http://dx.doi.org/10.14386/SIME.2023.31.4.60

머신러닝 기반 인공지능 특허 품질 예측

김성현* · 옥창훈** · 김영민***

<목 차>

국문초록 : 인공지능은 4차 산업혁명의 프레임이 소개된 이후 점차 보편적인 기술로 자리를 잡아가고 있으며, 인공지능 관련 특허 출원도 크게 증가하고 있다. 최근에는 특허 생태계가 출원 건수 위주의 양적 경쟁에서 고품질의 특허 확보라는 질적 경쟁으로 패러다임이 변화되면서, 저품질 특허로 인한 비용 손실에 관심이 높아지고 있다. 이러한 배경으로 본 연구에서는 머신러닝과 Doc2Vec 알고리즘을 활용하여 특허 품질을 예측하는 방법을 제안하고자 한다.

본 연구를 위해 WIPO에서 정의한 CPC 코드를 활용하여 미국 특허청(USPTO)에 등록된 인공지능 관련 특허 데이터를 추출하였고, 이를 통해 정형 데이터 기반 19개 변수, 비정형 데이터 기반 7개 변수를 개발하였다. 특히, 새롭게 제안하는 Doc2Vec 알고리즘을 이용한 제목과 초록 텍스트 유사도 변수는 고품질 특허를 예측하는데 영향을 미칠 것으로 판단된다. 이에 유사도 변수의 효과를 확인하기 위해 유사도 변수를 포함한 앙상블 기반 머신러닝 모델과 포함하지 않은 모델을 개발하여 비교하였다. 실험 결과, 유사도 변수를 포함한 모델이

* 한양대학교 기술경영전문대학원 박사과정 (shyunkim@hanyang.ac.kr)
** 한양대학교 기술경영전문대학원 박사과정 (changhunok@hanyang.ac.kr)
*** 한양대학교 기술경영전문대학원 교수 (yngmnkim@hanyang.ac.kr)

AUC 0.013, f1-score 0.025가 높게 나타나 더 우수한 성능을 보였다. 이는 유사도 변수가 고품질 특허 예측에 기여한다는 것을 시사한다. 또한, SHAP을 이용하여 블랙박스 형태의 머신러닝 변수 영향도를 설명하였다.

본 연구를 통해 핵심 기술 분야인 인공지능과 같은 영역에서 특허의 품질을 예측하고, 고품질 특허 개발을 장려함으로써 사회적 가치를 실현하는 데 기여할 수 있을 것으로 기대한다.

주제어 : 특허 품질 예측, 머신러닝, 인공지능, Doc2Vec, 유사도

Machine Learning Based Artificial Intelligence Patent Quality Prediction

Sunghyun Kim · Changhun Ok · Youngmin Kim

Abstract : Artificial intelligence has gradually become a ubiquitous technology since the introduction of the framework of the Fourth Industrial Revolution, and the number of patent applications related to artificial intelligence has also significantly increased. Recently, the paradigm of the patent ecosystem has shifted from a quantitative competition based on the number of applications to a qualitative competition focused on securing high-quality patents, due to the growing concern about the costs incurred by low-quality patents. Against this background, this study proposes a method for predicting patent quality using machine learning and the Doc2Vec algorithm.

For this research, we utilized CPC codes defined by WIPO to extract patent data related to artificial intelligence from the United States Patent and Trademark Office (USPTO). Through this process, we developed 19 variables based on structured data and 7 variables based on unstructured data. Particularly, we introduced a novel approach using the Doc2Vec algorithm to calculate similarity variables for the title and abstract texts, which are expected to influence the prediction of high-quality patents. To assess the impact of these similarity variables, we developed and compared an ensemble-based machine learning model that includes the similarity variables with a model that does not. The experimental results showed that the model incorporating the similarity variables exhibited superior performance with an AUC of 0.013 and an f1-score of 0.025, indicating their contribution to predicting high-quality patents. Additionally, we explained the variable importance of the black-box machine learning model using SHAP.

Through this study, we expect to contribute to the realization of social value by predicting the quality of patents and promoting the development of high-quality patents in the field of key technologies such as artificial intelligence.

Key Words : Patent quality prediction, Machine learning, AI, Doc2Vec, Similarity

Ⅰ. 연구배경 및 목적

4차산업혁명(Industry 4.0)의 프레임이 소개된 이후 인공지능(AI)은 보편적인 기술로 자리를 잡아가고 있다. 2022년 세계지적재산기구(WIPO)에서 발표한 보고서에 따르면 지난 5년 동안 인공지능 관련 특허는 미국, 중국, 일본, 한국, 유럽연합에서의 출원이 주를 이루었으며, 국가적으로 미국, 중국이 특허 출원을 주도하고 있는 것으로 나타났다. 이중 한국의 인공지능 분야 출원 증가가 가장 크다고 분석하였고, 이는 한국에서 인공지능 기술개발에 대한 관심과 투자가 증가하고 있음을 나타낸다.

그러나 한국에서는 특허 출원 이후 활용은 주변국들에 비해 낮은 것으로 나타났다. IBM에서 발표한 '2022년 AI 도입지수' 연례보고서에 따르면 한국은 2.81점으로 세계 40개국 평균 3점보다 낮은 것으로 평가되었다. AI 도입지수는 인공지능 기술을 사용하는 조직, 기술투자, 기술 재활용 등의 항목으로 국가별 수준을 평가한 것으로 한국의 경우 인공지능 기술을 사용하는 조직 비율, 투자 비율은 전세계 평균보다 높았으나, 특허 기술의 재활용 측면에서는 상대적으로 낮게 평가되었다. 실제로 2022년 특허청의 국내 특허 활용과 관련된 조사에서 국내 보유 기술 특허 활용 추이는 2014년부터 점차 증가하고 있지만 아직도 27.9% 수준에 그치고 있다는 결과가 보고되었다.

최근 특허 생태계가 출원 건수 위주의 양적 경쟁에서 고품질의 특허 확보라는 질적 경쟁으로 패러다임이 변화되면서, 저품질 특허로 인한 비용 손실에 관심이 높아지고 있다. 이에 특허 주요국들은 품질향상을 목표로 정책을 조정하거나, 품질개선을 위한 연구에 많은 투자를 진행하고 있다(김동준, 2010). 한국 또한 정책적인 개선을 위해 국내 산·학·연이 2020년 '지식재산 서비스 혁신위원회'를 조직하여 특허 품질을 평가하기 위한 프로세스를 정립하였으며, 특허청에서도 특허심판부서 조직을 개편하며 특허의 질적 향상을 위해 노력하고 있다.

또한, 특허 데이터를 이용한 품질 예측 모델 개발에 대한 연구도 활발하게 진행되고 있다. 특허 데이터는 인용, 출원인, 분류 등과 같이 구조적으로 관리되는 데이터와 초록, 청구항 등과 같이 비구조적으로 관리되는 영역으로 나눌 수 있다(Kim과 Lee, 2015). 특허 품질 예측에는 주로 구조적으로 관리되는 데이터가 많이 사용된다.

인용 데이터는 특허 품질 분석에 가장 많이 사용되며, 해당 특허가 인용하고 있는 특허를 분석하는 '후방인용분석', 해당 특허를 인용하고 있는 특허를 분석하는 '전방인용분석'으로 분석방법을 구분할 수 있다. 보통 특허 품질 예측 분석에는 전방인용분석 방법이 많이 쓰이나,

시간과 비용이 많이 들기 때문에 후방인용분석을 이용하여 특허 품질 예측을 진행하기도 한다(Kyebambe 외, 2017). 또한 출원인의 선행기술 정보를 바탕으로 품질을 예측하거나(김동준, 2010), 특허 분류를 통한 품질 예측에 대한 연구도 진행되었다(김성호와 김지표, 2019).

대부분의 연구는 구조적인 데이터를 이용하여 특허 품질 분석을 수행하고 있다. 그러나 초록과 같은 비정형 데이터도 특허 기술의 많은 내용을 담고 있기 때문에 품질 예측을 위한 입력으로 사용하면 품질 예측의 정확도를 높일 수 있다(Lee 외, 2018; Niemann 외, 2017). 최근에는 비정형 데이터의 처리기술 발달로 인해 특허 초록과 청구항 정보에서 의미를 추출하는 연구가 많이 진행되고 있으며, 머신러닝을 활용하여 데이터를 분류하고 cosine 유사도를 이용해 품질 예측의 척도로 사용하는 연구도 진행되고 있다(Chung과 Shon, 2020).

본 연구에서는 기존에 사용되었던 정형화된 특허 품질 관련 데이터와 비정형화 되어 있는 초록 데이터를 이용하여 특허 품질 예측 모델을 개발하였다. 특히, Doc2Vec 알고리즘을 활용하여 초록 데이터 기반 새로운 유사도 변수를 제시하였고, 배깅(Bagging) 및 부스팅(Boosting) 계열의 모델에 적용하여 성능을 개선하였다. 2장에서는 본 연구의 배경이 되는 기존 연구를 살펴보고 3장에서는 데이터와 변수를 제시한다. 그 다음 4장에서는 제시한 연구 방법의 모델 결과를 살펴보고 관련 시사점을 도출한다. 마지막으로 5장에서는 본 연구의 결론을 서술한다. 본 연구를 통해 제안된 특허 품질 예측 모델을 기반으로 사회적 비용 감소 효과 및 기술 경쟁력 평가의 지표로 활용할 수 있기를 기대한다.

Ⅱ. 관련 연구

특허의 품질이 중요해짐에 따라 인공지능 분야에서도 품질 연구가 활발히 이루어지고 있다. 특히, 특허 품질 예측은 머신러닝을 많이 활용하고 있으며 최근에는 부스팅 계열의 알고리즘을 적용하여 성능을 개선하기 위한 노력도 이루어지고 있다. 또한, 특허 텍스트를 이용한 인용 추천, 기술 식별, 유사도 등 다양한 연구가 진행되고 있다. 이번 장에서는 인공지능 분야에서의 특허 연구, 머신러닝을 기반으로 한 특허 품질 예측 및 특허 텍스트 기반 분석과 관련된 이전 연구들을 살펴보고자 한다.

1. 인공지능 분야 특허 품질 연구

광범위한 인공지능 분야 특허 기술 분석을 위해서는 인공지능 기술을 정의하는 것이 선행되어야 한다. 한국 특허청의 경우 인공지능 기술을 학습과 추론, 언어처리, 시각처리, 상황인식 4개의 대분류로 분류하여 정의하고 있고, 정보통신기술진흥센터(IITP)는 성장하는 AI기술(지도학습, 추론학습, 모델 경량화 등), 사회 친화적 AI기술(인지능력, 이해판단 능력 등)로 분류하여 정의하고 있다. 이처럼 인공지능의 기술 분야는 기관마다 다른 기준으로 관리되고 있다. 또한, 세계지적재산기구(WIPO)에서는 인공지능 관련 분류 기준인 CPC, IPC 코드를 제시하여 관련 특허 데이터 추출에 활용할 수 있게 하고 있다.

특허 데이터를 이용한 인공지능 기술 동향, 기술 예측 등의 연구에서는 기술 분야의 정의가 너무 광범위하여 특허 분류 코드나 키워드 검색으로 특허 데이터를 추출하고 있다. 이에 따라 국가별 특허 동향 추이, 네트워크 분석을 통한 기술 확산 패턴 등 기술 동향에 집중하여 특허 분석 연구가 진행되고 있다(이현상 외, 2022). 인공지능 기술에 대한 특허 분석 연구는 대부분 국가별 기술 경쟁력 비교나 기술 동향 분석에 중점을 두고 있다. 이는 인공지능 기술이 특정 도메인 기술과 결합하여 특허로 출원되고 있기 때문으로 분석된다.

2. 머신러닝을 통한 특허 품질 예측

기존 연구들은 피인용수를 기준으로 특허의 품질을 예측하는 것이 대부분이었다. 인용 비율이 높을수록 고품질 특허로 정의되며, 피인용수 상위 1%에 속하는 특허를 영향력 있는 획기적 발명(breakthrough inventions)으로 정의하고, 이는 인용이 매우 활발한 특허로 상업성 및 향후 기술 개발과 관련이 있다고 평가한다(Ahuja와 Morris, 2001). 또한, 피인용수 상위 5%에 속하는 특허는 고피인용 특허(Highly Cited Patents)로 분류되며, 후속 특허의 개발에 기초가 되는 중요한 기술적 진보를 담고 있다(Park과 Park, 2007). 한편, 특허나 연구 논문에서 해당 분야와 연도에 접수된 인용의 상위 5%를 'Home Run'으로 분류하여 고품질로 정의하고 있다(Ahmadpoor와 Jones, 2017).

피인용수와 관련된 연구로는 발명자수, 자기인용, 청구항수, 출원국가, 지리적 변수 등을 고려하여 특허 품질과의 관계를 분석하는 연구와(Lee 외, 2006), 문서 정보를 기반으로 바이오 산업에서 인용 예측을 위한 회귀모델 개발 연구(Lin 외, 2007) 등이 있다. 또한, 소송 만기, 재등록 여부, 청구항수, IPC Subclass 등 변수를 사용하여 피인용수를 예측하고

이를 통한 특허 검증 연구도 있다(Nathan과 Kenneth, 2016). 피인용수 외 다른 지표를 함께 사용한 연구로는 IPC 코드수, 특허 패밀리수, 평균 인용 빈도를 복합적으로 고려하여 특허의 가치를 평가하는 연구가 있다(Ernst, 2003). 또한, 이들 지표 간의 가중치를 고려한 품질 예측(Lanjouw와 Schankerman, 2004), 품질 평가 지표에 대한 연구도 진행되었다 (Squicciarini 외, 2013).

머신러닝 기술의 발전으로 특허 품질 예측에도 머신러닝 기술이 활용되고 있다. SOM, KPCA, SVM과 같은 분류 알고리즘을 이용하여 고품질, 중간품질, 저품질로 분류하는 연구와(Wu 외, 2016), 부스팅 계열의 XGBoost 분류 알고리즘을 이용하여 특허 품질을 나타내는 피인용수를 예측하는 연구가 진행되었다(조현진과 이학연, 2018). 기업의 올바른 R&D 투자에 기여하기 위해 MLP 알고리즘을 사용하여 고품질 특허를 예측하는 연구(Erdogan 외, 2022) 등도 진행되었다.

3. 특허 텍스트 분석

특허 텍스트에는 유용한 정보가 담겨있고, 이를 활용한 다양한 연구가 진행되고 있다. CPC 정보를 임베딩하기 위해 Diff2Vec 방법을 적용한 특허 인용 추천 프레임워크에 대한 연구가 진행되었다(Choi 외, 2022). 이 연구에서는 특허 심사관 또는 출원자가 이전 특허를 효과적으로 찾는 데 도움이 된다는 결과를 제시하고 있다. 그리고 기술 개발 초기 단계에서 아이디어를 선별하기 위한 분석적인 프레임워크에 관한 연구도 진행되었다(Hong 외, 2022). 특허 초록 텍스트를 Word2Vec을 사용하여 단어 간 의미적 관계를 파악하고, 특허 내에 함축된 아이디어의 기술적 내용을 나타내는 행렬을 구축하여 모델에 적용하였다. 또한, 미국의 모든 특허의 초록 텍스트를 기반으로 인공지능 기술을 식별하는 연구가 진행되었다(Milan 외, 2023). 이 연구에서는 텍스트 데이터를 벡터로 표현하기 위해 “bag-of-words” 방식과 임베딩 기반 방식을 사용하였다.

특허의 유사도 분석에도 텍스트를 활용한 연구가 진행되었다. 특허 텍스트를 빈도수 기반으로 TF-IDF에 적용하여 유사성을 찾아내고 관계를 비교하는 방법이 연구되었고(고광수 외, 2011), 다층 신경망을 이용한 Word2Vec 알고리즘을 적용하여 유사 특허 추천 연구가 진행되었다(이앞길 외, 2020). 또한, Doc2Vec 알고리즘을 이용하여 특허간의 유사도를 찾아 특허 문서를 자동으로 분류하는 연구와(송진주와 강승식, 2019), 초록 데이터의 cosine 유사도를 통해서 특허 간 유사성을 측정하는 연구도 진행되었다(Feng, 2020).

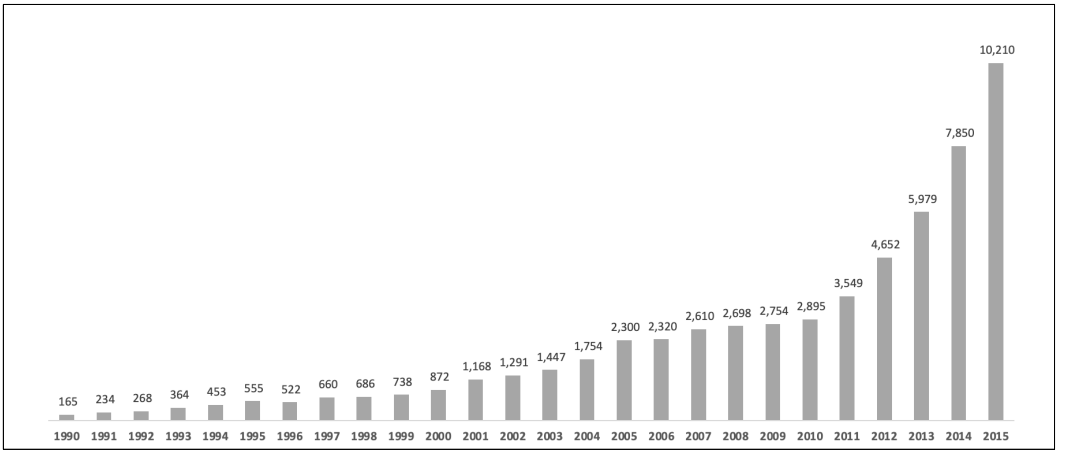
Ⅲ. 데이터 및 변수

1. 데이터

특허 품질 예측을 위해 본 연구에서는 PATSTAT 데이터베이스를 이용하여 미국 특허청(USPTO)에 등록된 특허 정보를 사용하였다. 미국 특허청(USPTO) 등록 특허는 국제 특허로서의 타당성을 인정받으며 지식재산 및 특허에 관련한 연구에서 그 대표성을 인정받고 있다(백서인 외, 2020). 분석 대상은 WIPO에서 정의한 인공지능 관련 CPC 코드를 이용하여 추출하였으며, 특허 출원 기간은 인공지능의 기초가 되는 기술의 등장시기와 특허 품질 판단을 고려하여 1990년부터 2015년까지로 설정하였다. 또한 텍스트 분석을 위해 제목과 초록의 언어가 영어인 특허만 사용하였으며 전체 분석 대상수는 총 58,994개의 특허이다. 주요 CPC 코드 분류별 출원건수 현황은 <표 1>과 같고, 연도별 특허 출원수를 보면 점진적으로 증가하는 추세를 보이다 2011년부터 크게 증가하는 것을 알 수 있다.

<표 1> 주요 AI 관련 CPC 코드 분류별 출원건수

CPC 코드	설 명	출원건수
G06K9/00	패턴 인식 방법 또는 배열	3,097
G06N7/005	확률 네트워크	2,490
A61B5/7264	생리적 신호 또는 데이터의 분류, 예. 신경망	2,335
G10L17/00	화자 인식 또는 검증	1,016
A61B5/7267	분류 장치의 훈련을 수반하는 것	1,005
G05D1/0088	자율적인 의사 결정 프로세스로 특징, 예. 인공지능	992
G06F17/16	행렬 또는 벡터 연산	989
G06T2207/20081	트레이닝; 학습	952
G10L15/00	음성(speech) 인식	707
H04L41/16	기계 학습 또는 인공 지능을 사용	582



<그림 1> 연도별 특허 출원건수

2. 변수

<표 2> 피인용수 상위 5% CPC 코드 분류별 출원건수

CPC 코드	설 명	출원건수
G05D1/0088	자율적인 의사 결정 프로세스로 특징, 예. 인공지능	179
A61B5/7264	생리적 신호 또는 데이터의 분류, 예. 신경망	173
G06K9/00	패턴 인식 방법 또는 배열	134
G06N7/005	확률 네트워크	95
A61B5/7267	분류 장치의 훈련을 수반하는 것	84
G10L15/00	음성(speech) 인식	57
G10L17/00	화자 인식 또는 검증	50
H04L41/16	기계 학습 또는 인공 지능을 사용	40
G06T2207/20081	트레이닝; 학습	25
G06F17/16	행렬 또는 벡터 연산	6

피인용수를 이용한 특허 품질 예측은 성능 및 활용성을 고려하여 분류 문제로 변환하여 수행하기로 한다. 이진 분류 형태로 변환하기 위해 출원 후 5년이내 피인용수 상위 5%를 target, 즉 '고품질'로 정의하였다. 피인용수 상위 5%의 CPC 코드 분류별 출원 현황은 <표 2>와 같고, 전체 분포와는 다르게 "인공지능", "신경망" 등이 가장 많은 것을 알 수 있다. 비선형 모델을 사용하였으며, 특허의 고품질 여부를 잘 분류할 수 있도록 다양한

입력변수를 고려하였다. PATSTAT 데이터베이스를 활용하여 두 가지 타입의 다양한 변수를 개발하였다. 첫 번째로는 청구항수, 개발자수 등 정형 데이터를 이용하여 19개의 변수를 개발하였고, 두 번째로는 본 연구에서 새롭게 제안하는 방식으로 초록과 제목의 텍스트 비정형 데이터를 이용하여 7개의 유사도 변수를 개발하였다. 청구항, 발명의 설명 등의 텍스트 정보는 PATSTAT 데이터베이스에 존재하지 않아 본 연구에서는 제외되었으나, 초록과 제목 텍스트만을 이용하여 우수한 성능을 나타내는 특허 분류 연구 사례가 존재한다(Li, 2018).

정형 데이터를 이용한 변수는 기존 연구에서 중요하다고 밝혀진 변수들로 구성되어 있다. 이 변수들은 인용 관련 6개, 패밀리수 관련 2개, IPC 코드 관련 5개, 청구항수 관련 2개, 발명가수 관련 2개, 그리고 특허의 독창성 관련 2개로 구분될 수 있다. 특히, 인용 특허의 총 피인용수(BC_FC_TOT)와 인용 특허의 평균 피인용수(BC_FC_AVG)는 이전 연구에서도 높은 중요도를 갖고 있는 변수로 알려져 있다. 또한, 독창성 관련 변수는 특허가 이전 기술과 상당히 다른지 여부를 판단할 수 있는 변수로 고품질 특허에 영향을 줄 수 있을 것으로 판단된다.

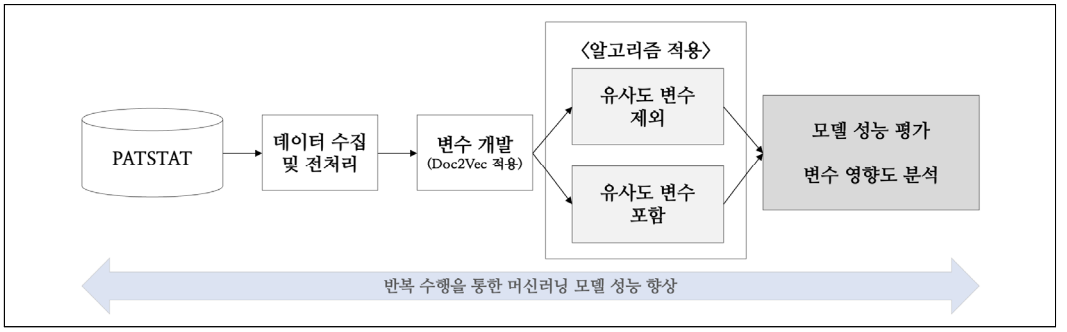
본 연구에서 새롭게 제안한 비정형 데이터 기반 변수는 제목과 초록의 유사도 변수 1개와 고품질 특허 그룹과의 초록 유사도 변수 6개로 구성되어 있다. 이러한 변수는 특허의 핵심 정보를 담고 있는 제목과 초록을 활용하여, 고품질 특허와 유사한 특허는 고품질로 분류될 가능성이 높다는 가정이다. 이를 통해 특허 품질 예측에 새로운 차원을 추가하고, 고품질 특허를 식별하는데 도움을 줄 수 있을 것이다.

특허 인용 시점에 따라 값이 변동될 수 있는 입력변수는 특허 출원년도를 기준으로 산정하였다. 예를 들면, 분석 대상 특허가 인용한 특허의 총 피인용수(BC_FC_TOT) 같은 경우는 분석 대상 특허 출원 이후에도 꾸준히 증가할 수 있기 때문에 분석 시점에 따라 달라질 수 있다. 따라서 분석 대상 특허의 출원년도까지의 총 피인용수를 계산하는 것이 타당하다. 최종적으로 <표 3>과 같이 26개의 입력변수, 1개의 target 변수를 사용하였다.

<표 3> 입력/Target 변수

구 분	변 수	설 명
특허 정형 데이터	BC	인용 특허수
	BC_BC_AVG	인용 특허의 평균 인용수
	BC_BC_TOT	인용 특허의 총 인용수
	BC_FC_AVG	인용 특허의 평균 피인용수
	BC_FC_TOT	인용 특허의 총 피인용수
	NPL	NPL(Non-Patent Literature) 인용수
	CNT_AID	INPADOC 패밀리수
	DOCDB_FAMILY_SIZE	DOCDB 패밀리수
	NB_IPC	IPC 코드수
	IPC4_CNT	4자리 기준 IPC 코드수
	IPC8_CNT	8자리 기준 IPC 코드수
	BC_IPC4_CNT_AVG	인용 특허의 4자리 기준 평균 IPC 코드수
	BC_IPC8_CNT_AVG	인용 특허의 8자리 기준 평균 IPC 코드수
	NB_CLAIMS	청구항수
	BC_NB_CLAIMS_AVG	인용 특허의 평균 청구항수
	NB_INVENTORS	발명가수
	BC_INVENTORS_AVG	인용 특허의 평균 발명가수
	ORIGINALITY	특허의 독창성(Jaffe's Originality) $ORIGINAL_i = 1 - \sum_{k=1}^{N_i} \left(\frac{NCITED_{ik}}{NCITED_i} \right)^2$
	BC_ORIGINALITY_AVG	인용 특허의 평균 독창성
특허 비정형 데이터	TITLE_ABSTRACT_SIM	제목과 초록의 유사도
	SIM_MIN	고품질 특허 그룹과의 초록 유사도 최소값
	SIM_25P	초록 유사도 25% 값
	SIM_50P	초록 유사도 50% 값 (중위수)
	SIM_75P	초록 유사도 75% 값
	SIM_AVG	초록 유사도 평균값
	SIM_MAX	초록 유사도 최대값
Target	TARGET	피인용수 상위 5% 이상이면 1, 아니면 0

Ⅳ. 모 델

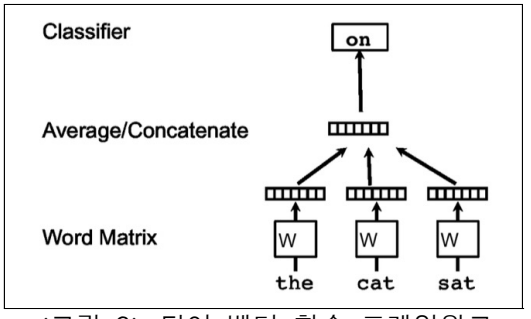


<그림 2> 모델 개발 프로세스

<그림 2>와 같이 특허 텍스트 데이터를 모델의 입력변수로 활용하기 위해 Doc2Vec 알고리즘을 이용하여 유사도 변수를 개발하였고, 최근 이진 분류 문제에 우수한 성능을 보이고 있는 배깅(Bagging) 계열의 Random Forest, 부스팅(Boosting) 계열의 LightGBM, XGBoost 알고리즘 등 앙상블(Ensemble) 방법을 사용하여 모델을 개발하였다. 또한, Doc2Vec 기반 유사도 변수의 효과를 검증하기 위해 해당 변수 포함 여부에 따른 동일한 하이퍼 파라미터를 사용한 모델 결과를 비교하였다. 특히, 2014년 출시된 XGBoost와 2016년 출시된 LightGBM은 다양한 머신러닝 경진대회에서 우승하면서 많이 활용되기 시작했다. 본 연구에서는 모델 개발을 위해 80%의 학습 데이터와 20%의 테스트 데이터로 분할하였으며, 학습 데이터 중 20%는 검증용 데이터로 사용하였다.

1. Doc2Vec 기반 유사도

본 연구에서는 특허 제목과 초록 비정형 데이터를 이용하여 새로운 변수를 제안하고, 이를 통해 모델의 성능 개선에 기여하고자 한다. 분석 대상 특허의 제목과 초록 텍스트를 Doc2Vec 알고리즘을 이용하여 벡터로 표현하고, cosine 유사도를 계산하여 다양한 입력변수로 개발한다. 이러한 유사도 변수는 고품질 특허를 예측하는데 중요한 요소 중 하나로 작용될 수 있다.

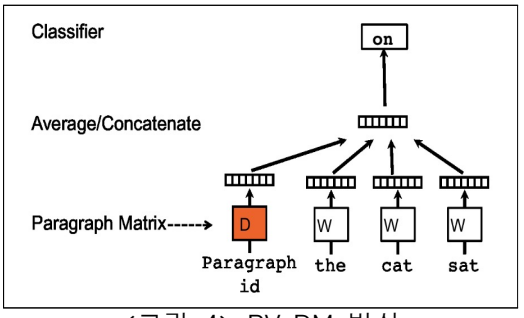


<그림 3> 단어 벡터 학습 프레임워크

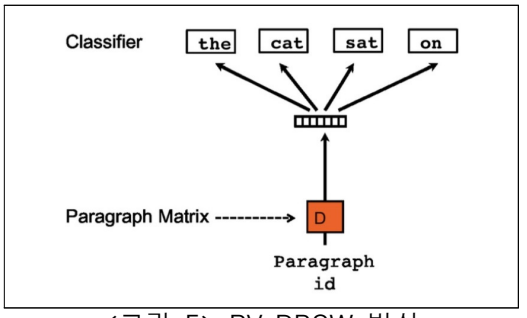
$$\frac{1}{T}\sum_{t=k}^{T-k}\log p(w_t \mid w_{t-k},\cdots,w_{t+k}) \quad <식\ 1>$$

$$y = b + Uh(w_{t-k},\cdots,w_{t+k};W) \quad <식\ 2>$$

$$P(w_t \mid w_{t-k},\cdots,w_{t+k}) = \frac{e^{y_{w_i}}}{\sum_i e^{y_i}} \quad <식\ 3>$$



<그림 4> PV-DM 방식



<그림 5> PV-DBOW 방식

Doc2Vec 알고리즘은 유사한 의미를 가진 단어들을 벡터 공간에 서로 가까이 위치시키는 Word2Vec 알고리즘을 기반으로 되어 있다. <그림 3>은 단어 벡터를 학습하기 위한 프레임워크로 “the”, “cat”, “sat” 단어를 이용하여 네 번째 “on” 단어를 예측한다. 단어 벡터 모델의 목적 함수는 <식 1>과 같이 평균 로그 확률을 최대화하는 것이고, <식 2>와 같이 softmax 파라미터 U, b를 통해 정규화되지 않은 로그 확률 yi가 계산된다. 그리고 <식 3>의 일반적인 softmax와 같은 다중 클래스 분류기를 사용하여 정규화된 확률값을 얻을 수 있다. 이러한 단어 벡터 학습 방법을 기반으로, Doc2Vec 알고리즘은 유사한 텍스트를 사용하는 문서를 벡터 공간에 가까이 배치시킬 수 있다. Doc2Vec 알고리즘은 <그림 4>와 같은 문맥적인 의미를 보존하기 위해 문장의 단어 순서 정보를 활용하는 PV-DM (Distributed Memory) 방식과 <그림 5>와 같은 문맥 정보를 배제하고 문서 전체의 의미를 담은 벡터를 학습하는 PV-DBOW(Distributed Bag Of Words) 방식이 있다. 두 가지 방식을 결합하여 사용하면 더 좋은 성능을 보인다는 결과도 있지만(Le, 2014), 본 연구에서는 각각의 방식을 개별적으로 모델에 적용하여 성능을 비교한 후, 최종적으로 PV-DBOW 방식이 적용되었다.

<표 4> Doc2Vec 기반 유사도 변수 개발 프로세스

단 계	설 명
Step1 : Target 정의	Target=1 : 고품질 특허의 초록 텍스트
Step2 : Doc2Vec 모델 개발 및 벡터 추정	1. 특허 제목, 초록 텍스트 학습을 통한 모델 개발 - vector_size=10, windows=6 (단, 리소스 문제로 인해 샘플 제목, 초록 5,000개 이용) 2. 개발된 모델을 활용하여 문서 벡터 추정

Step3 : 유사도 계산	1. i(전체 데이터)와 j(Target=1 그룹)의 cosine 유사도 계산
	2. 전체 데이터 i에 대해 분위수, 최소값, 최대값, 평균값
	* PTV : 특허 제목 벡터 / PAV : 특허 초록 벡터
	
	sim_result = []
	for i in range(len(전체 데이터)):
	cos_sim_result = []
	t_a_sim = cosine(PTVi, PAVi)
	for j in range(len(Target=1 그룹)):
	if i != j:
	cos_sim_result.append(cosine(PAVi, PAVj))
	sim_result.append(t_a_sim, (min, percentile, avg,
	max of cos_sim_result))

Doc2Vec 기반의 유사도 산출 프로세스는 <표 4>와 같이 진행되었다. 제목과 초록 텍스트를 이용하여 각각 Doc2Vec 모델을 개발하고, 이를 통해 각 특허 초록 벡터를 추정하였다. 이후 전체 특허를 고품질 특허 그룹과 1:1 cosine 유사도를 계산하여 유사도 최대값, 유사도 평균값 등의 새로운 변수를 생성하였다. 고품질 특허 또한 전체 특허에 속하기 때문에 동일한 특허간의 유사도 계산은 제외하였다. 만약 고품질 특허간의 유사도가 높게 나타난다면 이는 특허 초록의 유사도가 특허 품질 예측에 유의미한 변수가 될 수 있음을 시사한다.

2. 성능 평가

<표 5> Doc2Vec 기반 유사도 변수 제외 모델 성능

모델	알고리즘	Accuracy	AUC	Precision	Recall	F1-Score	MCC
①	Random Forest	0.738	0.711	0.129	0.682	0.217	0.211
②	LightGBM	0.750	0.801	0.137	0.696	0.228	0.227
③	XGBoost	0.831	0.788	0.168	0.554	0.258	0.236

모든 모델 성능 평가 결과는 테스트 데이터의 결과이다. <표 5>는 Doc2Vec 기반 유사도 변수를 제외한 모델의 성능 결과를 보여준다. Accuracy는 0.74~0.83 정도가 나왔지만 클래스가 불균형한 경우에는 accuracy로 해석하는 것은 타당하지 않다. 예를 들어, 테스트 데이터의 실제 레이블이 1:99로 고품질의 비율이 아주 낮다면, 모델이 모두 비고품질로 예측해도 accuracy는 0.99일 것이다. 본 연구에서는 고품질 특허의 비율이 약 5.3%로 클래스가 불균형하기 때문에, 모델의 성능을 올바르게 판단하기 위해서는 AUC(Area Under Curve)와 <그림 6>과 같은 다른 성능 지표를 활용해야 한다. AUC는 ROC curve의 밑면적으로 1에 가까울수록 ROC 그래프가 좌상단에 근접하게 되어 우수한 성능을 갖는 모델이라고 할 수 있다. 모델에서 '고품질'로 예측한 결과 중 실제 '고품질'의 비율이 precision이고, 전체 '고품질' 중 모델에서 제대로 예측한 '고품질'의 비율이 recall이다. Precision은 예측 정확도, recall은 커버리지 개념으로 설명할 수 있으며 이 둘의 조화평균이 f1-score이다. 또한, 클래스가 불균형한 이진 분류 문제에 적합하다고 알려진 성능 지표인 MCC(Matthews Correlation Coefficient)도 추가적으로 확인하였다. MCC는 모델이 얼마나 '고품질'과 'not 고품질'을 잘 분류하고 있는지 측정하기 위해 사용된다. 이 지표는 -1에서 1 사이의 범위를 가지며, 1은 완벽한 예측, 0은 무작위 예측, -1은 완벽하게 반대로 예측한 경우이다. <표 5>의 결과를 보면 클래스 불균형으로 인해 모든 알고리즘에서 recall 대비 precision의 값이 대체적으로 낮게 나타났다. 이에 따라 f1-score 기준으로 살펴보면, XGBoost 알고리즘을 사용한 모델의 값이 0.26으로 가장 높은 성능을 보이고 있다.

Confusion Matrix		Actual	
		Negative	Positive (고품질)
Predicted	Negative	True Negative (TN)	False Negative (FN)
	Positive (고품질)	False Positive (FP)	True Positive (TP)

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

$$\text{Precision} = \frac{TP}{FP+TP}$$

$$\text{Recall} = \frac{TP}{FN+TP}$$

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

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

<그림 6> Confusion Matrix 및 성능 지표

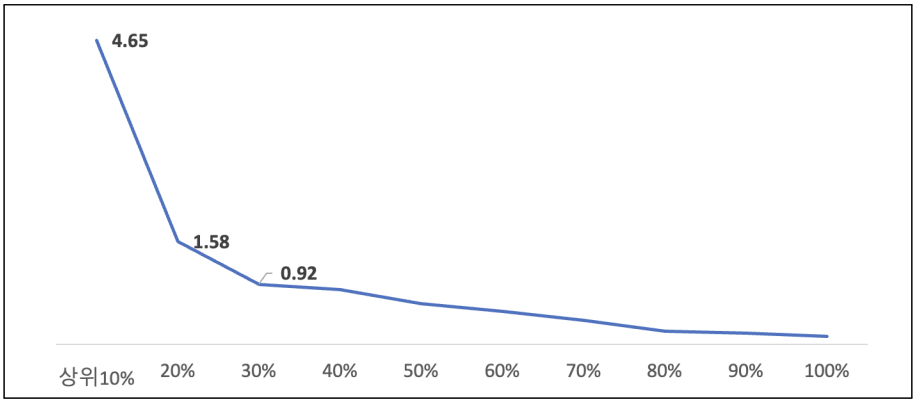
<표 6>은 Doc2Vec 기반 유사도 변수를 포함한 모델의 성능 결과를 보여준다. Doc2Vec 기반 유사도 변수를 포함한 경우에도 XGBoost 알고리즘을 사용한 모델의 f1-score가 0.28로 가장 높은 성능을 보이고 있다. 유사도 변수를 제외했을 때 보다 AUC는 0.013, f1-score는

0.025, MCC는 0.026이 각각 높게 나타났다. 따라서 Doc2Vec 알고리즘을 이용한 특허 초록 텍스트 정보는 모델 성능 개선에 기여함을 알 수 있다.

<표 6> Doc2Vec 기반 유사도 변수 포함 모델 성능

모델	알고리즘	Accuracy	AUC	Precision	Recall	F1-Score	MCC
④	Random Forest	0.748	0.732	0.138	0.715	0.232	0.233
⑤	LightGBM	0.751	0.814	0.140	0.717	0.235	0.237
⑥	XGBoost	0.849	0.801	0.189	0.559	0.283	0.262

모델의 예측 결과는 확률의 값으로 나오며 이는 고품질 특허가 될 가능성 점수로 활용될 수 있다. 따라서 분류 문제이지만 단순히 yes or no의 개념이 아닌 적정 수준의 threshold를 설정하여 활용성을 증대시킬 수 있다. <그림 7>의 lift curve를 보면 모델 스코어 상위 10%는 4.65, 20%는 1.58로 높게 나타나는 것을 알 수 있다. Lift는 랜덤(baseline rate) 대비 효율성을 판단할 수 있는 지표로, 모델 결과 상위 10%의 예측 정확도가 baseline rate 대비 4.65배 높다고 표현할 수 있다.



<그림 7> ⑥번 모델 Lift Curve

3. 변수 중요도

모델의 입력변수들은 각각 서로 다른 영향력을 가지며 중요도가 높을수록 해당 변수가 고품질 특허로 예측하는데 더 큰 영향을 미친다고 볼 수 있다. 유사도 변수 제외 모델 중

성능이 가장 좋은 ③번 모델과 유사도 변수 포함 모델 중 성능이 가장 좋은 ⑥번 모델은 모두 XGBoost 알고리즘을 사용한 모델이다. XGBoost 알고리즘에서 변수 중요도를 판단하는 기준으로 information gain이 사용되었으며, 이는 해당 변수가 모델 예측에 얼마나 영향을 미쳤는지 측정하는 방법으로 노드가 분기할 때 얻는 성능상의 이득이다.

<표 7>의 ③번 모델 변수 중요도를 살펴보면 CNT_AID(INPADOC 패밀리수), BC_ORIGINALITY_AVG(인용 특허의 평균 독창성), BC_NB_INVENTORS_AVG(인용 특허의 평균 발명가수), BC_FC_AVG(인용 특허의 평균 피인용수) 순으로 높게 나타나고 있다. 이를 통해 고품질 특허일수록 인용한 특허의 특성이 더 큰 영향을 미치는 것으로 추측할 수 있다. ⑥번 모델의 변수 중요도를 살펴보면 SIM_MAX(초록 유사도 최대값), CNT_AID(INPADOC 패밀리수), TITLE_ABSTRACT_SIM(제목과 초록의 유사도), BC_FC_AVG(인용 특허의 평균 피인용수) 순으로 높게 나타나고 있다. 이는 특허의 초록 텍스트 유사도 기반 변수들이 대체적으로 높은 중요도를 가지고 있어, 텍스트 데이터가 유용한 정보를 포함하고 있음을 보여준다. 그리고 NB_CLAIMS(청구항수)도 중요하게 나타났는데, 이는 고품질 특허일수록 상대적으로 많은 청구항수가 있는 것으로 추측된다.

모델의 성능 평가 결과, 유사도 변수를 포함한 ⑥번 모델이 포함하지 않은 ③번 모델보다 우수한 성능을 보였다. 이는 유사도 관련 변수들이 ⑥번 모델에서 중요한 역할을 하며, 이로 인해 ③번 모델과 중요 변수들 간의 차이가 있음을 나타낸다. ③번과 ⑥번 모델 공통적으로 BC_FC_AVG(인용 특허의 평균 피인용수), BC_NB_INVENTORS_AVG(인용 특허의 평균 발명가수) 등 인용 특허의 특성이 중요하게 나타났다.

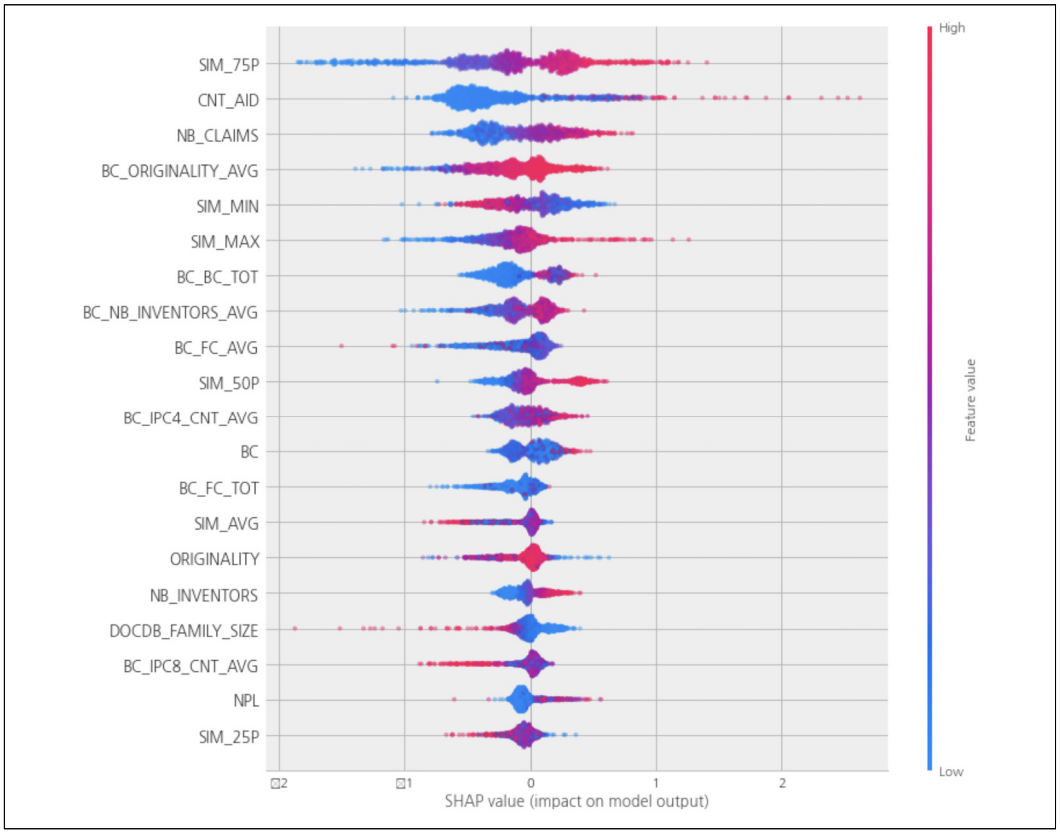
<표 7> 각 모델의 변수 중요도 Top 10

우선순위	③번 모델	⑥번 모델
1	CNT_AID	SIM_MAX
2	BC_ORIGINALITY_AVG	CNT_AID
3	BC_NB_INVENTORS_AVG	TITLE_ABSTRACT_SIM
4	BC_FC_AVG	BC_FC_AVG
5	BC_IPC4_CNT_AVG	BC_NB_INVENTORS_AVG
6	ORIGINALITY	BC_BC_AVG
7	BC_NB_CLAIMS_AVG	SIM_MIN
8	BC_FC_TOT	SIM_75P
9	BC_BC_AVG	NB_CLAIMS
10	BC_IPC8_CNT_AVG	BC_NB_CLAIMS_AVG

일반적으로 앙상블 모델에서 산출되는 변수 중요도는 영향력의 크기만 알 수 있지만 어떻게 영향을 미치는지는 알기가 어렵다. 이에 Lundberg와 Lee(2017)가 제안한 SHAP

을 통해 성능이 가장 좋은 ⑥번 모델의 각 변수가 어떻게 영향을 미치는지 설명한다. 이 방법은 특정 변수에 대한 영향도를 해당 변수 포함 여부로 인해 발생되는 예측값 변화의 평균으로 판단한다. 그래프에서는 변수의 값이 클수록 빨간색, 작을수록 파란색으로 표시되며, 왼쪽으로 갈수록 모델 결과에 음의 영향을 미치고, 오른쪽으로 갈수록 양의 영향을 미친다. 따라서 그래프를 보면 각 변수가 모델 예측에 어떤 영향을 미치는지 파악할 수 있다.

⑥번 모델의 결과는 <그림 8>과 같고, SIM_75P(초록 유사도 75% 값) 변수의 SHAP value 그래프를 보면 값이 클수록 모델 결과에 양의 영향을 미치는 것을 알 수 있다. 또한 SIM_MIN(초록 유사도 최소값)은 값이 작을수록 모델 결과에 음의 영향을 미치고, SIM_MAX(초록 유사도 최대값)는 값이 클수록 모델 결과에 양의 영향을 미친다. 따라서 고품질 특허 그룹과의 초록 유사도가 높을수록 고품질 특허가 될 가능성이 높은 것을 알 수 있다. 그리고 NB_CLAIMS(청구항수)도 값이 클수록 모델 결과에 양의 영향을 미치므로, 청구항수가 많을수록 고품질 특허가 될 가능성이 높은 것을 알 수 있다.



<그림 8> ⑥번 모델의 SHAP value

Ⅴ. 결 론

본 연구에서는 인공지능 분야 특허 품질 예측을 위해 PATSTAT 데이터베이스의 미국(US) 특허 데이터를 활용하여 머신러닝 모델을 개발하였다. Doc2Vec 기반의 특허 초록 유사도 변수 포함 총 26개의 입력변수를 활용하였고, 이진 분류 문제로 변환하기 위해 피인용수 상위 5% 이상을 '고품질'로 정의하였다. 모델의 성능 평가 결과, 유사도 변수 제외 대비 유사도 변수를 포함했을 때 AUC는 0.013, f1-score는 0.025가 각각 높게 나타나 더 우수한 성능을 보였다. 모델 결과인 확률 값은 점수로 활용할 수 있으며, 점수에 따라 lift curve가 완만하게 내려감으로써 모델의 일반화 성능을 확인할 수 있었다. 이렇게 본 모델을 활용하여 인공지능 분야 특허 출원 시 향후 5년이내 고품질 특허가 될 가능성을 예측할 수 있다.

기존의 비정형 데이터를 이용한 연구는 특허 품질 예측보다는 특허 분류에 더 많이 집중되었으며, 특허 품질 예측은 정형 데이터 위주로 연구되었다. 본 연구는 정형 데이터에 비정형 데이터를 추가하여 비선형적인 머신러닝 모델을 통한 예측으로 더 나은 성능을 기대할 수 있다. NLP 기법의 Doc2Vec 알고리즘을 활용하여 특허 텍스트 정보를 모델에 반영하여 성능 개선에 기여하였고, SHAP을 활용하여 블랙박스 형태인 머신러닝 모델의 변수 영향도를 설명하였다. 이러한 접근방식은 향후 유사한 연구 분야에 좋은 사례가 될 수 있을 것이고 특허 품질 예측의 새로운 시각을 제시하였다.

이러한 연구 방법론을 통해 인공지능과 같은 핵심 기술 분야에서 특허의 품질을 예측하고, 고품질 특허 개발을 장려함으로써 사회적 가치를 실현하는 데 기여할 수 있을 것이다. 개발자는 고품질 특허를 개발하는 데 더욱 집중할 수 있고, 이는 기술 혁신과 산업 발전을 촉진하는 데 도움이 될 것이다. 따라서 특허 생태계에서의 혁신과 사회적 가치 실현을 위한 중요한 역할을 수행할 것으로 기대된다. 그러나, 고품질의 특허가 충분히 축적되지 않은 신규 부상 기술 분야에서는 활용 가능성이 낮다는 한계점도 존재하다.

향후 연구로 모델의 성능 향상을 위해 새로운 데이터를 확보하거나, 임베딩 값을 직접 변수로 활용하는 등 다양한 변수 개발을 고려할 수 있다. 그리고 transformer 기반의 더 효과적인 문서 임베딩 모델을 사용하여 유사도 변수를 더 발전시킬 수 있을 것이다. 또한, 고품질 특허 데이터의 특성 분석 및 저품질 특허에 대한 사전 필터링 방안 등 다양한 응용 가능성도 살펴볼 필요가 있다.

참 고 문 헌

(1) 국내문헌

고광수., 정원교., 신영근., 박상성., 장동식. (2011). 텍스트 마이닝을 이용한 특허정보검색 개발에 관한 연구. 한국산학기술학회논문지, 12(8), 3677-3688.

김동준. (2010). 특허출원인 선행기술정보 개시제도에 대한 비교법적 고찰. 전남대 법학 연구소, 91-132.

김성호., 김지표. (2019). QFD를 이용한 우수특허 선별에 관한 연구. 기술경영경제학회, 27(3), 83-112.

문진희., 권의준., 금영정. (2017). 특허 동시분류분석과 텍스트마이닝을 활용한 사물인터넷 기술융합 분석. 기술혁신연구, 25(3), 1-24.

백서인., 이현진., 김희태. (2020). 인공지능의 기술 혁신 및 확산 패턴 분석: USPTO 특허 데이터를 중심으로. 한국콘텐츠학회논문지, 20(4), 86-98.

송진주., 강승식. (2019). Doc2Vec을 이용한 특허문서 자동 분류. 한국정보처리학회 춘계학술발표대회 논문집, 239-241.

이앞길., 최근호., 김건우. (2020). LDA 토픽 모델링과 Word2vec을 활용한 유사 특허문서 추천연구. Information Systems Review, 22(1), 17-31.

이현상., 차오신., 신선영., 김규리., 오세환. (2022). 특허데이터 기반 한국의 인공지능 경쟁력 분석: 특허지표 및 토픽모델링을 중심으로. 정보화정책, 29(4), 43-66.

임홍래. (2019). 특허 인용 네트워크 분석을 활용한 국가연구개발사업 특허의 평가 방안. 기술혁신연구, 27(4), 1-19.

조현진., 이학연. (2018). 머신러닝 기법을 활용한 특허 품질 예측. 대한산업공학회 춘계공동학술대회 논문집, 1343-1350.

(2) 국외문헌

Ahmadpoor, Mohammad., Jones, Benjamin F. (2017). The dual frontier: Patented inventions and prior scientific advance. Science, 357(6351), 583-587.

Ahuja, G., & Morris Lampert, C. (2001). Entrepreneurship in the Large Corporation: A longitudinal Study of how established firms create breakthrough inventions. Strategic Management Journal, 22(6 7), 521-543.

Choi, J., Lee, J., Yoon, J., Jang, S., Kim, J., Choi, S. (2022). A two-stage deep learning-based system for patent citation recommendation. Scientometrics, 127, 6615-6636.

Chung, P., Sohn, S. Y. (2020). Early detection of valuable patents using a deep learning model: Case of semiconductor industry. Technological Forecasting and Social Change, 158, 120146.

Erdogan, Z., Altuntas, S., Dereli, T. (2022). Predicting Patent Quality Based on Machine Learning Approach. IEEE Transactions on Engineering Management, 1-14.

Ernst, H. (2003). Patent information for strategic technology management. World Pat. Inf., 25(3), 233-242.

Feng, S. (2020). The proximity of ideas: An analysis of patent text using machine learning. PLoS ONE, 15, 1-19.

Hong, S., Kim, J., Woo, H., Kim, Y., Lee, C. (2022). Screening ideas in the early stages of technology development: A word2vec and convolutional neural network approach. Technovation, 112.

Kim, J., Lee, S. (2015). Patent databases for innovation studies: A comparative analysis of USPTO, EPO, JPO and KIPO. Technological Forecasting and Social Change, 92, 332-345.

Kyebambe, Moses Ntanda., Cheng, Ge., Huang, Yunqing., He, Chunhui., Zhang, Zhenyu. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis, 125, 236-244.

Lanjouw, J., Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. The Economic Journal, 114(495), 441-465.

Le, Q., Mikolov, T. (2014). Distributed Representations of Sentences and Documents. Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2), 1188-1196.

Lee, C., Kwon, O., Kim, M., Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. Technological Forecasting and Social Change, 127, 291-303.

Li, S., Hu, J., Cui, Y., Hu, J. (2018). DeepPatent: patent classification with convolutional neural networks and word embedding. Scientometrics, 117, 721-744.

Lee, Y. G., Lee, J. D., Song, Y. I. (2006). An analysis of citation counts of ETRI-Invented US patents. ETRI Journal, 28(4), 541-544.

Lin, B. W., Chen, C. J., Wu, H. L. (2007). Predicting citations to biotechnology patents based on the information from the patent documents. International Journal of Technology Management, 40(1-3), 87-100.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Milan, M., Nan, J., Kenneth, G, H. (2023). Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. Strategic Management Journal, Wiley Blackwell, 44(2), 491-519.

Nathan, F., Kenneth, T. (2016). Patent Valuation with Forecasts of Forward Citations. Journal of Business Valuation and Economic Loss Analysis, 12(1), 101-121.

Niemann, H., Moehrle, M. G., Frischkorn, J. (2017). Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application. Technological Forecasting and Social Change, 115, 210-220.

Park, G., Park, Y. (2006). On the Measurement of Patent Stock as Knowledge Indicators. Technological Forecasting and Social Change, 73(7), 793-812.

Squicciarini, M., Dernis, H., Criscuolo, C. (2013). Measuring Patent Quality : Indicators of Technological and Economic Value. OECE Science, Technology and Industry Working Papers.

Wu, J. L., Chang, P. C., Tsao, C. C., Fan, C. Y. (2016). A patent quality analysis and classification system using self-organizing maps with support vector machine. Applied Soft Computing. 41. 305-316.

□ 투고일: 2023.07.17. / 수정일: 2023.09.07 / 게재확정일: 2023.09.12.