Journal of Next-generation Convergence Information Services Technology Vol.13, No.1, February (2024), pp.93-102
ISSN 2384-101X(p) 2672-1163(e) *http://dx.doi.org/10.29056/jncist.2024.02.09*

자연어처리와 기계학습을 활용한 기술 특허 분류

Classification of Technology Patents Using Natural Language Processing and Machine Learning Models

이우식[1]*, 이예진[2]

Woosik Lee[1]*, Ye Jin Lee[2]

요 약

최근 빅데이터 시대의 도래로 인공신경망을 포함한 기계학습 모델들이 의학, 유전체 연구, 기업 경영 등 다양한 분야에 광범위한 영향을 미치고 있음에도 불구하고, 기술 특허 분석에 자연어 처리와 기계학습을 적용한 국내 리걸테크 연구는 충분히 발전하지 못한 상황이다. 본 연구는 이산화탄소 포집·활용에 대한 특허 데이터, 자연어 전처리 기법 그리고 기계학습모형 기반의 기술 특허 분류 시스템을 설계하고, 정확도, 카파 상관계수 그리고 F1-점수를 비교·분석하였다. 주요 결과를 요약·정리하면 다음과 같다. 첫째, 다섯 가지 이산화탄소 포집 및 활용 기술 분류에서 그래디언트 부스팅, 랜덤 포레스트, 의사결정나무 순으로 성능이 나타났다. 이를 통해 단일 결정 나무보다 배깅과 부스팅 기법을 적용한 랜덤포레스트 모형과 그래디언트 부스팅 모형이 더 우수한 학습 성능을 제공함을 확인할 수 있었다. 둘째, 특허의 요약과 제1 청구항을 활용한 기술 분류에서 비슷한 성능이 관찰되었다. 이는 자연어 처리 과정에서 중요한 키워드를 명사로만 추출한 것이 주요 요인으로 보인다. 본 연구는 자연어 전처리와 기계학습 모형을 이산화탄소 포집 및 활용 기술 특허 분류에 처음으로 적용한 의미 있는 연구로 사무 로봇 기술을 통해 반복적인 업무를 자동화하는 데 응용될 수 있는 가능성을 제시한다.

핵심어 : 비즈니스 애널리틱스, 자연어 처리, 특허, 비즈니스 의사 결정, 로봇 프로세스 자동화

Abstract

With the advent of the big data era, machine learning models, including artificial neural networks, have had a wide-ranging impact on various fields such as medicine, genomics research, and corporate management. Despite this, domestic research in legal tech, particularly applying natural language processing and machine learning to technical patent analysis, has not sufficiently developed. This study designs a system for classifying patents on Carbon Dioxide Capture and Utilization (CCU) based on patent data, natural language pre-processing techniques, and machine learning models, and compares and analyzes accuracy, kappa coefficient, and F1-score. The main findings are summarized as follows: First, in classifying five types of CCU technologies, the performance was observed in the order of gradient boosting, random forest, and decision trees. This confirms that random forest and gradient boosting models, which apply bagging and boosting techniques, respectively, provide superior learning performance over single

1 College of Business Administration, Gyeongsang National University, Jinju, Korea [Professor]
e-mail: woosiklee@gnu.ac.kr (Corresponding author)
2 College of Law, Gyeongsang National University, Jinju, Korea [Undergraduate Student]
e-mail: lein1224@gnu.ac.kr

Received(November 21, 2023), Review Result(1st: December 11, 2023), Accepted(February 9, 2024), Published(February 29, 2024)

decision trees. Second, similar performance was observed in classifying technologies based on the abstract and first claim of patents. This suggests that the extraction of important keywords as nouns during the natural language processing is a significant factor. This research is meaningful as it applies natural language pre-processing and machine learning models to the classification of CCU technology patents for the first time, presenting the potential for applying robotic automation technology to automate repetitive tasks.

Keyword : Business Analytics, Natural Language Processing, Patent, Business Decision-Making, Robotic Process Automation

1. 서론

최근 기후변화를 연구하는 비영리단체 클라이밋 센트럴의 발표에 따르면, 2022년 11월부터 23월 10월까지의 지구 평균기온이 산업화 이전 시기인 1850년~1900년보다 섭씨 1.32도 높아 '가장 더운 12개월'로 기록되었다고 밝혔다 [1]. 또한, 국제적으로 기후변화에 대해 심각성을 인지하고 1980년대부터 국제사회 차원에서 지속적으로 노력했으나 아직까지 전 세계적 노력이 가시적인 효과를 거두지 못하였다 [2]. 이로 인해 2015년 12월 프랑스 파리에서 열린 제 21차 유엔기후변화협약 당사국총회에서 파리기후변화협약이 발표되었는데 [3], 파리협정(Paris Climate Agreement)에 거의 대부분 국가들이 온실가스 감축에 참여하여 지구 평균 온도 상승을 2도 아래에서 억제한다고 명시되어 있다. 이러한 목표를 달성하기 위해 전 세계 국가들이 노력 중이며, 대한민국도 2030년까지의 온실가스 감축목표를 2017년 대비 24.4%로 제시하였다 [4].

온실가스 중 가장 많은 배출량을 차지하는 이산화탄소(CO_2)를 감축하는 것이 지구온난화를 방지하는데 효율적이기 때문에 이산화탄소를 직접 제거할 수 있는 이산화탄소 포집·활용·저장(Carbon Capture Utilization & Storage, CCUS) 기술이 주목받고 있다. 국제에너지기구(International Energy Agency)는 '2070 글로벌 탄소중립' 과정에서의 CCUS의 기여도를 이산화탄소 전체 감축량의 15% 수준으로 제시하고, CCUS 없이 넷제로(Net Zero)에 도달하는 것이 불가능하다고 전망하였다 [5].

하지만 CCUS에 대한 지속적인 투자와 연구성과물이 창출됨에도 불구하고 상용화가 더디게 진행되고 있다 [6]. 이에 따라, CCUS 프로젝트의 양적·질적 성장을 모색하기 위한 측면에서 축적된 특허 데이터에 근거한 분석이 필요한 시점이다. 최근 국내에서 수행된 CCUS 관련 특허 연구는 김정민 외 5인의 연구 [2] 와 서재호 및 이동명의 연구 [7] 등이 있으나, 이 분야에 대한 분석은 아직 충분하지 않은 상태다. 김정민 외 5인의 연구 [2]는 2000년부터 2020년까지의 해당 기술에 대한 주요 국가별, 출원인별, 기술 분야별 특허 동향을 분석하였으며, 특히 이산화탄소 포집 분야에서 상당한 특허 출원이 이루어졌음을 밝혔다. 서재호와 이동명의 연구 [7] 는 시멘트, 철강, 발전 산업에 초점을 맞추고 광물탄산화 기술에 대한 기술 개발 활동성 및 향후 기술의 성숙기를 바스 확산 모형(Bass Diffusion Model)와 곰페르츠 모형(Gompertz Model)을 통해 추정하였다. 그러나 대부분의 선행 연구는 주로 특허 출원의 집중도와 광물탄산화 기술의 성숙기 예측에 국한되어 있다. 이에 본

연구는 CCUS의 하위 분야이자 이산화탄소를 유용한 물질로 전환하는 기술인 이산화탄소 포집·활용(Carbon Capture Utilization, CCU)의 특허 자료를 자연어처리(Natural Language Processing, NLP)와 기계학습 모형으로 분석, 기술 특허 분류를 통해 선행연구와의 차별성을 강조하고자 한다.

본 논문은 다음과 같이 4개 장(章)으로 구성되어 있다. 제1장 서론에서는 이 연구를 수행하게 된 배경과 필요성에 대해 설명한다. 제2장에서는 본 연구의 핵심인 기계학습 알고리즘과 자연어 처리에 대한 이론적 근거를 검토하고 요약하여 제시한다. 제3장에서는 실증 연구를 위한 방법론을 정립하고, 실제 분석을 통해 얻은 결과를 밝힌다. 마지막으로, 제4장에서는 연구 결과의 의의와 시사점을 도출함으로써 연구의 결론을 마무리한다.

2. 이론적 배경

2.1 기계학습 알고리즘

본 연구에서는 의사결정나무의 응용모형을 중심으로 자료 분석을 진행하고자 한다. 의사결정나무 모형은 독립변수 공간을 계층적으로 분할된 공간 내의 종속변수의 대표값을 종속변수의 추정치로 활용한다. 이 모형의 학습과정은 추정오차를 가장 작게 만드는 독립변수 공간을 결정하는 과정이다. 이러한 의사결정나무 모형은 학습표본에 과적합 되는 경향이 강하기 때문에, 여러 개의 소규모 의사결정나무를 만들어 각각의 의사결정나무 추정치의 대표값을 추정값으로 활용하는 의사결정나무 응용모형인 랜덤포레스트(Random Forest) 모형과 그래디언트 부스팅(Gradient Boosting) 모형을 사용한다 [8].

다양한 분류 및 예측 문제에 널리 사용되는 랜덤포레스트 모형은 독립변수를 임의로 선택하여 학습하는 방식인 배깅(Bagging) 기법을 채택하고 있다. 배깅은 '부트스트랩 집계(Bootstrap Aggregating)'의 줄임말로, 원본 데이터로부터 같은 크기의 표본을 여러 번 복원 추출하는 방법이다. 이렇게 복원 추출된 표본을 '부트스트랩 표본'이라고 한다. 각 부트스트랩 표본에 대해 독립적으로 생성된 분류기들의 결과를 종합하여 최종 결정을 내리 것이 특징이다. 문선영의 연구 [9]에 따르면, 랜덤 포레스트의 가장 큰 특징은 무작위성(Randomness)을 통해 서로 조금씩 다른 특성을 갖는 여러 의사결정나무를 구성한다는 점이다. 이는 의사결정나무들 간의 상관성이 낮추고, 각 의사결정나무의 편향을 유지함으로써 전체 모형의 정확성과 안정성을 향상시킨다.

그래디언트 부스팅은 여러 개의 예측력이 약한 모형을 결합하여 더 강력한 모형을 만드는 앙상블(Ensemble) 모형 중 하나로서 역시 예측과 분류 등에 자주 활용된다. 본 모형은 앞서 학습된 모형이 잘못 예측한 데이터에 대해 가중치를 증가시키면서, 연속적으로 모형을 학습시켜 전체 모형의 성능을 향상시킨다. 초기에 모든 관측치는 동일한 가중치를 받지만, 모형이 예측을 수행한 후 오분류된 데이터에는 더 높은 가중치를, 정확히 분류된 데이터에는 낮은 가중치를 부여한다. 이렇

게 조정된 가중치를 바탕으로 데이터를 다시 샘플링하고, 새로운 모형을 학습하는 과정을 반복한다. 최종적으로는 각 모형의 예측 결과에 가중치를 적용하여 가중 평균을 계산함으로써, 전체 앙상블 모델의 최종 예측값을 도출하게 된다.

2.2 자연어 처리

자연어 처리는 인간의 언어를 기계학습 또는 딥러닝이 처리하고 의미를 분석할 수 있도록 하는 기술 분야이다. 자연어 처리 적용 주요 분야로 기계 번역, 감정 분석, 자동 요약 등으로 GPT(Generative Pre-training Transformer) 등장 이후 가장 이목이 쏠리는 분야 중 하나이다. 자연어 처리 모형은 단어 수준에서 의미를 포착하는 워드 임베딩(Word Embeddings)에서 대규모 텍스트 데이터셋에서 사전 학습되어, 언어의 광범위한 문맥적 패턴을 학습한 사전 학습된 언어 모델(Pre-trained Language Models)로 고도화 되고 있다. 본 연구에서는 제한된 기술 특허 데이터와 특정 문서 내에서 의미 있는 단어를 식별해야 하는 경우로 카운트 기반의 워드 임베딩인 TF-IDF(Term Frequency-Inverse Document Frequency) 방법을 적용하였다.

$$TF-IDF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \frac{|D|}{|d_j : t_i \in d_j|} \quad (1)$$

식(1)에서 i는 단어 순서별 번호, j는 문서 순서별 번호, $n_{i,j}$는 문서 d_i에서 단어 t_i가 나오는 빈도수, $\sum_k n_{k,j}$는 문서 d_i에서 나오는 모든 단어빈도 수이다. 그리고 $|d_j : t_i \in d_j|$는 t_i가 나오는 문서 수이고 $|D|$는 전체 문서 수이다. TF는 문서 내에서 특정 단어가 나타나는 빈도수이고, IDF는 전체 문서 중 해당 단어를 포함하는 문서의 수에 역수를 취한 값이다. TF-IDF 값은 TF 값과 IDF 값의 곱으로 계산되며, 특정 단어가 소수의 문서에서 자주 등장하고 전체 문서 중 그 단어를 포함하는 문서 수가 적을수록 TF-IDF 값은 커지고, 이는 단어의 중요도가 높음을 나타낸다 [10].

2.3 평가지표

본 연구에서는 분류 모형의 성능 지표로 정확도(Accuracy), 카파 상관계수(Kappa Coefficient), F1-점수(Score)를 고려하였다. 정확도는 모든 관측치의 관측값과 모형의 예측값이 일치하는지를 나타내는 비율이다. 만약 모형의 정확도가 90%이면, 100개의 관측치 중 90개가 정확하게 분류되었다는 것을 의미한다. 카파 상관계수는 일반적으로 코헨의 카파 상관계수를 가리키며, 관측값과 예측값의 일치도를 측정하는 방법이다. 카파 상관계수의 등급은 [표 1]와 같이 Landis와 Koch의 해석을 주로 따른다 [11]. F1-점수는 정확도(Precision)와 재현율(Recall)의 조화평균을 이용하여 분류 모형의 성능을 평가하는 지표이다.

[표 1] 코헨의 카파 통계량을 통한 해석
[Table 1] Interpretation of Cohen's Kappa Statistic

코헨의 카파 통계량	해석
< 0	Poor
0.0 ~ 0.2	Slight
0.2 ~ 0.4	Fair
0.4 ~ 0.6	Moderate
0.6 ~ 0.8	Substantial
0.8 ~ 1.0	Almost Perfect

3. 실증분석

3.1 자료의 구성

실증 분석을 위해 사용한 표본은 CCU 기술을 [표 2]와 같이 다섯 가지 대분류로 나누었다. 이후, 특허 검색 프로그램인 WIPS ON(윕스 온)을 사용하여 [표 3]와 같이 총 1,190건의 특허를 검색했다. 이후 노이즈 데이터를 제거하여 최종적으로 한국의 등록 특허 329건의 '제1 청구항'과 '요약'을 수집하였다.

[표 2] 5가지 CCU 기술 분류
[Table 2] Five Types of CCU Technologies

기술 분류	기술 정의
화학적 전환 기술	이산화탄소를 환원 혹은 개질하는 화학적 전환을 통해 고분자의 원료,고분자,플라스틱, dmc, polycarbonate polyol,폴리올,합성가스등을 제조하는 기술.
전기화학적 전환 기술	전기화학전지에 전기에너지를 공급하여 양 전극에 생기는 전위차를 이용하여 이산화탄소를 환원시켜 기체 혹은 액체의 화합물로 전환하는 기술.
광화학/광전기화학적 전환 기술	태양에너지를 반도체 광촉매를 이용하여 전기에너지로 전환 후,이산화탄소가 용해된 전해질과 빛을 흡수한 반도체 광촉매 계면에서 이산화탄소를 환원시키는 기술.
생물 전환 기술	이산화탄소를 고정하는 미생물을 이용하여 이산화탄소를 바이오연료,유용물질로 전환하는 기술. 전기에너지로 환원력(전자)를 공급받는 비광합성,전기생합성 미생물을 이용하는 기술과 광합성 미생물,미세조류를 이용하여 식품용 색소,영양 보충제,화장품,의약품의 원료,자외선차단제,생분해성 원료를 만드는 기술.
광물탄산화 기술	고체 혹은 액체의 알칼리성 부산물,수용액,알칼리 토금속을 포함하는 천연광물을 이산화탄소와 직접 반응시켜 탄산마그네슘,탄산칼슘,중조 등의 탄산염광물로 만들어 이산화탄소를 고정화하고,무기탄산염을 얻는 기술,혹은 원료에서 알칼리 이온을 용출하여 탄산화하는 기술.

[표 3] 특허 검색식

[Table 3] Patent Search

검색 대상 기술 및 키워드 / 검색식 기호	검색식
보편적 ccu 키워드 (A1)	((((이산화탄소 카보온다이옥사이드 carbondioxide carbondioxid carbon-di-oxide 카본다이옥싸이드 카본다이옥시드 carbon-dioxi carbon-dioxid carbon-di-oxid 카본다이옥씨드 carbondioxi 이산화카본 co2 씨오투 이산화-탄소 2산화탄소) near 3 (전환 변환 변화 변경 체인지 utilize* utilization change* use* apply*)) or ccu)) AND (((이산화탄소 카보온다이옥사이드 carbondioxide carbondioxid carbon-di-oxide 카본다이옥싸이드 카본다이옥시드 carbon-dioxi carbon-dioxid carbon-di-oxid 카본다이옥씨드 carbondioxi 이산화카본 co2 씨오투 이산화-탄소 2산화탄소) and (전환 변환 변화 변경 체인지 utilize* utilization change* use* apply*)) or ccu).TI.
이산화탄소의 화학적 전환 (A2)	(이산화탄소 카보온다이옥사이드 carbondioxide carbondioxid carbon-di-oxide 카본다이옥싸이드 카본다이옥시드 carbon-dioxi carbon-dioxid carbon-di-oxid 카본다이옥씨드 carbondioxi 이산화카본 co2 씨오투 이산화-탄소 2산화탄소) and ((((수소화 수소처리 hydrogenated hydrogenolysis 수소첨가 하이드로제네이션 히드로게네이션 환원)) 개질 reforming reform refrom 바꾸 바꿔 change transform convert shift 전환 화학전환 (chemical adj1 (change transform convert)) 변환 환원 ccu)) AND ((이산화탄소 카보온다이옥사이드 carbondioxide carbondioxid carbon-di-oxide 카본다이옥싸이드 카본다이옥시드 carbon-dioxi carbon-dioxid carbon-di-oxid 카본다이옥씨드 carbondioxi 이산화카본 co2 씨오투 이산화-탄소 2산화탄소) and (바꾸 바꿔 change transform convert shift 전환 화학전환 (chemical adj1 (change transform convert)) 변환 환원 ccu)).TI
이산화탄소의 전기화학적 전환 (A3)	((이산화탄소 카보온다이옥사이드 carbondioxide carbondioxid carbon-di-oxide 카본다이옥싸이드 카본다이옥시드 carbon-dioxi carbon-dioxid carbon-di-oxid 카본다이옥씨드 carbondioxi 이산화카본 co2 씨오투 이산화-탄소 2산화탄소) and (전기화학 electrochemi 일렉트로케미컬 electrochemical 일렉트로케미칼 electrochem electro-chmi electro-chemi 전극반 전위차 전위 전극 battery 배터리) and ((전환 change transform* utiliz* 변환) or (제조 방법 장치 device 공정 생산 make method product produce producing))) AND (전기화학 electrochemi 일렉트로케미컬 electrochemical 일렉트로케미칼 electrochem electro-chmi electro-chemi 전극반 전위차 전위 전극 battery 배터리) and (전환 change transform* utiliz* 변환).TI.
이산화탄소의 광화학/광전기화학적 전환 (A4)	((이산화탄소 카보온다이옥사이드 carbondioxide carbondioxid carbon-di-oxide 카본다이옥싸이드 카본다이옥시드 carbon-dioxi carbon-dioxid carbon-di-oxid 카본다이옥씨드 carbondioxi 이산화카본 co2 씨오투 이산화-탄소 2산화탄소) and (광화학 빛화학 photoelectrochemical aficial-photosynthesis 광전기-화학 photoelectrode solar-panel 태양에너지 태양광 태양력 태양열 sun-power sunpower solar solor 솔라 쏠라 솔러 photovoltc photovolataic sunlight photovol 인공광합성 광촉매 photocatalyst photo-catalyst 빛촉매 광화학촉매 photocatalysis)) AND ((이산화탄소 카보온다이옥사이드 carbondioxide carbondioxid carbon-di-oxide 카본다이옥싸이드 카본다이옥시드 carbon-dioxi carbon-dioxid carbon-di-oxid 카본다이옥씨드 carbondioxi 이산화카본 co2 씨오투 이산화-탄소 2산화탄소)).TI.

이산화탄소의 생물 전환 (A5)	((이산화탄소 카보온다이옥사이드 carbondioxide carbondioxid carbon-di-oxide 카본다이옥싸이드 카본다이옥시드 carbon-dioxi carbon-dioxid carbon-di-oxid 카본다이옥씨드 carbondioxi 이산화카본 co2 씨오투 이산화-탄소 2산화탄소) and (((생물 bio 생물화학) adj1 (전환 변환 trans*)) 미세조류 microalgae algae 남조류 phytoplankton 미생물 microorganism 박테리아 비광합성 전기생합성 bacteria micro-organism electrobio* 원생동물 microbic microorgan bioconversion bioconvert)) AND ((이산화탄소 카보온다이옥사이드 carbondioxide carbondioxid carbon-di-oxide 카본다이옥싸이드 카본다이옥시드 carbon-dioxi carbon-dioxid carbon-di-oxid 카본다이옥씨드 carbondioxi 이산화카본 co2 씨오투 이산화-탄소 2산화탄소)).TI.
이산화탄소의 광물탄산화 기술 (A6)	((이산화탄소 카보온다이옥사이드 carbondioxide carbondioxid carbon-di-oxide 카본다이옥싸이드 카본다이옥시드 carbon-dioxi carbon-dioxid carbon-di-oxid 카본다이옥씨드 carbondioxi 이산화카본 co2 씨오투 이산화-탄소 2산화탄소) and (슬래그 철강* 제강* financ* industria* 부산물 by-product byproduct) and (알칼리성 알카리성 alkari* 알카리용액 알칼리수 알칼리용액 탄산화 중탄산염 탄산칼슘 황산암모늄 중조 caco3 nahco3 무기탄산염 inorganic) and (광물탄산화 광물화 mineraliz* 광물자원화 carbonate*))
Cooperative Patent Classification(CPC) (A7)	(B01J* C12M* H01M* C12P* C12M* C01F* C12Y* A01G* B01D* C25B* C04B* Y02C* Y02E* Y02P* C07C* C12N* C02F* C01B* C08B* C08L* C10* C01F* Y02W* C08G* H01L* C01D* C25C*).CPCM.
발명의 명칭 (A8)	(이산화탄소 카보온다이옥사이드 carbondioxide carbondioxid carbon-di-oxide 카본다이옥싸이드 카본다이옥시드 carbon-dioxi carbon-dioxid carbon-di-oxid 카본다이옥씨드 carbondioxi 이산화카본 co2 씨오투 이산화-탄소 2산화탄소).TI.

3.2 모형의 추정 및 분석

[표 2]에 나타난 바와 같이, CCU 기술을 다섯 가지 주요 분류로 나누었으며, 각 기술에서 의미 있는 단어를 추출하기 위해 특허 요약과 특허 제 1청구항에서 명사만을 추출하였다. 이어서, 텍스트 데이터를 수치화하기 위해 TF-IDF(Term Frequency-Inverse Document Frequency) 방법을 적용하였다. 이 과정에서 계산의 효율성과 모델 성능 최적화를 위해, TF-IDF 값이 가장 높은 상위 2,000개의 단어를 선정하도록 하였다. 또한, 5개 미만의 문서에서만 등장하는 단어는 TF-IDF 벡터화 과정에서 제외했다. 그리고 기술 특허 분류 모형을 구축하기 위해 5겹 교차 검증을 실행하였다. 분석에 사용된 모형은 의사결정나무, 랜덤포레스트 그리고 그래디언트 부스팅 모형으로 최대 뿌리 깊이는 각각 10, 20 그리고 30으로 설정하였고, 각 노드에서 분리 가능한 최소 잎의 수는 8, 12, 18로 설정하였다. 노드가 분할하기 위한 최소 샘플 수는 2, 5, 10개이며, 마지막으로 결정 나무 개수는 10과 200으로 설정하여 기계학습 모형을 학습시켰고, [표 4]과 같이 각 기계학습에 따른 최적의 매개변수를 확인할 수 있었다.

[표 4] 기계 학습 모형의 최적 하이퍼파라미터

[Table 4] Optimal Hyperparameters of Three Machine Learning Models

하이퍼 파라미터	하이퍼 파라미터 범위	의사결정나무		랜덤 포레스트		그래디언트 부스팅	
		요약	제1청구항	요약	제1청구항	요약	제1청구항
트리의 최대 깊이	[10, 20, 30]	20	10	10	10	30	20
리프 당 최소 샘플 수	[8, 12, 18]	8	8	8	8	8	8
최소 샘플 수	[2, 5, 10]	2	5	5	5	10	2
트리 수	[10, 200]	-	-	200	200	200	200

마지막으로 [표 5]에는 의사결정나무, 랜덤 포레스트, 그래디언트 부스팅 모형의 성능을 정확도, 카파 상관계수, F1-점수를 기준으로 성능을 살펴볼 수 있다. 특허 요약과 특허 제1 청구항 부문에서 모두 그래디언트 부스팅, 랜덤 포레스트 그리고 의사결정나무 모형 순으로 성능을 보여주었다. 이를 통해 단일 결정 나무보다 배깅과 부스팅 기법을 적용한 랜덤포레스트 모형과 그래디언트 부스팅 모형이 우수한 학습 성능을 보임을 확인할 수 있었다. 더불어 특허 요약 분류와 제1 청구항 분류에서 비슷한 성능이 나타난 것을 확인할 수 있었는데, 이는 자연어 전처리 과정에서 CCU 기술을 다섯 가지 주요 분류로 나누고 중요한 키워드로 명사만을 추출한 결과로 판단된다.

[표 5] 분류 모형의 성능 평가

[Table 5] Performance of Classification Models

요약	Accuracy	Kappa	F1-Score
Decision Tree	0.64	0.53	0.64
Random Forest	0.74	0.66	0.72
Gradient Boosting	0.84	0.80	0.85

제1 청구항	Accuracy	Kappa	F1-Score
Decision Tree	0.62	0.50	0.61
Random Forest	0.74	0.65	0.72
Gradient Boosting	0.83	0.78	0.83

4. 결론

최근 빅데이터 시대의 도래로 인공신경망을 포함한 기계학습 모델들이 의학, 유전체 연구, 기업 경영 등 다양한 분야에 광범위한 영향을 미치고 있음에도 불구하고, 기술 특허 분석에 자연어 처리와 기계학습을 적용한 국내 리걸테크(Legal Technology) 연구는 충분히 발전하지 못한 상황이다.

본 연구는 CCU 기술 특허 데이터, 자연어 전처리 기법 그리고 기계학습모형 기반의 분류 시스템을 설계하고, 정확도, 카파 상관계수 그리고 F1-점수를 비교·분석하였다. 주요 결과를 요약·정리

하면 다음과 같다. 첫째, 다섯 가지 CCU 기술 분류에서 그래디언트 부스팅, 랜덤 포레스트, 의사결정나무 순으로 성능이 나타났다. 이를 통해 단일 결정 나무보다 배깅과 부스팅 기법을 적용한 랜덤포레스트 모형과 그래디언트 부스팅 모형이 더 우수한 학습 성능을 제공함을 확인할 수 있었다. 둘째, 특허의 요약과 제1 청구항을 활용한 기술 분류에서 비슷한 성능이 관찰되었다. 이는 자연어 처리 과정에서 중요한 키워드를 명사로만 추출한 것이 주요 요인으로 보인다.

본 연구는 자연어 전처리와 기계학습 모형을 CCU 기술 특허 분류에 처음으로 적용한 의미 있는 연구이다. 이는 사무 로봇 기술(Robotic Process Automation)를 통해 반복적인 업무를 자동화하는 데 응용될 수 있는 가능성을 제시한다. 그러나 본 연구가 중요한 성과와 의의를 보여줌에도 불구하고, 여전히 향후 후속 연구를 통해 보완해야 할 부분이 몇 가지 존재한다. 본 연구 수행에 활용한 특허 데이터는 요약과 제 1청구항으로 구분되는데 특허 분석의 질적 성장을 모색하기 위한 측면에서 선행기술, 도면 그리고 IPC(International Patent Classification) 등이 포함된 데이터에 근거한 분석 및 예측이 필요하다. 더불어 자연어 전처리 과정에서 명사 추출에만 국한된 점을 고려할 때, 형태소 분석, 품사 태깅, 어구 추출 등 다양한 기법이 포함된 연구가 필요하다.

References

[1] U. J. Shin, “The hottest year in history...1.32℃ ↑ than before industrialization”, www.ytn.co.kr, https://www.ytn.co.kr/_ln/0104_202311101529244270.

[2] J. M. Kim, S. Y. Kim, J. H. Bae, Y. J. Shinn, E. Y. Ahn, J. W. Lee, “Analysis of Patent Trends on the CCUS Technologies”, Economic and Environmental Geology, vol. 53, no. 4, August 2020, pp. 491-504, doi: 10.9719/EEG.2020.53.4.491.

[3] Ministry of Foreign Affairs, “Paris Agreement Significance and Features”, www.mofa.go.kr, https://www.mofa.go.kr/www/wpge/m_20150/contents.do.

[4] Y. M. Kim, “Greenhouse Gas Reduction Goals and Carbon Neutrality Policy Measures in the Building Sector by 2050”, The Planning and Policy, Korea Research Institute for Human Settlements, vol. 479, September 2021, pp. 12-19.

[5] International Energy Agency, “Technology Perspectives 2020”, International Energy Agency, Paris, France, Septemer 2020. [Online]. Available: https://www.iea.org/reports/ccus-in-clean-energy-transitions/ccus-in-the-transition-to-net-zero-emissions.

[6] H. D. Lim, “CCUS that has high expectations on carbon neutrality to public-private, but has long way to 'commercialization'”, www.greenpostkorea.co.kr, https://www.greenpostkorea.co.kr/news/articleView.html?idxno=207285.

[7] J. H. Suh, D. M. Lee, “Technology Forecasting Analysis of Mineral Carbonation from High CO2 Emission Industries”, The Journal of Next-generation Convergence Technology Association, vol. 7, no. 1, January 2023, pp. 120-131, doi: 10.33097/JNCTA.2023.07.01.120.

[8] S. W. Kang, H. C. Kang, “Who Gets Government SME R&D Subsidy? Application of Gradient Boosting Model”, The Journal of Society for e-Business Studies, vol. 25, no. 4, November 2020, pp. 77-109, doi: 10.7838/jsebs.2020.25.4.077.

[9] S. Y. Moon, “Performance comparison of classification methods based on the random forest in class imbalanced data”, Master's Thesis, Department of Biostatistics, Graduate School of Korea University, Republic of Korea, 2018. [Online] Available: https://www.riss.kr/search/detail/DetailView.do?p_mat_type=be54d9b8bc7cdb09&control_no=751424b80ded7d13ffe0bdc3ef48d419&keyword=Performance%20comparison%20of%20classification%20methods%20based%20on%20the%20random%20forest%20in%20class%20imbalanced%20data.

[10] J. Y. Kim, “Information Visualization Methods for Personalized Product Reviews”, Master's thesis, Department of Industrial · Information System Engineering, Graduate School of Soongsil University, Republic of Korea, 2017. [Online] Available: https://www.riss.kr/search/detail/DetailView.do?p_mat_type=be54d9b8bc7cdb09&control_no=affc201378187dd0ffe0bdc3ef48d419&keyword=Information%20Visualization%20Methods%20for%20Personalized%20Product%20Reviews.

[11] J. R. Landis, G. G. Koch, “An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers”, Biometrics, vol. 33, no. 2, June 1977, pp. 363-374, doi: 10.2307/2529786.