위한 지원이 필요하다. 또한, 제4차 산업혁명과 인공지능 분야의 중요성이 높아지면서 전 분야에 활용성이 높은 AI 분야 등에서의 핵심 SW 기술 확보 방안이 중요하게 대두되고 있다.

3. 한국의 SW 융합 및 AI 정책 현황

한국은 과학기술 혁신을 통한 국가발전을 도모하기 위해 대통령령으로 설치된 과학기술전략회의에서 AI를 비롯한 9대 국가전략 프로젝트[22] 추진계획을 발표하였다(제2차 과학기술전략회의, 2016.08.). 특히, 「지능정보사회 선도 AI 프로

젝트」를 통해 AI 핵심기술 자립기반 확보와 AI 기술·산업 성장의 기반을 조성하고자 하고 있다.

[그림 6-29] 지능정보사회 선도 AI 프로젝트 주요 내용

[지능정보사회 선도 AI 프로젝트 주요 내용]

구분	내용
AI 공통플랫폼	민간의 AI 제품 · 서비스 개발을 지원하기 위한 AI 요소기술(언어 · 시각인지, 학습, 추론기술 등)을 민관 협력으로 개발 및 제공
차세대 AI 기술	초기 단계의 국내 AI 기술력을 극복하여 세계적 기술 수준을 달성하기 위한 장기 원천기술 연구
AI 선도서비스	공공분야(국방, 치안, 노인복지) 우선 적용으로 민간 AI 수요 창출

출처: 해외 ICT R&D 정책동향 2017-01호, 정보통신기획평가원

제4차 산업혁명 및 지능정보기술로 인해 나타날 경제·사회의 혁신적 변화에 대응한 종합적인 국가전략으로서 「제4차 산업혁명에 대응한 지능정보사회 중장기 종합대책」을 발표하여 다음과 같이 추진한다(2016.12.).

그간 정부는 4차 산업혁명 및 AI 등장에 대한 각 부처의 대응력 제고를 위해 범부처 조직인 '지능정보사회추진단(이하 추진단)'을 조직하는 등 AI 국가전략 수립을 위한 지속적 노력을 기울이며, 특히, 추진단은 AI 개발에 관한 중장기적인 대책과 구체적인 실행계획 마련을 위해 전문가 및 국민 의견을 수렴하여 「제4차 산업혁명에 대응한 지능정보사회 중장기 종합대책」을 발표하였다. 동 종합대책은 한 세대 이상의 미래를 조망하고 혁신적 변화에 대응하기 위해 글로

22) (성장동력 확보분야) AI, 가상증강현실, 자율주행차, 경량소재, 스마트시티, (삶의질 제고 분야) 정밀의료, 탄소자원화, 미세먼지, 바이오 신약

벌 수준 기술 기반 확보, 전 산업 지능정보화 촉진, 사회 정책 개선 및 제도 정비 등 3개 정책 방향 및 12개 전략 과제로 구성되어 있다.

세계 주요국은 제4차 산업혁명 시대를 대비해 인공지능, IoT, 빅데이터 등 SW 융합 기술 개발 및 인력확보가 시급한 것으로 인식하고 이를 위한 정책과 전략을 수립하여 앞다투어 추진하고 있다. 한국은 '지능정보사회추진단'을 설치하여 AI 정책을 적극적으로 추진하고 있으나 역할이나 투자 규모가 부족한 것으로 보여 다음과 같은 점들을 보완할 필요가 있다.

가. 상시 전담조직 필요

일본은 총리 산하에 '인공지능기술전략회의'를 설치하여 범정부 차원의 역량과 혜안을 집중하여 AI 정책을 추진 중이며, 동 조직은 총리 주도로 만들어진 만큼 부처를 초월하여 강력한 권한 행사를 통해 AI 산업화 추진 가속화하고 있다. 반면 한국도 '지능정보사회추진단'을 설치하여 AI 정책을 적극적으로 추진하고 있으며, 향후 범부처 컨트롤타워의 역할 강화를 위한 제반 조건 정비가 필요하다. 특히 각 분야에 활용하기 융합 관점의 AI 기술 개발 및 개발 역량 강화를 위한 범부처 AI 정책 컨트롤타워의 역할을 강화하고, 이를 위한 제반 조건 등을 정비할 필요가 있다.

나. 사회적 수용성

한국과 일본 모두 AI 기술에 대한 부정적 인식 제거 등 사회적 수용성 제고 노력을 AI 정책 안에 포함하고 있다. 다만 일본의 경우, 불안감 제거 등 부정적 인식 개선에 더해 AI R&D 필요성, AI에 대한 기술적 이해 등 포지티브한 수용 노력을 포함하고 있다. AI 기술 발전의 성과와 결과물이 일반 국민에게까지 확산되기 위해서는 부정적 인식 개선 노력뿐만 아니라 AI의 긍정적 활용 확대 방안도 추진할 필요가 있으며, AI 기술 확산과 사회적 수용성 제고를 위해 부정적 인식 제거 및 각 분야에 AI 기술의 긍정적 활용 확대를 위한 다양한 인식개선 방안 모색이 필요하다.

다. AI/SW 융합 분야

글로벌 시장 선점을 위해서는 원천 기술 개발뿐만 아니라 시장의 니즈를 반영할 수 있는 전략이 필요하며, 일본은 고령화라는 사회문제를 해결해왔던 노하우에 AI 기술을 적극적으로 반영하여 경쟁력 확보를 도모하고 있으며, 한국도 파급효과를 고려하여 AI 기술 융합 분야로 공공서비스, 제조업 분야 등을 선정하고 로드맵 계획 중이다. 더 나아가 사교육, 남북문제 등 노하우 축적이 가능한 한국만의 문제 혹은 전염병, 고령화, 자연재해 등 글로벌 공통으로 부상하고 있어 시장 확보가 가능한 사회문제 등 문제 해결 방향으로 AI 융합을 시도하는 전략도 고려할 필요가 있다.

라. 인프라 구축 및 접근성 확대

일본은 AI 활용 극대화를 위해 AI 전용 슈퍼컴퓨터를 확보하고 기업이 활용할 수 있도록 구체적인 R&D 사업(거점 확보 사업)을 시작하였다. 한국도 기상청이 보유하고 있는 미리, 누리 등의 슈퍼컴퓨터를 포함하여 총 7대가 세계 슈퍼컴퓨터 고성능 500위 안에 들지만, AI 시장 선도와 각 산업으로의 활용 극대화를 위해서는 고성능화를 위한 국가 차원의 지원이 요구된다. 특히, 구축된 슈퍼컴퓨터에 대한 접근성이 떨어지는 스타트업이나 중소기업들이 AI 기술을 보다 쉽게 활용할 수 있도록 하는 지원 정책의 확대가 필요하다.

마. 정부주도 정책

민간이 대규모로 AI 개발을 진행하고 있는 미국과 다르게 한·일 양국은 정부 주도로 AI 개발 계획을 추진 중이며, 한국과 일본은 AI 추진체계 안에 민간 참여 확대 등 집단지성을 활용하고 있기 때문에 산·학·연의 다양한 시각이 정책에 반영이 가능한 장점이 있으므로 이를 충분히 활용한 정책 확장이 필요하다.

제5절 연구 의의 및 한계, 향후 방향

1. 연구의 의의

본 연구의 목적은 국가연구개발과제에서 일어나고 있는 SW 융합 현상을 파악하기 위해 NTIS에서 제공하는 약 5만여 개의 과제 중 2,000개의 과제를 표본으로 추출한 뒤, 상당한 시간과 비용을 투자하여 30명의 SW R&D 전문가의 3차례의 합의를 거치는 정성적 방법의 한계점을 개선하는 것이다. 이를 위해 기계학습을 활용한 자동 분류 모델을 개발하였으며, NTIS에서 매년 제공하는 국가연구개발과제 정보를 대상으로 SW 융합 R&D의 자동 분류가 단기간에 가능하며 표본 추출이 아닌 전수 조사 역시 가능하다. 즉, 한 번 개발된 모델을 활용하여 일부의 개선 작업을 거쳐 향후 제공되는 국가연구개발사업을 대상으로 반복적인 수행이 가능하므로 표본에 대한 조사를 통한 모수 추정 방식인 기존의 연구와 대비하여 투입 시간과 노력, 비용 절감이 가능하다는 장점을 가지고 있다.

연구의 결과물인 개발한 자동 분류 모델을 활용하면, 연구개발의 수행 부처, 국가과학기술표준분류체계의 대분류, 연구개발 단계 및 연구수행 주체 등 NTIS에서 제공하는 다양한 유형에 따른 SW 융합 R&D 현황을 파악할 수 있으며 연차별 SW 융합 R&D에 대한 추세 분석 등이 가능할 것이다. 또한, 미국의 연구기관인 NSF나 DARPA, EU의 Horizon 2020과 같은 국외 연구기관은 자신들이 수행하는 과제 정보를 제공하고 있는데, 이를 분류 모델에 적용하면 해당 기관이 수행하는 SW 융합 R&D의 단계별 현황에 대한 파악이 가능하며, 이를 국내 현황과 비교할 수 있다.

본 연구에서 개발한 자동 분류 모델을 활용한 SW 융합 R&D 현황 분석 자료는 향후 정부가 마련하는 SW 융합 R&D 혁신 정책을 도출하기 위한 기초 자료로 활용할 수 있다. 또한, 본 연구에서 수행한 국외 SW 관련 R&D 정책 분석과 국내 국가연구개발과제의 SW 융합 R&D 현황 분석을 통하여 SW 기술이 기반핵심이 되고 있는 제4차 산업혁명 시대에 맞춰 전반적인 국가연구개발과제의 SW 융합 정도를 적절한 수준으로 향상하기 위한 정책 수립의 근거를 제시하였다고 할 수 있다.

2. 연구의 한계

물론 본 연구에서 활용한 학습 데이터 중 중간 단계에 속하는 Med High, Medium, Med Low 단계에 대한 예측 결과가 High 및 Low 단계에 비해 재현율과 정밀도가 낮다는 한계점은 존재한다. 이를 극복하기 위해 SW R&D 전문가가 추가로 분류한 1,124개의 학습 데이터를 확보하여 분석에 활용하였다. 하지만 근본적으로 중간 단계의 데이터는 전문가가 판단할 경우에도 의견이 일치하지 못한 부분이기 때문에 여전히 분류가 어렵다는 제약사항이 있다. 뿐만 아니라 tf-idf와 Linear SVM 모델의 조합에서 5가지 단계의 SW 융합 R&D 분류가 가지는 예측력은 72.78%로 해당 예측력을 높이기 위한 추가적인 노력이 필요할 것이다. 이러한 부분을 해결하기 위해서 보다 다양한 텍스트 분류 모델에 대한 실험 및 예측 결과에 대한 추가 검증을 수행하여 모델을 보완하면 더 예측력이 높고 일반화된 연구 결과를 얻을 것으로 기대한다. 또한, 본 연구에서 개발한 코드의 일부 수정을 통해 SW 융합 R&D 단계 분류 외에 타 분야에서 이미 단계가 구분된 텍스트 형태의 학습 데이터를 가지고 있다면 이로부터 단계를 예측하는 문제 등의 이슈에 적용하여 문제를 해결할 수 있을 것이다.

3. 향후 연구 방향

본 연구의 학습 데이터는 2016년 국가연구개발사업의 총 과제 수인 54,827건에서 표본 추출한 2,000건의 데이터를 기본으로 하였다. 여기에 SW R&D 전문가의 검증을 통해 추가로 분류한 과제 데이터를 추가하여 총 3,124건으로 구성하였다. 분류 모델의 예측력을 높이기 위해서는 충분한 학습 데이터의 크기가 확보되어야 하지만 시간과 비용의 제약 때문에 본 연구에서 확보한 추가 데이터의 크기를 통해서는 예측력을 높이는데 한계가 있었다. 또한, 최근 요구되는 SW 관련 기술들이 빠르게 변화하고 있으며 국가연구개발사업 역시 이를 반영하기 때문에 2016년 국가연구개발사업의 학습 데이터로 현재와 앞으로 진행될 국가연구개발사업의 예측이 어려울 수 있다. 그러므로 이러한 상황을 극복하기 위해서는 추가로 최근에 수행된 국가연구개발사업에 대한 추가 분류 작업을 통해 학습 데이터의 보강 및 분류 모델의 개선 작업이 하나의 방법이 될 수 있다. 그리고 앞서 언급한 것과 같이 다양한 텍스트 임베딩 방법과 분류 모델의 실험을 통해 예측력을 높이기 위한 추가 연구가 필요하다.

또한, 본 연구에서 개발한 분류 모델을 2018년 국가연구개발과제에 적용하고 2017년 국가연구개발과제에 대한 예측 결과와 비교 분석하고, 추후 지속적으로 예측하여 추이 분석을 통한 현황 및 현안 파악 역시 요구된다. 마지막으로 본 분류 모델은 영문 텍스트를 입력 값으로 활용하고 있으므로 국외에서 수행되는 국가연구개발사업 관련 텍스트 데이터를 확보하고 자동 분류 모델을 적용할 수 있다. 이를 통해 국내외 SW 융합 R&D 단계에 대한 비교·분석을 수행하여 국내 SW 융합 R&D의 발전방향을 제시하는 추가 연구가 필요하다.

참 고 문 헌

[국내 문헌]

공영일・서영희(2019), 「국가연구개발사업에서 SW융합 연구의 현황과 분석」, 소프트웨어정책연구소

과학기술정보통신부(2016), 「제4차 산업혁명에 대응한 지능정보사회 중장기 종합대책」

과학기술정보통신부(2018), 「2019년도 정부 연구개발 투자방향 및 기준」

금융보안원(2018), 「설명 가능한 인공지능(eXplainable AI, XAI) 소개」

김동성(2019), 「Doc2Vec 단어 임베딩 언어 모델을 활용한 텍스트 장르 구분」, 한국언어정보학회, 언어와 정보 23권2호, pp.23-43.

김승희(2015), 「체계적 문헌고찰을 위한 텍스트 분류 모델링 방법에 대한 연구」, 서울대학교, pp.28-29.

김진형・배두환 외(2011), 「SW R&D체계 개편방안 연구」, 한국과학기술원

서영희(2019), 「기계학습을 활용한 SW 융합 R&D 자동 분류 및 현황 분석에 관한 연구」

서영희・공영일(2018), 「SW 융합 R&D 현황 분석 및 시사점」, 소프트웨어정책연구소, pp.23-25.

소프트웨어산업진흥법 제2조 2항

이상기・이병섭・박병용・황혜경(2010), 「나이브베이즈 분류모델과 협업필터링 기반 지능형 학술논문 추천시스템 연구」, 정보관리연구, vol.41, no.4, pp.227-249.

이인규(2015), 「층화표본에서의 표본 배분에 대한 연구」, 한국통계학회, 응용통계연구 28권 6호, pp.1047-1061.

정근하(2010), 「텍스트마이닝과 네트워크 분석을 활용한 미래예측 방법 연구」, 한국과학기술기획평가원, pp.42-46.

정보통신기획평가원(2016), 「미국의 인공지능 R&D 전략계획」

정보통신기획평가원(2017), 「독일 (신)하이테크전략 10년 성과 리뷰」
정보통신기획평가원(2017), 「일본의 인공지능 정책동향과 실행전략」
정보통신기획평가원(2017), 「중국의 AI 정책 동향」
정보통신기획평가원(2017), 「해외 ICT R&D 정책동향 2017-01호」
한국경제연구원(2018), 한국, 「글로벌 시총 500대 포함기업 수…10년 간 제자리」

[국내 사이트]
국가과학기술지식정보서비스(NTIS) 홈페이지, https://www.ntis.go.kr/

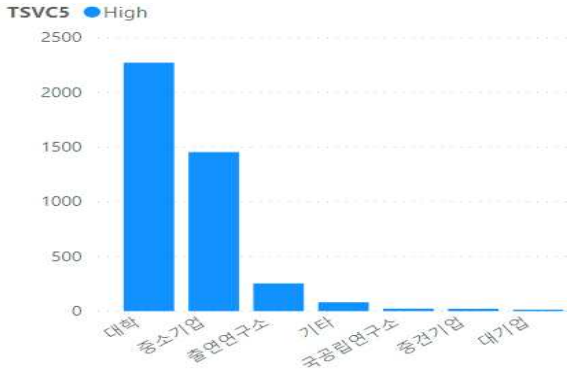
[국외 문헌]
Corinna Cortes. and Vladimir Vapnik. (1995). 「Support-Vector Networks, Machine Learning」, 20, pp.273-297.
Hearst, M.A., S.T. Dumais, E. Osman, J. Platt, and B. Schölkopf. (1998), 「Support vector machines」, IEEE Intelligent System, Vol. 13, No. 4, , pp.18-28.
Jiang Su, Jelber Sayyad-Shirabad and Stan Matwin. (2011), 「Large Scale Text Classification using Semi-spervised Multinomial Naive Bayes」, ICML.
Marco T. R. and Sameer S. (2016), 「 "Why Should I Trust You?" : Explaining the Predictions of Any Classifier」, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
Quoc Le. and Tomas Mikolov. (2014), 「Distributed Representations of Sentences and Documents, in Proceedings of the International Conference on Machine Learning」, pp.1188-1196.

[국외 사이트]
http://scikit-learn.org/

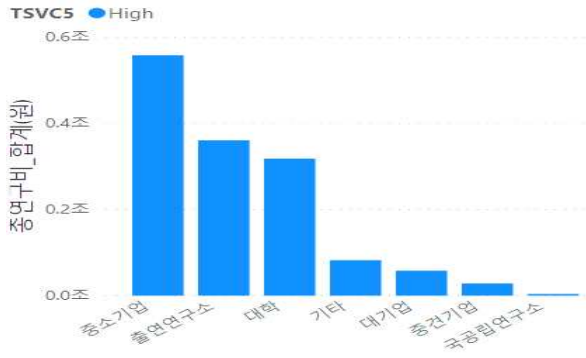
[부록]

1. 연구개발(R&D) 수행 주체 - SW 융합 R&D 5단계 구분 현황

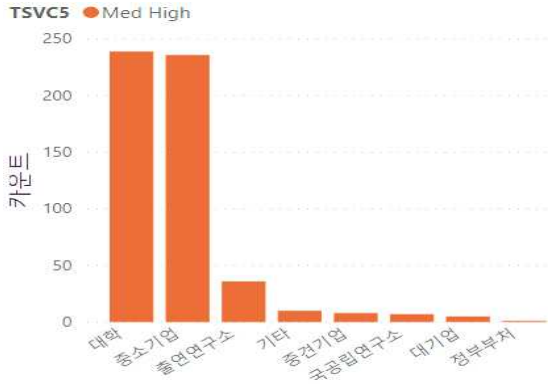
[그림 1] High-연구수행 주체(과제 수)



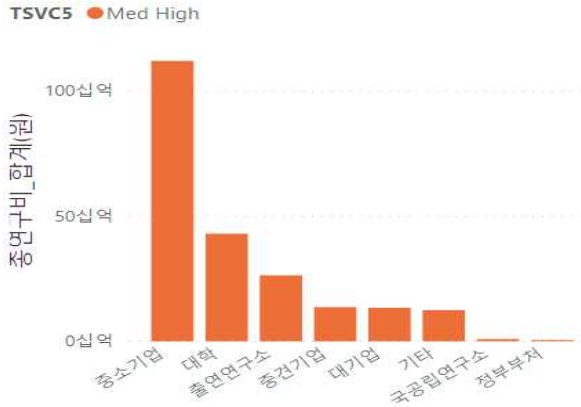
[그림 2] High-연구수행 주체(연구비)



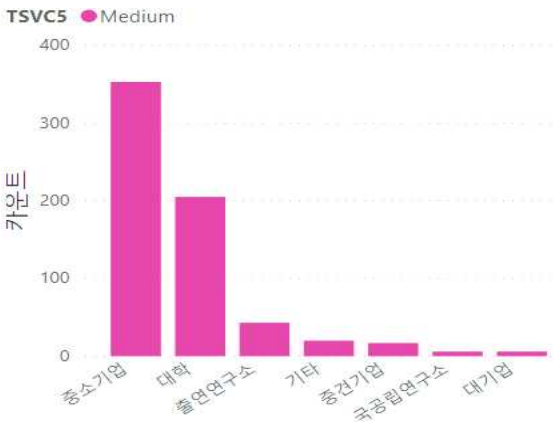
[그림 3] Med High-연구수행 주체(과제 수)



[그림 4] Med High-연구수행 주체(연구비)

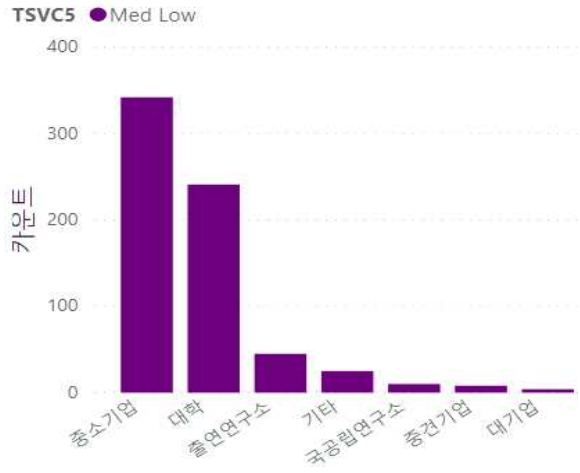


[그림 5] Medium-연구수행 주체(과제 수)

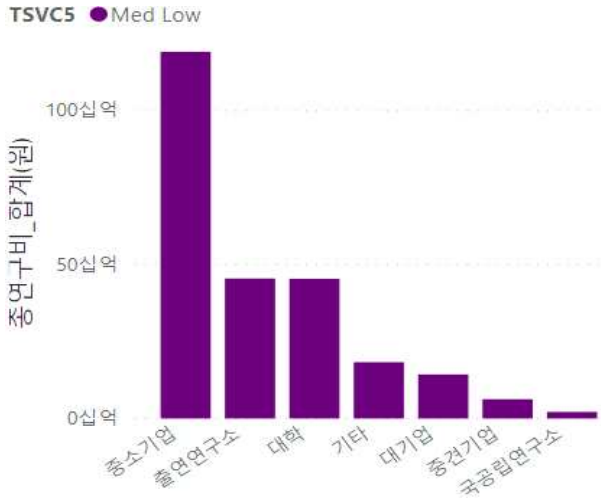


[그림 6] Medium-연구수행 주체(연구비)

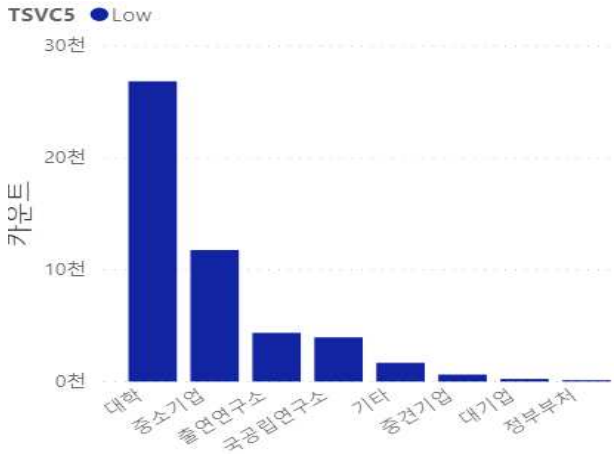
[그림 7] Med Low-연구수행 주체(과제 수)



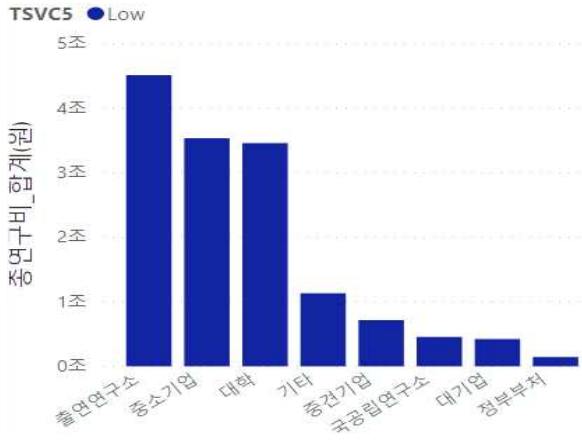
[그림 8] Med Low-연구수행 주체(연구비)



[그림 9] Low-연구수행 주체(과제 수)

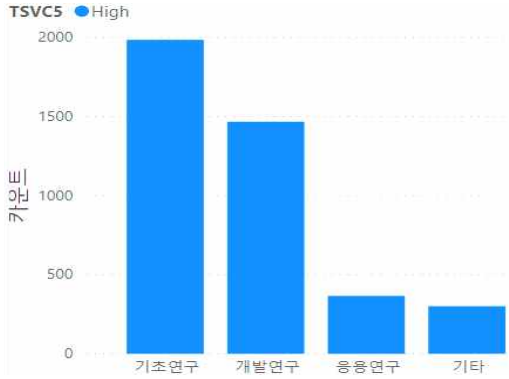


[그림 10] Low-연구수행 주체(연구비)

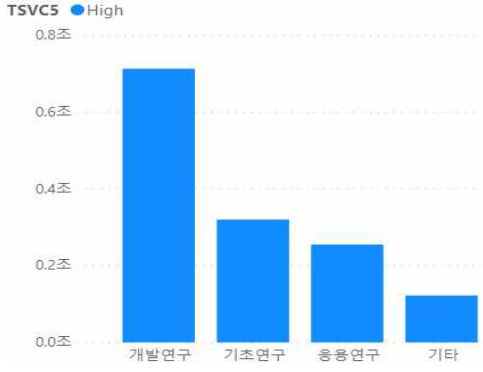


2. 연구개발(R&D) 단계 - SW 융합 R&D 5단계 구분 현황

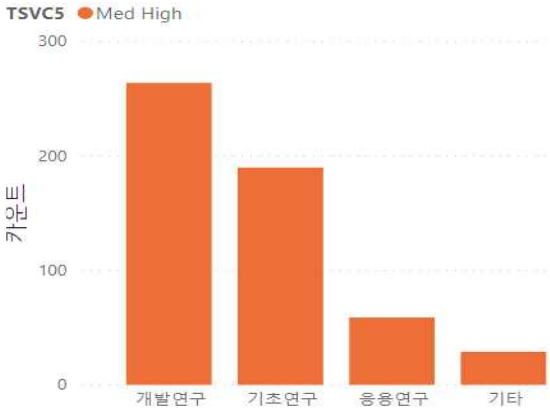
[그림 11] High-R&D 단계(과제 수)



[그림 12] High-R&D 단계(연구비)



[그림 13] Med High-R&D 단계(과제 수)

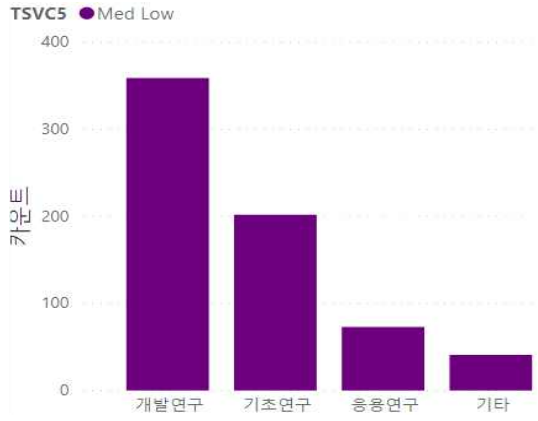


[그림 14] Med High-R&D 단계(연구비)

[그림 15] Medium-R&D 단계(과제 수)

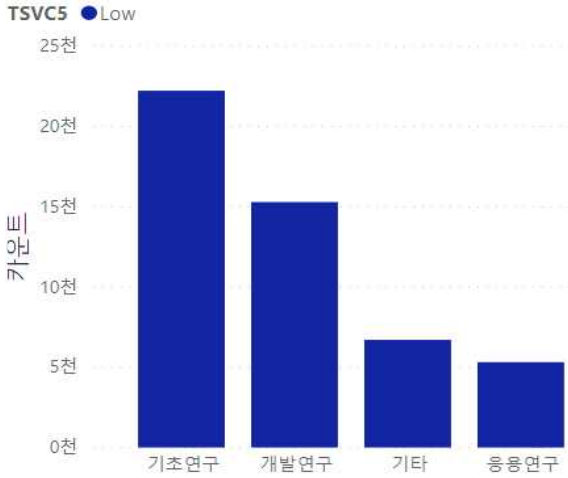
[그림 16] Medium-R&D 단계(연구비)

[그림 17] Med Low-R&D 단계(과제 수)

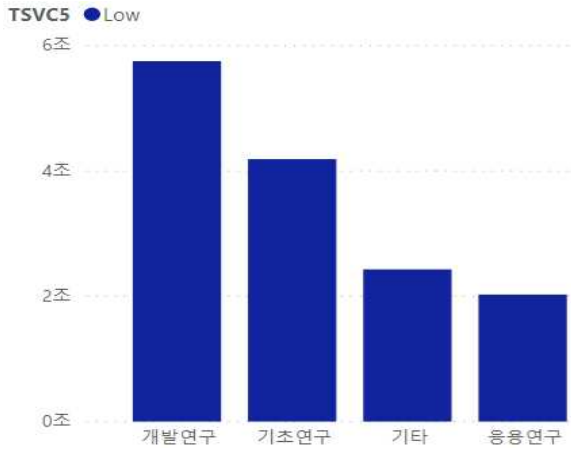


[그림 18] Med Low-R&D 단계(연구비)

[그림 19] Low-R&D 단계(과제 수)



[그림 20] Low-R&D 단계(연구비)



주 의

1. 이 보고서는 소프트웨어정책연구소에서 수행한 연구보고서입니다.
2. 이 보고서의 내용을 발표할 때에는 반드시 소프트웨어정책연구소에서 수행한 연구결과임을 밝혀야 합니다.

비매품/무료
93500
9 788961 084703
ISBN 978-89-6108-470-3