산업융합연구 제20권 제7호 pp. 1-11, 2022
ISSN 2635-8875 / e-ISSN 2672-0124
DOI : https://doi.org/10.22678/JIC.2022.20.7.001

토픽모델링 분석을 활용한 국가연구개발사업과제와 국회 상임위원회 사이의 정책 인식 비교 : ICT 분야를 중심으로

송병기[1], 김상웅[2*]
[1]정보통신정책연구원 재무회계팀 팀장, [2]정보통신정책연구원 기획전략실 행정원

Comparison of policy perceptions between national R&D projects and standing committees using topic modeling analysis : focusing on the ICT field

Byoungki Song[1], Sangung Kim[2*]
[1]Director, Finance & Accounting Team, Korea Information Society Development Institute
[2]Staff, Department of Planning&Strategy, Korea Information Society Development Institute

요 약 본 논문에서는 여러 연구기관에서 논의하고 있는 데이터 기반 평가 방법론 중 토픽모델링 기법을 이용하여 계량적인 값을 도출하고 그 과정에서 실제 전문가들이 수행하는 국가연구개발사업과제와 이를 법률과 정책실무에서 다루는 국회 상임위원회 간의 정책적 인식 차이가 있는지 ICT 분야를 중심으로 파악해 보고자 한다.
먼저 HAN 모델로 사업과제 데이터를 학습하여 ICT 문서를 분류하는 모델을 만들고, 해당 모델을 통해 분류된 ICT 문서를 대상으로 LDA 토픽모델링 분석을 수행하여 국가연구개발사업과제 데이터와 국회 상임위원회 회의록에서 도출된 토픽과 분포를 비교한다. 구체적으로 총 26개의 토픽이 도출되었으며, 각 토픽이 포함하는 단어와 문서 분포 비율을 살펴봤을 때, 국가사업과제는 상대적으로 전문적인 주제의 문서가 많았으며, 국회 상임위원회는 상대적으로 사회적이고 대중적인 문제를 다루는 것으로 나타나 인식에 다소 차이가 있는 것으로 보였다. 인식의 차이를 수치적으로 확인할 수 있는 만큼, 향후 정책이나 과제 평가에 사용할 수 있는 지표에 대한 기초연구로 활용 가능할 것이다.

키워드 : 국가연구개발사업과제, 국회 상임위원회, 회의록, 문서분류, 토픽모델링

Abstract In this paper, numerical values are derived using topic modeling among data-based evaluation methodologies discussed by various research institutes. In addition, we will focus on the ICT field to see if there is a difference in policy perception between the national R&D project and standing committee. First, we create model for classifying ICT documents by learning R&D project data using HAN model. And we perform LDA topic modeling analysis on ICT documents classified by applying the model, compare the distribution with the topics derived from the R&D project data and proceedings of standing committees. Specifically, a total of 26 topics were derived. Also, R&D project data had professionally topics, and the standing committee-discuss relatively social and popular issues. As the difference in perception can be numerically confirmed, it can be used as a basic study on indicators that can be used for future policy or project evaluation.

Key Words : National R&D projects, Standing committee, Proceeding, Document classification, Topic modeling

*Corresponding Author : Sangung Kim(tkddnd0214@kisdi.re.kr)
Received April 13, 2022 Revised June 16, 2022
Accepted July 20, 2022 Published July 28, 2022

1. 서론

2021년 9월, 기획재정부에서 발표한 보도자료에 따르면 정부는 2022년도 국가연구개발(R&D) 예산을 29.8조 원 규모로 편성하였고, 이는 2021년도 대비 약 8.8% 증가한 수준이다. 이에 따라 우리나라는 전체 R&D 투자 세계 2위, GDP 대비 R&D 투자 세계 1위의 연구개발 투자 강국으로 자리매김하였다. 국가연구개발사업이란 중앙행정기관이 법령에 근거하여 연구개발을 위해 예산 또는 기금으로 지원하는 사업을 말하는 것이다[1]. 즉, 국가연구개발사업은 큰 규모의 정부예산과 세금이 투입되고 있기 때문에, 우리나라의 발전에 이바지하고 국민의 기대에 부응하는 방향으로 수행되어야 한다. 따라서 해당 사업을 수행하는 연구기관은 매년 설립목적에 부합하는 국가사업과제를 잘 시행하고 있는지, 연구를 통해 정책을 생산하는 과정 및 결과에 대한 기여 정도에 대한 성과평가를 받는다[2].

국가사업과제는 결과와 성과를 통해 국회에 올바른 정책적 방향 설정에 선도적인 제언을 하는 역할을 하고 있고, 국회 상임위원회는 국가사업과제의 목표를 효율적으로 달성할 수 있도록 국가사업과제의 세부 지원 분야 또는 지원 대상, 지원 규모 및 지원 기간, 과제선정 및 과제 평가 방법 등을 수립하는 역할을 하고 있기 때문에 상호보완적인 관계로 볼 수 있다[3]. 특히 주요 국회 상임위원회에서는 소관 부처의 R&D사업에 대하여 추진체계와 투자전략은 물론 주요 사업의 예산 집행과 성과관리 관련 문제점과 개선과정을 도출하고자 노력하고 있다[3]. 따라서 실제로 수행되고 있는 사업과제와 정책의 주제와 방향이 맞아야 사회에 전략적이고 올바른 영향을 끼칠 수 있을 것이며, 이에 대해 파악하고 평가할 프로세스를 연구할 필요가 있다.

현재 연구 평가의 지표들이 서지학 분석에 초점을 맞추고 있고, 전문가 판단으로 평가되고 있기 때문에 개개인의 주관에 따라 성과평가에 대해 단편적인 면만 보여질 수 있다[4]. 그리고 빅데이터에 대한 접근이 쉬워짐에 따라 여러 연구단체에서 데이터 기반 평가 방법론에 대한 논의가 이루어지고 있다. 한국법제연구원에서는 데이터에 기반한 입법 평가방법론에 대한 연구가 수행되었고[5], 국회예산정책처에서는 데이터에 기반한 사업평가방법론에 대해 논의하였다[6]. 이외에도 과학기술정책연구원, 한국과학기술기획평가원, 한국지능정보사회진흥원 등의 기관들도 데이터에 기반한 연구개발 관리 방안, 연구혁신 정책 의사결정 지원 방안, 기술기획 방안, 공공데이터 성과평가 방안 등을 개발하고자 연구를 진행하였다[7-10]. 하지만 텍스트마이닝, 오피니언 마이닝, 네트워크 분석 등의 방법론 적용 가능성에 대해서만 논할 뿐, 구체적으로 어떤 자료에 실제로 적용하여 어떤 계량적인 값을 이용하여 어떻게 지표를 만들고 평가할지에 대하여 발표된 연구는 찾기 쉽지 않다.

따라서 본 논문에서는 5년(`16~`20년) 동안 수행된 국가연구개발사업이 얼마나, 어떤 토픽에 대해 국회 상임위원회에서 정책적으로 논의되었는지 텍스트데이터를 기반으로 한 토픽모델링을 통해 비교·분석하고자 한다. 또한 국가연구개발과제와 국회 상임위원회가 다루는 내용의 분야가 많아 모든 내용을 언급하기는 힘들기 때문에, 국가연구개발과제 10대 중점 분야 중 하나인 디지털 뉴딜 사업과 관련된 ICT 분야를 중심으로 살펴보고, 계량적인 값을 도출 및 비교하고 분석 과정상에서의 이슈에 대하여 살펴보고자 한다. 다음절에서는 본 연구에 사용된 모형과 이론에 대해 간략히 소개하고, 3절과 4절에서는 본 연구를 위해 수집된 자료와 연구 과정 및 결과를 소개한다. 그리고 5절에서는 본 연구의 결론 및 시사점에 대해 언급하며 마무리하겠다. 분석자료로는 과학기술정보통신부에서 운영하는 국가과학기술지식서비스(NTIS : National Science & Technology Information Service)에서 제공받은 사업과제정보 데이터와 20대, 21대 국회 상임위원회 회의록 자료를 수집하여 사용하였으며, 분석 tool로는 파이썬의 gensim[1] 패키지를 활용하였다.

2. 이론적 배경

2.1 Hierarchical attention networks(HAN)

문서분류는 임의의 문서를 이미 정해져 있는 범주에 따라 분류하는 문제이다. 문서는 다양한 길이의 연속된 문장들로 구성되어 있고, 각각의 문장들은 다양한 단어 성분으로 구성되어 있다. HAN 모델은 단어-문장-문서로 이어지는 문서의 계층적 특성을 반영시키고자 Yang 등(2016)에 의해 제시된 모델로 어텐션 메커니즘(attention mechanism) 아이디어를 적용한 문서분류 모델이다. 어텐션 메커니즘이란 전체 입력 문장을

1) 자연어를 수치화 벡터로 변환하는데 필요한 대부분의 편의 기능을 제공해주는 Python 라이브러리

동일한 비중으로 참고하는 것이 아니라, 예측 시점에서의 단어와 연관 있는 문장과 단어에 가중치를 주어 집중적으로 반영하는 방법이다. 특히, RNN(Recurrent neural network) 기반의 문서 분류 모델인 HAN은 문서의 계층적 구조를 반영하고, 중요한 단어나 문장에 가중치를 더해줄 수 있어, 문서분류 연구에 활발하게 사용되고 있다[11]. 예를 들어, 온라인 리뷰나 민원 자료를 이용하여 그 카테고리를 분류하는 문제나, 특허문서를 이용하여 분류 체계를 정립하는 문제 등에 적용되는 활용성을 보였다[12-14].

HAN 모델은 Fig. 1과 같이 크게 워드 시퀀스(word sequence), 문장 시퀀스(sentence sequence) 두 부분으로 구성되어 있으며, 각각의 시퀀스는 인코더(encorder)와 가중치를 부여하는 부분으로 구성된다[11]. 워드 시퀀스에서 계산된 값은 문장 인코더의 입력값(input value)으로 사용되고 softmax 함수로 구성되어 있는 출력층을 거쳐 각각 문서 범주로 분류될 확률을 도출하고, 가장 확률이 높은 범주로 분류한다. HAN의 인코더는 일반적으로 GRU(Gated recurrent unit) 기반의 양방향 순환 신경망(Bidirectional RNN)을 사용한다[15]. HAN 모델은 다양한 길이의 문장에 대해 효과적인 분석이 가능하고, 문서들의 계층적 특성을 반영할 수 있어 여러 방면에서 기존 RNN 기반 문서 분류 방식에 비해 우수한 분류 성과를 보인다[16].

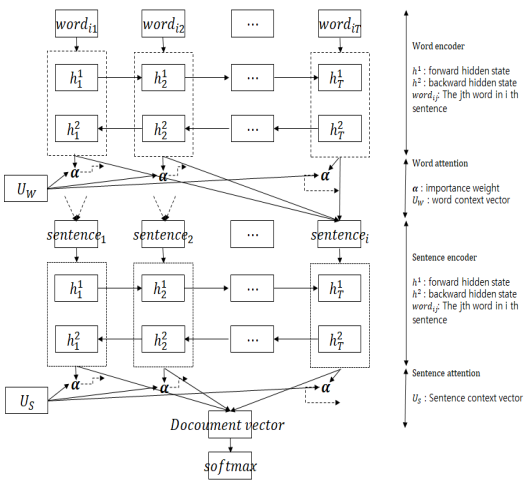

Fig. 1. Structure of HAN model

2.2 Latent dirichlet allocation(LDA)

토픽모델링이란 텍스트데이터에서 단어들의 동시 사용 규칙을 기반으로 해당 텍스트데이터를 대표하는 특정 토픽이나 주제를 자동으로 발견해내는 비지도 분석기법이다. 그 중 LDA 모델은 Blei 등(2003)에 의해 제안된 토픽모델링 기법으로 하나의 문서가 하나의 주제로만 분류되던 이전의 유니그램 혼합모형(mixture of uni-grams)이나 PLSI(Probablistic latent semantic indexing)와 달리 하나의 문서 내에 여러 주제가 서로 다른 가중치로 혼합되는 것을 허용하며, 훈련 데이터(training data)에 대한 과적합 문제를 잘 보이지 않는다는 장점이 있다[20]. 이러한 장점으로 국내에서는 LDA모형을 이용하여 코로나19 관련 연구 동향, 해외 건설시장 동향, 문헌정보학 연구 동향 등 폭넓은 분야에서 활용되고 있다[17-19].

Fig. 2에서 α는 θ를 생성하는데 사용되는 디리클레 분포(Dirichlet distribution)의 모수이며, θ는 문서가 각 토픽에 속할 확률 분포 벡터를 의미한다[21]. 그리고 벡터 내 확률값 중 가장 큰 값을 가지는 토픽 z에 단어 w가 분류된다. 여기서 w는 문서 내에 구분된 하나하나의 단어 변수를 의미하며, β는 각 토픽에 속한 단어들의 확률 분포 벡터를 의미한다[21]. 최종적으로 문서의 개수가 M개이고, 사용된 단어의 개수가 N개일 때, 다음과 같은 과정을 거쳐 가장 큰 값을 가진 토픽에 관측된 단어와 문서가 분류된다[20].

(1) 연구자의 판단에 따라 토픽 개수 K개를 결정
(2) 모든 단어를 K개의 토픽 중 랜덤으로 하나의 토픽에 할당
(3) 아래 과정을 반복 :
① 하나의 단어 w를 제외한 다른 단어들은 올바른 토픽에 할당되어 있고, 단어 w만 잘못 할당되어 있다고 가정
② 문서에 대한 토픽의 확률 분포 $(\theta_i|\alpha \sim P(topic\ t_i|document\ d_j))$를 계산,
$i=1,2,\cdots,K,\ j=1,2,\cdots,M$
③ 토픽에 대한 단어의 확률 분포 $(b_j \sim P(word\ w_i|topic\ t_j))$를 계산,
$i=1,2,\cdots,N,\ j=1,2,\cdots,K$

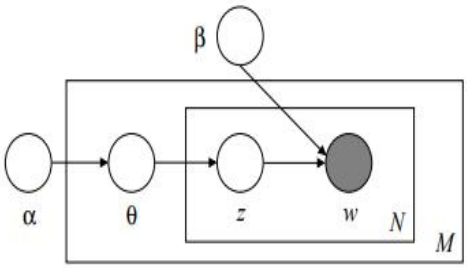

Fig. 2. Structure of LDA model

3. 연구방법 및 절차

3.1 연구 절차

본 연구를 위하여 NTIS에서 5개년(2016~2020년) 국가연구개발과제 데이터를 제공받았으며, 국회에서 운영하는 국회회의록 사이트에서 국회 상임위원회 회의록을 수집하였다. 국가연구개발과제 데이터를 사용하여 ICT분야의 문서를 판별하는 HAN모델을 학습시켰으며, 학습된 모델에 회의록을 적용하여 마찬가지로 ICT분야 문서를 구분하였다. 이후, ICT문서로 판별된 국가연구개발과제 데이터를 이용하여 LDA 토픽모델링을 실시하여 토픽 분포와 토픽 내 단어를 파악해 보았으며, 해당 모델에 ICT문서로 판별된 회의록 데이터를 적용시켜 토픽 분포를 확인하고 이를 국가연구개발과제의 토픽 분포와 비교해 보았다(Fig. 3 참조).

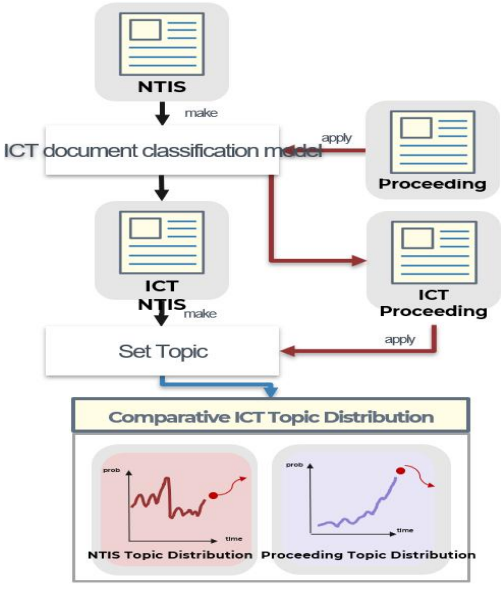

Fig. 3. Research procedure

3.2 데이터 수집·정제 및 용어사전 구축

3.2.1 국가연구개발사업 데이터

본 연구에서는 R&D과제 기본정보, 과제 정보, 요약서, 수행과제 내용 등을 변수로 가지고 있는 연도별 사업과제정보 데이터가 각각의 과제에 대해 과학기술표준분류에 따라 명확하게 분류되어 있고, 데이터의 형태가 토픽모델링을 수행하는데 적절하다고 보여 해당 데이터를 연구의 기준 자료로 결정하였고, '과학기술표준분류1-대', '요약문_연구목표', '요약문_연구내용', '요약문_기대효과' 변수를 분석에 사용하였다. 수집된 연도별 국가연구개발사업 데이터의 현황은 Table 1과 같으며, 총 323,593건의 데이터 중에 결측치, 중복데이터, 전문이 영어인 데이터, 데이터가 너무 짧거나 단어가 적은 데이터 등을 제거하고 총 199,870건의 국가연구개발사업 데이터를 분석에 사용하였다.

Table 1. Number of national R&D project data by year

Year	count
2016	54,827
2017	61,280
2018	63,697
2019	70,288
2020	73,501
Total	323,593

3.2.2 국회 상임위원회 회의록 데이터

국회 회의록은 국회 또는 회의체의 회의 시작에서 끝까지 모든 의사에 관한 발언 등을 사실대로 기록한 문서로, 회의에 대한 공식 기록이며 회의에 관한 쟁점이 있을 때 유력한 증거가 되는 자료이다. 특히 국회 상임위원회는 각 전문 분야별로 정부가 제출한 법률안에 대해 담당하는 의안이나 청원을 심사하기 위한 상설적으로 운영하는 위원회 조직으로 그 회의록에는 의정활동에 대한 앞으로의 비전 등 정부에서 중요하게 여기는 계획과 평가가 담겨 있다. 국회 상임위원회 회의록은 그 자체의 양이 적고, 국가사업과제에 대한 위원회 논의 반영에 시간차가 있다고 판단하여 좀 더 긴 기간의 자료를 수집하였다. 따라서 본 연구에서는 2020년 7월 정부 발표한 '한국판 디지털 뉴딜 종합계획'과 31개 대표과제에 대한 유관 20대, 21대(~'21.11) 국회 상임위원회 회의록 자료 총 877건을 수집하였으며, 자료의 현황은 Table 2와 같다.

Table 2. Number of proceeding of standing committees

# / committees	20th	21th (~2021.11)	Total
Science and ICT Committee	60	34	94
the Board of Education	25	26	51
Land, Infrastructure, Transport and Tourism Commission	67	38	105
Committee on Agriculture, Food, Marine and Fisheries	101	34	135
Culture, Sports and Tourism Committee	33	47	80
Culture, Sports and Tourism Committee	85	30	115
Trade and Industry Support Committee	29	-	29
Trade and Industry Support Small and Medium Venture Business Committee	72	44	116
Safety and Administration Committee	34	-	34
Public Administration and Security Commission	70	48	118

3.2.3 형태소 분석기 및 용어사전 구축

Python에서는 한국어 텍스트나 형태소 처리를 위한 KoNLPy 패키지를 사용할 수 있고, KoNLPy 패키지는 Kkma, Hannanum, Mecab-ko, Komorna 등의 여러 분석기를 제공하고 있다. 본 연구에서는 한국어 명사의 분석 성능이나 실행시간 면에서 다른 분석기에 비해 우수하다고 알려져있는 Mecab-ko 패키지를 사용하였다[21].

텍스트 분석은 일반적으로 단어의 출현 여부 혹은 출현 빈도에 따라 텍스트를 수치로 변경하는 작업을 하게 된다. 하지만 문서를 구성하는 모든 용어를 사용하기에는 양이 방대하기 때문에, 분석 주제에 집중하고 향상된 품질의 분석 결과를 도출하기 위해 용어 사전 또는 불용어 사전을 사용하여 단어를 제한한다. 용어 사전이란 분석에 사용할 단어를 정의한 목록을 의미하며, 불용어 사전은 분석에서 배제되는 용어의 목록으로, 주로 문서의 내용 파악이나 주제 식별에 영향을 주지는 않지만 자주 출현하는 용어들로 구성된다[22].

본 연구에서는 분석을 위한 용어 사전으로 K-ICT 빅데이터센터와 지능정보사회진흥원(NIA)에서 개발한 NIADic이라는 형태소 사전을 기반으로 구축하였으며 그 프로세스는 다음과 같다.

(1) NIADic 형태소사전 확보(929,160단어)

(2) (1)에서 1차로 3가지 사전으로 구분

① 1차 용어 사전 : 명사이거나 정보·통신으로 분류되어 있는 단어(740,494 단어)

② 불용어 사전 : 1차 용어 사전을 제외한 모든 단어(형용사, 동사 등)(188,666단어)

③ 정보통신 사전 : Category가 정보· 통신으로 분류되어 있는 단어(6,313단어)

(3) (2)-①,(2)-②에서 각각 중복된 단어 제거

① 1차 용어 사전(-76,962단어 → 663,532단어)

② 불용어 사전(-21,207단어 → 167,459단어)

(4) (3)-① 에서 불용어 사전에 포함된 단어와 동음인 단어 제거(-7,535단어 → 655,997단어)

(5) 최종 용어사전 : (4)에 정보통신 사전에 포함된 단어 포함 후 중복단어 제거(+90단어 → 656,087단어)

4. 연구 결과

4.1 ICT 문서 분류기 활용(HAN)

분석을 위해 수집한 국가사업과제 데이터와 국회 상임위원회 자료는 여러 분야의 내용이 혼합되어 있다. 본 연구에서는 ICT 분야에 대한 내용의 비교를 중점으로 다루고 있기 때문에, ICT 분야에 대한 자료만을 판별해 내야 할 필요가 있다. 따라서 본 연구에서는 문서 분류 모델 중 하나인 Hierarchical attention networks(HAN) 모델을 사용하여 ICT 문서를 분류해 추출하였다. 본 연구에서 사용된 HAN 모델의 구조는 문장 인코더와 워드 인코더를 각각 100층의 GRU(Gated recurrent unit) 은닉층으로 구성하였고, 2개의 뉴런을 가지는 출력층에서는 데이터 마이닝의 분류 문제에 주로 이용되는 Softmax 함수를 사용하였다. 여기서 최종적으로 출력층에서 도출되는 두 뉴런의 값은 해당 문서가 ICT 문서일지, Non ICT 문서일지에 대한 확률이며, 최종적으로 계산되는 두 확률에 따라 값이 더 큰 쪽의 문서로 분류된다. 손실함수는 교차엔트로피 오차를 사용하였으며, 학습률은 0.001, 모든 Random seed는 42로 설정하였다.

최근 ICT 분야는 단순 ICT분야 하나에 한정되어 있는 것이 아니라 다양한 분야와 융합되어 있는 경우가 많다. 따라서 다양한 분야에 대한 문서를 좀 더 ICT 문서로 구분해내기 위해 ICT 문서 분류 모델 학습을 위

해 앞서 구축하였던 정보통신 사전에 포함되는 단어가 15개 이상 포함되는 국가연구개발사업 자료를 ICT문서라고 정의하였다. 그 결과 ICT 문서로 82,484건, Non ICT 문서로 117,386건으로 라벨링되었다. ICT 문서와 Non ICT 문서의 클래스 균형을 맞추기 위해 ICT 문서와 똑같이 82,484건의 Non ICT 문서를 랜덤 샘플링 하였고, 각각 60%(총 98,980건), 20%(총 32,994건), 20%(총 32,994건)씩 랜덤 샘플링하여 Train data, Validation data, Test data를 구성하여 모델을 학습시키고 평가하였다.

Train data를 학습한 모델에 Validation data를 적용하여 threshold에 따른 분류 정확도와 재현율, ROC curve를 확인하였다. 모델이 한쪽으로 치우친 정확도를 보이는 것보다, ICT 문서를 정확하게 판별하는 정도와 Non ICT 문서를 정확하게 판별하는 정도가 비슷한 것이 본 연구에 더 적합하다고 판단하여 최적의 threshold를 0.5491로 설정하였으며, 최종 출력값이 0.5491보다 높으면 ICT 문서로, 낮으면 Non ICT 문서로 분류하였다. 이때 분류 정확도와 재현율은 약 84.2%였으며(Fig. 4 참조), AUC(Area under the curve)의 값은 약 0.93으로 나타났다(Fig. 5 참조).

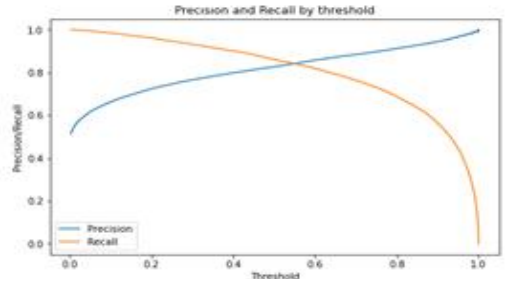

Fig. 4. Precision and Recall by threshold for validataion data

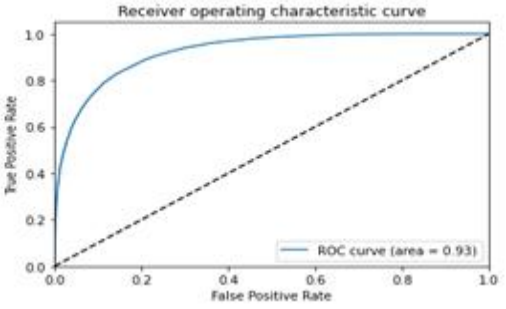

Fig. 5. ROC curve by threshold for validataion data

그리고 threshold까지 정해진 최종 모델에 test data를 적용하여 판별해 본 결과, Table 3과 같은 결과를 보였으며, 약 85%의 정확도를 보였다.

Table 3. Discriminant result for test data

Real \ Pred	Non ICT	ICT	Total
Non ICT	14,068	2,429	16,497
ICT	2,502	13,995	16,497
Total	16,570	16,424	32,994

또한 Table 4는 test data에서 〈과학기술표준분류 1-대〉에 따른 분류 예측 결과이다. 정보/통신 부분은 약 91%의 문서를 ICT분야로 구분하고, 그 외 분야는 약 38%만을 ICT분야로 구분하여 모델이 ICT 문서를 어느정도 잘 구분해 내는 것으로 확인할 수 있었다.

Table 4. Discriminant result by 〈Science and Technology Standard Classification-Large〉 for test data

Science and Technology Standard Classification-Large	Non ICT count (%)	ICT count (%)	Total
information/ communication	229 (8.28%)	2,536 (91.72%)	2,765
others	18,476 (61.12%)	11,753 (38.88%)	30,229

최종 모델을 통해 전체 국가연구개발사업 데이터와 국회 상임위원회 회의록 데이터에서 ICT 문서를 추출하였으며, 각각 국가연구개발사업 데이터에서는 총 87,363건, 회의록 데이터에서는 총 150건의 문서가 ICT 문서로 추출되었으며, 이후 분석 과정은 추출된 ICT 문서를 대상으로 분석을 진행하였다.

4.2 LDA 기반 토픽모델링 분석

토픽모델링의 결과를 평가하는 기준으로 Coherence score와 Perplexity score를 주로 사용한다. Coherence score는 D.Newman 등(2010)이 제안한 척도로 토픽모델링 결과로 나온 주제에 대해 각각의 주제에서 상위 N개의 단어를 뽑고 난 후, 상위 단어 간의 유사도를 계산하여 측정한다. 주제의 일관성을 측정하는 척도로써, 높을수록 해당 모델이 잘 만들어졌고, 토픽 내에 의미적으로 유사한 단어가 많이 모여있다는 의미이다[23]. Perplexity score는 만들어진 토픽모델이 문서 내 토픽의 출현 확률과 토픽 내 단어의 출현 확률

을 실제 값과 얼마나 유사하게 예측하는지를 나타낸다. Perplexity score가 작을수록 생성된 모델이 문서를 잘 파악했다고 볼 수 있다. 하지만 모델이 문서를 잘 학습했다는 의미일 뿐, 사람이 그 결과를 해석하기는 어려울 수도 있다는 단점이 있다.

토픽모델링의 분석의 경우 반복할 때마다 토픽의 순서가 일정하지 않아 이후의 내용은 각 Random seed는 42로 설정한 하나의 모델의 결과이며, 타 모델의 경우, 토픽의 주제 순서는 일정하지 않았지만, 토픽 내에 단어들은 전반적으로 비슷한 결과를 보였음을 알린다.

앞 절에서 ICT 문서로 판별된 87,363건의 국가사업과제 데이터에 대해 토픽의 개수를 1에서 100개까지 설정하여 LDA 모델을 생성하였고 Fig. 6과 Fig. 7은 각각 모델들의 Coherence score와 Perplexity score를 나타내고 있다. Coherence score는 급격히 증가하다가 26개 토픽에서 가장 높은 값을 가지고 이후로 서서히 감소한다. 또한, Perplexity score는 급격히 증가하다가 다시 급격히 감소하고, 기울기가 완만해진 후에는 비슷한 기울기로 감소하고 있는 모양을 볼 수 있다. 본 연구에서는 Coherence score가 가장 높고, 상대적으로 기울기가 안정화된 26개의 토픽 개수를 최종적으로 설정하였다.

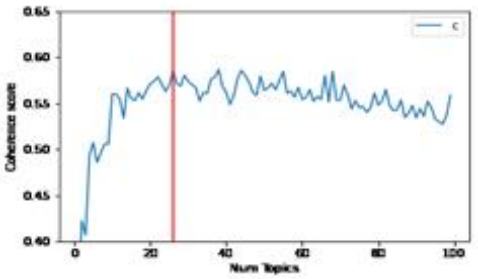

Fig. 6. Coherence score by topic number

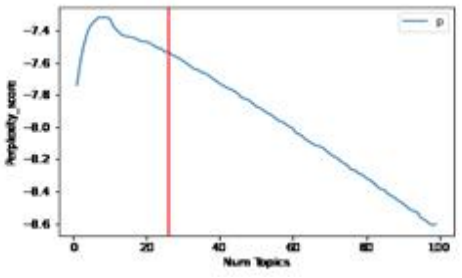

Fig. 7. Perplexity score by topic number

Table 5는 국가사업과제 데이터에 26개의 토픽 개수를 가진 LDA 모델링을 적용하여 만들어진 모델의 토픽에 대한 연관성 지표(relevance) 상위 10개 단어를 나타낸 표이다. 이때 연관성 지표는 아래 식과 같은 형태로 나타낸다. Table 5는 영어로 표기하였으나, 실제로는 한글 자료를 분석한 자료임을 알린다.

$$relevance(wvertt) = \lambda * \log(p(wvertt)) + (1-\lambda) * \log(\frac{p(wvertt)}{p(w)})$$

λ는 토픽 t에서 단어 w가 나올 조건부확률인 $p(w|t)$와, 전체 문서에서 단어 w가 나올 확률보다 토픽 t에서 단어 w가 나올 확률이 상대적으로 얼마나 향상되는지를 나타내는 향상도 $\frac{p(wvertt)}{p(w)}$ 사이의 가중치를 나타낸다. 본 연구에서 λ는 0.5로 가정하였다. LDA 토픽모델링을 사용하여 산출한 토픽들은 각각의 레이블이 자동적으로 만들어지지는 않기 때문에, 토픽별 상위 키워드와 해당 토픽에 배정된 문서들을 기준으로 연구자가 직접 해당 토픽들의 명(Label)을 설정하였다.

Table 5. Top10(relevance) word by topic for R&D project data

Topic	Word in descending relevance(Label)
1	design, module, power, control, circuit, drive, fabrication, communication, performance, motor (System Design and Development)
2	space, logistics, underwater, tunnel, twin, city, residential, ecology, circulation, algae (Port logistics)
3	Language, Korean, English, Satellite, Vocabulary, Vibration, Sentence, Propagation, Sound, Speaker (Translation and Corpus)
4	data, algorithms, information, deep learning, platform, learning, security, artificial intelligence, autonomous, cloud (ICT Theory and Background)
5	cells, nerves, proteins, animals, control, stimulation, drugs, genes, expression, immunity (Life Science)
6	Material, Battery, Electrode, Semiconductor, Sun, Electric, Thin Film, Energy, Cell, Electronics (new electrical energy)
7	education, society, activities, music, sports, policies, schools, teachers, welfare, classes (Elementary and secondary ICT education)
8	marine, ship, climate, weather, data, accidents, nuclear power plants, surveillance, systems, forecasting (Smart Ship)
9	content, behavior, media, games, personal, advertising, recommendations, movies, finance, virtual reality (Personal information and customized services)
10	patient, medical, clinical, treatment, health, hospital, disease, rehabilitation, dementia, disability (Smart Hospital)
11	dimension, walking, robot, artificial, posture, exercise, elderly, band, structure, audio (Walking Robot)

Table 5. Continued

Topic	Word in descending relevance(Label)
12	diagnosis, prediction, genomic, data, database, model, mutation, virus, infection, disease (Digital Bio Healthcare)
13	theory, method, quantum, problem, calculation, function, model, math, suggestion, case (Computing System)
14	fire, detector, cut, matrix, tag, sign, collision, arc, alarm, detection (Fire Safety)
15	image, measurement, signal, laser, detection, ultrasonic, optical, light source, noise, shooting (Image and signal processing)
16	Culture, Literature, History, Works, Politics, Korea, Modern, Women, Philosophy, Thought (Humanities ICT)
17	smart, sensor, design, device, mobile, function, app, smartphone, interlock, service (Smart Device)
18	agency, participation, gas, process, university, supervising, hydrogen, evaluation, parts, manufacturing (Converged Smart Process)
19	standards, safety, testing, certification, disaster, assessment, transportation, procedures, railways, management (Smart Mobility)
20	enterprise, business, support, operations, industry, linkage, drive, build, password, strategy (ICT industrial support)
21	printing, filter, printer, fiber optic, edge, fingerprint, shape, disk, hologram, screen (3D printing)
22	training, education, manpower, convergence, field, professional, international, competency, university, team (training of ICT professionals)
23	product, production, automatic, processing, work, us, quality, sales, process, sales (Manufacturing Smart Factory)
24	food, microorganisms, plants, agriculture, crops, medicines, ingredients, cosmetics, cultivation, organisms (Smart Farm)
25	nano, matter, chemical, particle, synthesis, surface, reaction, molecule, polymer, fluid (ICT Nano Tech)
26	energy, structure, construction, building, construction, construction, construction, housing, wind, solar, building (Renewable energy)

다음 쪽의 Fig. 8은 국가사업과제 데이터와 회의록 문서가 LDA 모델에 의해 각 토픽별로 분류되는 분포를 그래프로 나타낸 것이다. 국가사업과제 데이터로 도출된 토픽들에 대해 회의록 문서는 Topic9, 16, 20, 19, 13, 7 등의 특정 주제에 대해서 이외 토픽에 비해 눈에 띄게 높은 문서 분포를 나타냈다. 이는 각각 '개인정보 및 맞춤 서비스 등', '인문ICT콘텐츠', 'ICT 산업 지원', '스마트 모빌리티 시스템', '컴퓨팅 시스템', '초/중등 ICT 관련 교육' 등에 관한 토픽으로 타 토픽에 비해 사회적이고 대중적인 내용에 해당한다. 이에 비해 실제로 수행되는 국가사업과제는 Topic4나 1과 같이 ICT 분야에대한 기초적인 이론 및 배경과 시스템 설계 및 개발과 같은 부분의 연구를 많이 수행하고 있고, Topic23, 18 등과 같은 스마트 공장 및 공정 시스템 관련 내용을 많이 다루고 있다는 차이를 보였다. 두 데이터에서 함께 상위권을 차지하고 있는 토픽은 Topic20인 ICT 산업 지원에 관한 토픽뿐이었다.

5. 결론

본 연구에서는 국가 ICT 산업 정책과 긴밀한 접점이 있는 서로 다른 두 집단에 대해 데이터에 기반한 ICT 분야 인식 차이를 비교해 보는 프로세스에 대해 다루었다. 먼저 HAN 모델에 국가사업과제 데이터를 학습시켜 ICT 문서와 Non ICT 문서로 분류하는 모델을 만들었고, 해당 모델에 전체 국가사업과제 데이터와 국회 상임위원회 회의록을 대입하여 ICT 분야와 연관이 있는 국가사업과제 데이터와 회의록을 추출하였다. 다음으로 ICT 문서로 분류된 국가사업과제 데이터에 LDA 토픽모델링 모델을 적용하여 토픽을 도출하였으며, 해당 모델에 회의록 또한 대입하여 토픽별 주요 단어를 살펴보고 국가사업과제와 회의록의 수치적인 문서 확률 분포를 도출하여 국가사업과제와 회의록이 주로 어떤 토픽을 다루었는지 확인하고 비교해 보았다.

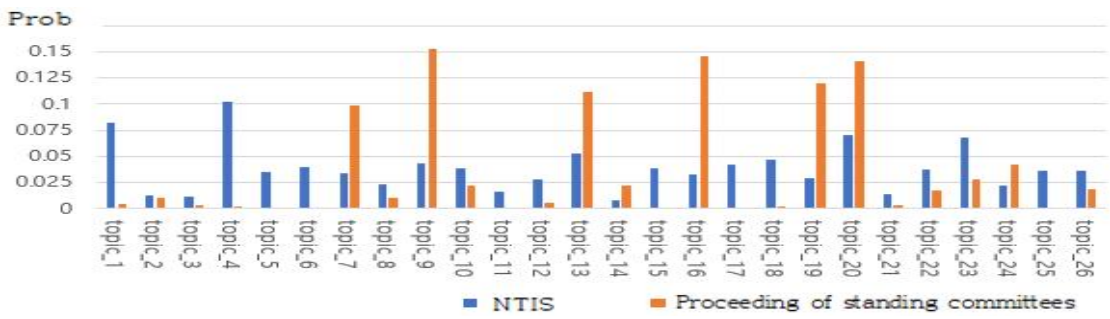

Fig. 8. Distribution of document by Topic

구체적으로 HAN 모델 분류 결과, 국가사업과제 데이터에 대해 실제 정보/통신 분야의 문서는 약 91% 정도를 ICT분야로 구분하였고, 그 외 분야는 약 38% 정도를 ICT분야로 구분하였다. 세계적으로 ICT 기술의 중요성이 확대되고 있는 흐름에 맞게 여러 분야의 사업과제들이 ICT 분야와 융합하여 수행되고 있는 것으로 보인다.

ICT 문서로 분류된 국가사업과제 데이터를 LDA 토픽모델링 기법을 적용하여 총 26개의 토픽을 도출하였으며, 각 토픽에 속한 주요 단어와 그 문서의 비중을 비교해 봤을 때, 국가사업과제는 이론적이고 전문적인 내용에 비중을 두고 여러 토픽에 대하여 과제를 수행하고 있지만, 국회 상임위원회에서는 대중적인 내용에 치우쳐져 있는 모습을 보였다.

이는 몇가지 이유를 추측해 볼 수 있다. 첫째는, ICT 이론이나 기법과 관련해서는 빠르게 변화하고, 적용범위가 급격히 넓어지고 있기 때문에 관련한 연구는 각 분야의 정부 출연 연구원이나 부처에서 사업과제로 수행하고 있지만, 이를 정책적으로 다루는 집단의 수는 상대적으로 적고 실제 회의에서 논의가 되기까지 오랜 시간이 걸리기 때문에 차이가 난다고 볼 수 있다. 둘째는, 실질적 이론이나 기법에 대해서는 연구 국회 상임위원회에서 관련 내용에 대해 논의하는 과정에서 분야·부문이나 프로그램 등의 큰 구분에 관한 내용까지는 논의가 되지만 전문적인 내용이 많이 포함되어있는 세부사업을 모두 비중 있게 다루기에는 한계가 있기 때문에 분석 결과에 잘 나타나지 않는 것으로 볼 수 있다. 실제 실무적으로 수행하는 연구와 정책적으로 다루어지는 분야가 차이를 보이는 이유에 대해 향후 분석을 수행하여 유의한 근거를 찾아낼 수 있다면 해마다 정책적으로 혹은 과제 실무적으로 어떤 주제에 투자와 지원을 하고 발전을 시켜나가야 할지에 대한 의사결정을 하는 데 객관적인 증거 자료로 활용할 수 있는 참고자료가 만들어질 수 있을 것이다.

본 연구에서 파악된 연구 이슈는 다음과 같다. 첫째로 용어사전을 구축하는 과정에서 최대한 신뢰성있고 많은 단어를 포함하고 있는 사전을 활용하였지만, ICT 분야에서 전문적으로 활용되고 있는 단어들이나, 결합단어들이 용어사전에 포함되어 있지 않아 관심 대상임에도 불구하고 실제 분석에서는 배제되는 경우가 있었다. 예를 들어, '통합검색'이나 '보안드라이브'같은 단어의 경우 각각 '통합'/'검색', '보안'/'드라이브'로 나뉘어 분석에 활용되는 경우가 있었다. 두 번째로 본 연구에서 연구자가 임의로 모델의 Shape를 결정하고, ICT 문서의 정의를 ICT 단어의 포함 개수만을 기준으로 정하는 등 경험적 결과를 바탕으로 편의상 조작적 정의를 하고 넘어간 부분이 있어 연구 타당도가 다소 부족하다는 점이 있다. 해당 부분은 추후 실무적 활용을 위해서 반드시 신뢰성 있는 연구를 통하여 해결해야 할 문제이다.

본 논문의 결과로 5년에 대한 전체 연구 주제에 대한 결과만을 보여주었지만, 그 결과가 계량적인 수치로 나타나는 만큼, 연도별 분포와 그 선후관계에 대한 추가적인 연구가 이루어 진다면 좀 더 연구가 정책이나 과제에 영향도를 객관적으로 평가하는 데 도움이 될 것이다. 또한 본 논문에서는 국가연구개발과제와 국회회의록 자료만을 비교하였지만, 연구기관의 보고서 및 정기간행물, 뉴스, 학술 논문, 다양한 온라인 채널(SNS, 유튜브 등), 해외 정책자료 등의 문서를 수집하여 비교분석하는 프로세스에 대하여 추가적으로 연구해 본다면 사회 전반, 학계, 정책연구기관을 대표하는 평가 지표를 개발하는데 참고할 수 있는 자료가 될 것이다.

REFERENCES

[1] *NATIONAL RESEARCH AND DEVELOPMENT INNOVATION ACT.* (2020). Act No. 17343.

[2] H. S. Kim. (2015). *Study of program characteristics by performance criteria for evaluation of national research and development projects.* Jincheon-gun, Chungcheongbuk-do : KISDI.

[3] National Assembly Budget Office. (2020). *A study on the evaluation system of project planning for national R&D projects.* Seoul : NABO.

[4] E. Jimenez-Contreras, F. M. Anegon & E. D. Lopez-Cozar. (2003). The evolution of research activity in Spain: The impact of the National Commission for the Evaluation of Research Activity(CNEAI). *Research Policy, 32(1),* 123-142.

[5] Korea Legislation Research Institute. (2019). *The study for data-based legislative assessment methodology.* Sejong : KLRI.

[6] National Assembly Budget Office. (2007). *A study on the project evaluation methodology.* Seoul : NABO.

[7] Korea Institute of Science and Technology Evaluation and Planning. (2018). *A study on the*

establishment of decision support system for research innovation policy based on bigdata. Eumseong-gun, Chungcheongbuk-do : KISTEP.

[8] Science and Technology Policy Institute. (2020). *Innovation Strategy for the data-based R&D management system of the Korean government.* Sejong : STEPI.

[9] G. Y. Rhee, S. C. Park & S. Y. Ryoo. (2020). *Performance measurement model for open bigdata platform.* Daegu : NIA.

[10] H. Y. Yang. (2012). *Technology Planning Methodology Using Big Data.* Eumseong-gun, Chungcheongbuk-do : KISTEP.

[11] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola & E. Hovy. (2016). Hierarchical attention networks for document classification. *Proceedings of the Conference of the North American Chapter of the Association for Computaional Linguistics: Human Language Technologies,* 1480-1489. DOI : 10.18653/v1/N16-1174

[12] I. H. Jang, K. Y. Park & J. K. Lee. (2018). Analysis of the online review based on the theme using the hierarchical attention network. *Journal of information technology services: Korea Society of IT Services, 17(2),* 165-177. DOI : 10.9716/KITS.2018.17.2.165

[13] S. Y. Woo. (2019). *Classification of civil appeals using hierarchical attention network focusing on Seoul civil appeas data.* Thesis of Master's Degree, Yonsei University, Seoul.

[14] H. C. Jang, D. H. Han, T. S. Ryu, H. K. Jang & H. S. Lim. (2018). Patent Document Classification by Using Hierarchical Attention Network. *Proceedings of the Korea Information Processing Society Conference : Korea Information Processing Society,* 369-372. DOI : 10.18653/v1/N16-1174

[15] D. Bahdanau, K. H. Cho & Y. Bengio. (2015). Neural machine translation by jointly learning to align and translate. *Proceedings of the 3rd International Conference on Learning Representaions (ICLR 2015).* DOI : arXiv:14090473

[16] N. Pappas & A. P. Belis. (2017). Multilingual hierarchical attention networks for document classification. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* DOI : arXiv:170700896

[17] S. M. Heo & J. Y. Yang. (2020). Analysis of research topics and trends on COVID-19 in Korea using latent dirivhlet allocation(LDA). *Journal of The Korea Society of Computer and Information : Korea Society of Computer Information, 25(12),* 83-91. DOI : 10.9708/jksci.2020.25.12.083

[18] S. H. Moon, S. H. Chung & S. H. Chi. (2018). Topic modeling of news article about international construction market using latent dirichlet allocation. *Journal of the korean society of civil engineers: Korean Society of Civil Engineers, 38(4),* 595-599. DOI : 10.12652/Ksce.2018.38.4.0595

[19] J. H. Park & M. Song. (2013). A Study on the Research Trends in Library & Information Science in Korea using Topic Modeling. *Journal of the Korean Society for Information Management: Korea Society for Information Management, 30(1),* 7-32. DOI : 10.3743/KOSIM.2013.30.1.007

[20] D. M. Blei, A. Y. Ng & M. I. Jordan. (2003). Latent dirichlet allocation. Journal of Machine *Learning Research, 33,* 993-1022.

[21] J. H. Park & H. J. Oh. (2017). Comparison of Topic Modeling Methods for Analyzing Research Trends of Archives Management in Korea: focused on LDA and HDP. *Journal of Korean Library and Information Science Society (JKLISS), 48(4),* 41-61.

[21] H. S. Kang & J. H. Yang. (2018). Selection of the Optimal Morphological Analyzer for a Korean Word2vec Model. Korea Information *Processing Society's 2018 Autumn Academic Conference*, 376-379.

[22] H. K. Jung & N. K. Kim. (2018). Analyzing the Effect of Characteristics of Dictionary on the Accuracy of Document Classifiers. *Management & information systems, 37(4),* 41-61.

[23] D. Newman, J. H. Lau, K. Grieser & T. Baldwin. (2010). Automatic Evaluation of Topic Coherence. *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, 100-108.*

송 병 기(Byoungki Song) **[정회원]**

· 1997년 3월 : 한양대학교 경영학과 (경영학사)
· 2013년 2월 : 한양대학교 경영학과 (경영석사)
· 1997년 4월~현재 : 정보통신정책연구원 책임행정원

· 관심분야 : 경영, 마케팅
· E-Mail : gaebak@kisdi.re.kr

김 상 웅(Sangung Kim) [정회원]

· 2018년 2월 : 경북대학교 통계학과 (통계학 학사)
· 2021년 2월 : 경북대학교 통계학과 (통계학 석사)
· 2021년 4월~현재 : 정보통신정책 연구원 행정원

· 관심분야 : 통계학, ICT, 빅데이터
· E-Mail : tkddnd0214@kisdi.re.kr