융합기술의 효과적 분류를 위한 CPC 특허 체계의 적합성 검토: 탐색적 접근

임수진[1], 이현진 [2], 박채나 [3], 금영정*

[1 2 3]서울과학기술대학교 산업공학과 학사과정

*서울과학기술대학교 산업공학과 부교수

[1]sujin990722@seoultech.ac.kr, [2]lhj9918@seoultech.ac.kr, [3]chaena7@seoultech.ac.kr

* 교신저자: yjgeum@seoultech.ac.kr

2022년 대한산업공학회 추계학술대회

Contents

01 서론 연구배경, 연구동기, 연구목적

- **연구배경**

  - 융합기술의 중요성
    - 최근 기술융합의 영향이 크게 확대되면서 미래 시장의 주도권을 선점하기 위해 융합기술이 중요 속성으로 작용하고 널리 활용되고 있음 (강희종, 엄미정, 김동명. 2006)
      - ➢ 혁신 기술의 등장으로 융합기술이 산업경쟁력을 제고시키는 중요 속성으로 작용 (유재흥, 조원영, 신정우, 2022)
    - 따라서 새로운 혁신을 위해 융합기술을 효과적으로 파악하고 분석하는 것이 중요한 활동으로 간주되어 왔음

  - 특허와 기술혁신과의 관계
    - 특허는 오랫동안 기술의 대리지표로 활용되어 왔으며, 혁신을 측정하기 위한 핵심 지표로 고려되어 왔음 (권오진, 노경란, 이방래, 고병열, 문영호. 2007, 윤민호. 2011)
      - ➢ 특허정보는 기술의 확산을 관찰하고 평가할 수 있는 지표로 활용되고 있음 (전상규. 2021)
      - ➢ 기술혁신의 중요한 지표인 특허정보를 활용해서 기술융합의 구조를 파악할 필요가 있음 (백현미, 김명숙. 2013)
    - 따라서 융합기술의 효과적 분석을 위해 특허분석이 널리 활용되어 왔으며, 융합연구를 위한 다수의 특허분석 연구가 수행되어 왔음 (전상규. 2021)
      - ➢ 특허를 기반으로 인용분석, 교차영향분석 등의 방법론으로 바이오 인포매틱스 분야의 원천 기술 흐름 및 융합 양상 파악, 향후 산업 융합의 가능성 살펴봄 (유준상, 이희상. 2013)

01 서론 연구배경, 연구동기, 연구목적

- **연구동기**
  - 대부분의 특허기반 기술융합 연구는 특허 CPC 분류체계를 바탕으로 융합현상을 분석하는 데 초점을 맞추고 있음
    (강희종 외 2인. 2006, 배영임, 신혜리. 2017, 정명석, 정소희, 이주연. 2018, 백서인, 이현진, 김희태. 2020)
    - 인공지능(AI) 관련 지식의 구성 요소와 AI 지식의 진화 궤적과 향후 연구개발 투자 방향에 대해 특허 CPC 코드를 활용 (김경외, 이준민, 이창준, 2021)
    - 특허는 보통 하나의 특허가 다수의 CPC 분류코드를 가지고 있기 때문에 CPC를 통한 기술융합 분석이 다수 연구 (심재륜. 2018)에서 이루어짐
    - CPC코드는 최소 1개부터 제한 없이 부여될 수 있으나 1개인 경우 50%, 2개인 경우 31.7%, 3개인 경우 12.3% (심우철, 민재옥, 조유정, 고봉수, 노한성, 2020)
    - CPC는 세분화된 분류를 제공해 기술 특성을 더 세밀하고 정확하게 표현하며, 특허분류코드를 활용해 기술융합현상을 분석하는 경우가 많음
      (강지호, 김종찬, 이준혁 박상성, 장동식, 2015)
  - 그럼에도 불구하고 현 CPC 분류체계가 융합기술 분류에 적합한가라는 근본적 질문을 다룬 연구는 거의 없음
    - 융합기술 분류를 다루고 있는 Y 섹션이 존재하기는 하지만, Y 섹션에 포함된 특허가 극히 제한적임 (부가정보로만 활용됨)
    - 최근 기술혁신의 다수를 차지하고 있는 기술이 대부분 다양한 산업배경을 가진 것을 고려할 때, 현 CPC 분류체계의 적합성을 검토할 필요가 있음
      (강지호 외 4인. 2015)
    - 융합기술을 분석함에 있어, 정적인 특허 분류코드를 활용하는 기존 연구방법들은 새로이 나타나는 기술들을 정의하고 이들 간의 파급관계를 분석하는 데 한계점을 지님 (정병기, 김정욱, 윤장혁. 2016)
- **연구목적**
  - 현재 CPC 분류체계의 융합기술 반영 현황 파악
  - 융합기술의 효과적 분류를 위한 CPC 특허 체계 재정립에 대한 탐색적 연구
  - 특허 문서들의 분류 과정을 기계학습을 통해 검토해보며 현 CPC 분류체계의 적합성 검증

02 선행연구 (1/3)

- **IPC, CPC 특허 체계를 이용한 기술융합 분석 관련 연구**
  - 특허 기반 기술융합 연구는 대부분 IPC, CPC 특허 체계를 이용한 기술 동향 파악 및 융합 패턴 탐색 중심으로 이루어져 있음
    - 융합기술 및 융합지수, 유망기술 및 유망지수를 정의하여 유망 융합기술을 예측하는 정량적인 방법을 제시함 (강희종 외 2인. 2006)
    - IPC를 활용하여 산업분류, 기술분류 코드를 매칭시키고, 인공지능 기술의 산업 간, 기술 간 등의 융합패턴을 분석함 (배영임, 신혜리. 2017)
    - 인공지능 기술에 대한 IPC 분류 기준 공백기술 분석 및 네트워크 분석을 통해 기술 혁신, 확산 패턴을 분석하며 인공지능의 발전 방향성을 제시함 (정명석 외 2인. 2018, 백서인 외 2인. 2020)
    - CPC를 활용하여 연관규칙분석 기반의 사물인터넷과 웨어러블 기술융합동향을 분석함 (강지호 외 2인. 2015)
    - 특허정보를 활용하여 계량데이터에 기반한 산업의 융합성 평가 방법론을 제안함 (김지은, 이성주. 2013)
      - ➢ 인용정보 활용한 기술연관분석
      - ➢ IPC 클래스를 하나의 기술로 보고 각 기술들 간의 인용 건수를 활용하여 융합유발계수 측정
    - 특허에 부여된 기술분류(IPC)를 활용하여 특정 기술의 주변기술과 융합되는 경향과 상대적인 집중도를 평가하고 예측할 수 있고, 융합기술의 재확산 강도와 규모를 예측할 수 있는 모형을 제시함 (전상규. 2021)

- **현 CPC 특허 체계의 재정립, 제안 관련 연구**
  - 현 CPC 분류체계가 특허 분류에 적합한가라는 근본적인 질문을 다룬 연구는 거의 없으며, 융합기술 분류에 초점을 맞춘 연구 또한 거의 존재하지 않음
    - 인공지능 기술을 대상으로 프로세스적 관점에서의 분류체계를 제시하며 특허 분석에 이를 적용함 (김지혜, 김병초. 2017)
    - 사용자 분류체계에 따라 특허문헌을 자동으로 분류하는 분류 모델 및 분류기 아키텍처를 설계함 (김성훈. 2021)
    - 일본의 CS-term 및 Facet ZIT 특허분류체계의 도입의 검토 필요성 제안함 (권지현, 2020)

02 선행연구 (2/3)

- **특허 자동 분류에 딥러닝 모델을 적용한 연구**
  - 특허의 특성을 고려한 자동 분류 모델 생성을 위해 다양한 연구가 진행됨
    - CNN을 이용하여 특허 분류 체계가 갖는 계층과 입력 데이터의 의미 관계를 고려한 모델을 제안함 (한동희. 2019)
    - 다중 레이블을 가지며 클래스에 따른 분포가 매우 불균형한 특허 데이터의 특성을 고려한 분류 모델 생성을 위해 BERT 기반의 향상된 극한 다중 레이블 분류 모델을 제안함 (정구익. 2020)
    - 클래스 불균형 상태의 특허데이터로부터 생성된 분류기의 문제점을 해결하기 위하여 MLP 모델에 모델 재귀적 오버 샘플링이라는 기법을 제안함 (김성훈, 김승천. 2021)
    - 언어 모델인 BERT를 이용하여 특허상품연계정보의 발명의 명칭 정보를 학습하고, 19개의 국제 상품 유사군명으로 다중 분류하는 방법을 제안함 (이관용. 2022)
    - 비교실험 결과를 통해 효과적인 분류모델과 단일 필드와 복합 필드에서 유효할 수 있는 특허문헌 필드 조합을 제안함 (심우철 외 4인. 2020)
      - ➢ 최근 자연어처리 분야에서 각광받고 있는 pre-trained BERT 모델을 관련연구에서 사용한 동일한 파라미터 값을 적용함
      - ➢ 한국어 특허 자동연구 발전에 기여하고자 함
    - R&D 과제 정보를 활용하여 BERT 기반 으로 자동적으로 문서 특징들을 추출하고 분류에 직접 활용하는 모델을 제안함 (황상흠, 김도현, 2020)

02 선행연구 (3/3)

- **텍스트 분류 모델에 XAI를 적용한 연구**
  - 도박사이트 분류 모델을 XAI 기법으로 분석하여 주요 키워드를 탐색하는 방법에 대하여 제안함 (이경석, 임규민, 조호묵. 2022)
    - 웹사이트의 키워드(텍스트 데이터)를 학습한 분류 모델에 XAI 적용함
  - Self-Attention 방법으로 기계 번역 알고리즘의 번역 오류 요인을 설명하는 방법을 제안함 (장청롱, 안현철. 2022)
  - 다중 레이블 모델에 적용할 수 있는 해석 가능성 접근 방식을 이용한 학생의 학습 스타일과 입력 활동의 상관관계를 연구함
    (Daiva Goštautaite, Leonidas Sakalauskas. 2022)

- **기존 연구의 한계점**
  - 특허 기반 기술융합 연구는 대부분 IPC, CPC 특허 체계를 이용한 기술 동향 파악 및 융합 패턴 탐색 중심으로 이루어져 있음
    - 융합기술은 다양한 산업으로의 파급이 활발히 이루어져 앞으로 국가와 사회에 큰 영향을 미칠 예정임 (정다운. 2018)
    - 따라서, 융합기술 관련 특허의 분류 체계가 명확하게 정의되어야 특허 검색 및 혁신 측정의 지표로 잘 활용될 수 있음
  - 특허가 융합기술 연구에 많이 사용되고 4차 산업혁명 관련 분류코드가 생겨나는 등 융합기술에 대한 특허의 발전이 이루어지고 있음
  - 그에 비해, 그 분류체계인 CPC가 융합기술이 잘 반영된 체계인지에 대한 적합성을 입증하는 연구는 부족한 상황임
  - CPC 분류코드가 발전하는 융합기술의 속도를 따라가지 못한다는 한계점을 밝힌 연구(정병기 외 2인. 2016) 가 있었지만,
    XAI / 토픽모델링 등을 사용하는 낮은 분류 성능에 대한 원인 탐색에 관한 연구는 존재하지 않음
    - 따라서, 자동 분류 모델에 XAI를 결합하여 분류 특허와 미분류 특허의 키워드를 비교함으로써 현재 CPC 체계의 적합성을 검토하고자 함

03 연구 프레임워크

데이터 수집

- 수집 범위 선정
- USPTO PatentView

데이터 전처리

- 레이블 원핫인코딩
- 특수문자 제거
- 불용어 제거
- 토큰화
- 명사 추출
- 명사 원형 복원

특허 탐색 기준 설정

BERT

분류 특허/미분류 특허 구분
성능에 따른 CPC subclass 구분

CPC 분류체계 적합성 검토

분류 특허 vs 미분류 특허 비교

MultiClassification Explainer

T-test

Y섹션 포함 비율 차이 검정

키워드 빈도 분석

CPC 예측에 높은 영향을 준 키워드 비교

Pretrained Fasttext

4차 산업 키워드와의 유사도 계산을 통한
융합기술적 특성 비교

미분류 CPC subclass 탐색

BERT Classification Report

↓

미분류 CPC subclass 정의

토픽모델링

토픽별 키워드 파악 및 미분류 원인 탐색

Rule 기반 분석

기타 미분류 원인 탐색

03 연구 프레임워크 BERT

- **BERT(Bidirectional Encoder Representations from Transformers)**

  Transformer의 인코더만을 다중으로 쌓은 다중 레이어 양방향 트랜스포머 인코더

  - 기존의 Transformer와 달리 양방향에서 Attention을 사용하는 모델
  - 위키피디아(25억 단어)와 BooksCorpus(8억 단어)와 같은 레이블이 없는 텍스트 데이터로,
    2가지 task [Masked Language Model, Next Sentence Prediction] 에 대해 사전 훈련된 언어 모델

- **BERT 선정 이유**

  Finetuning한 BERT 모델은 NLP 11개 분야에서 state-of-art 성능을 달성할 정도로 좋은 성능을 가짐

  → 현재 CPC 분류 체계를 가장 잘 설명할 수 있는 분류 모델을 적용하는 것이 목적이며,
  Input 이 특허 abstract 이므로 긴 문장을 처리하는데 특화된 BERT 모델을 선정하여 진행

- **본 연구에서의 활용**

  현재 CPC 분류체계로 분류되지 않는 특허를 구분하기 위해 BERT를 사용해 분류 특허와 미분류 특허를 구분하고자 함

03 연구 프레임워크 eXplainable Artificial Intelligence, XAI

- **Explainable AI (XAI)**
  - 인공지능의 행위와 판단을 사람이 이해할 수 있는 형태로 설명할 수 있는 인공지능

- **XAI 적용 방법**
  - transformers-interpret 라이브러리의 MultiLabelCalssificationExplainer 사용
  - transformers-interpret 라이브러리는 Captum으로부터 설계됨 → pytorch에서 model interpretability를 위해 설계된 패키지
  - Multi-label 모델 출력에 대한 단어 속성 및 시각화 출력 가능

- **본 연구에서의 활용**
  - BERT Classification에 적용
  - 분류 특허, 미분류 특허 대상으로 MultiLabelClassificationExplainer (XAI) 적용
  - 특허 분류에 영향을 주는 특허의 키워드 파악 및 특성 추출
  - 분류 특허들의 키워드와 미분류 특허들의 키워드 비교

03 연구 프레임워크 LDA 토픽모델링

- **LDA(Latent Dirichlet Allocation)**

  주어진 문서에 대하여 각 문서에 어떤 주제들이 존재하는지를 서술하는 확률적 토픽 모델 기법 중 하나

- **LDA 선정 이유**
  - 사전 확률을 고려하여 분류 성공률 높고 해석 용이
  - 노이즈가 많은 데이터이더라도 원하는 단어나 문장 추출 가능
  - 미분류 CPC subclass의 원인 해석 용도로 사용 (BERT 모델의 미분류 CPC subclass의 특허들이 노이즈가 높을 수 있다고 판단)

- **LDA 적용 방법**

  1. 토픽의 개수 k 정하기 : Perplexity(혼란도)와 Coherence(일관성) 고려하여 적정 passes와 k선정
     - Perplexity(혼란도) : 특정 확률 모델이 실제로 관측되는 값을 얼마나 잘 예측하는지를 의미. 값이 작아지는 것이 중요
     - Coherence(일관성) : 상위 단어 간의 유사도 계산으로 해당 주제가 의미론적으로 일치하는 단어들끼리 모여 있는지 파악
  2. 모든 문서의 단어를 k개 중 하나의 토픽에 랜덤 할당
  3. 어떤 문서의 각 단어 w는 자신은 잘못된 토픽에 할당되어져 있지만, 다른 단어들은 전부 올바른 토픽에 할당되어져 있는 상태라고 가정
     - p(topic t | document d) : 문서 d의 단어들 중 토픽 t에 해당하는 단어들의 비율
     - p(word w | topic t) : 각 토픽들 t에서 해당 단어 w의 분포
     - 이를 통해 각 단어 w가 어떤 토픽에 할당될지 결정
  4. 3번 과정 반복

- **본 연구에서의 활용**
  - 토픽 모델링에 활용된 단어들을 기반으로 특허 문서들의 내용 파악
  - CPC subclass에 적합하지 않는 토픽의 존재 여부 검토
  - Input : 특허의 abstract
  - Output : CPC subclass별로 선정한 토픽 개수 기반으로 각 토픽 생성에 영향을 끼친 키워드

03 연구 프레임워크 여러 방법론들

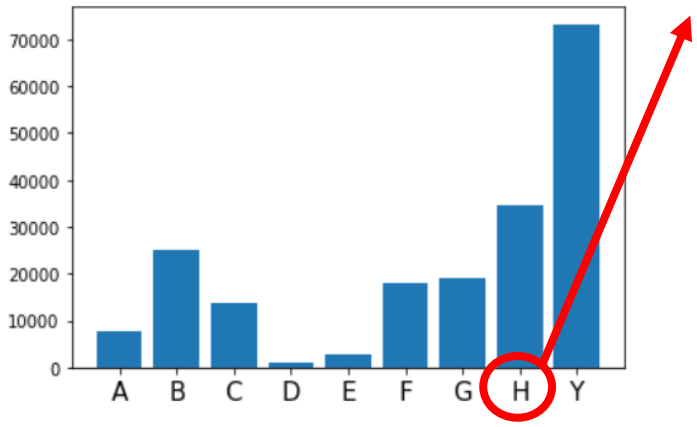
- **빈도 분석**
  - 본 연구에서의 활용 : 빈도에 기반한 CPC 예측에 높은 영향을 준 키워드 비교
- **Pretrained FastText**
  - FastText : Facebook 의 AI Research(FAIR) 연구소에서 만든 단어 임베딩 및 텍스트 분류 학습을 위한 라이브러리
    - 157개국의 언어에 대해 common crawler와 wikipedia의 데이터를 학습한 pre-trained model
  - 본 연구에서의 활용 : 4차 산업 키워드와의 유사도 계산을 통한 융합기술적 특성 비교
- **T-test**
  - 모집단의 분산이나 표준편차를 알지 못할 때, 모집단을 대표하는 표본으로부터 추정된 분산이나 표준편차를 가지고 검정하는 방법
  - '두 모집단의 평균간의 차이는 없다'라는 귀무가설과 '두 모집단의 평균 간에 차이가 있다'라는 대립가설로 판단하는 통계적 검정방법
  - 본 연구에서의 활용 : 분류 특허와 미분류 특허의 Y섹션 포함 비율 차이 검정
- **Rule 기반 분석**
  - 본 연구에서의 활용
    - 다양한 방법론으로 설명되지 않는 기타 미분류 CPC subclass의 원인 탐색
    - 정량적인 기준과 정성적인 기준을 포함한 프레임워크 설계

04 분석 및 결과 해석 데이터 수집 – 수집 범위 선정

- **Y섹션 크롤링**
  - 사이트 : USPTO PatentsView (https://datatool.patentsview.org/query/)
  - 수집기간 : 등록일자 기준 2020.01.01 ~ 2021.12.31 (2개년)
  - 데이터개수 : 73,246개

- **Y 섹션 내 다중 레이블 특허 개수**

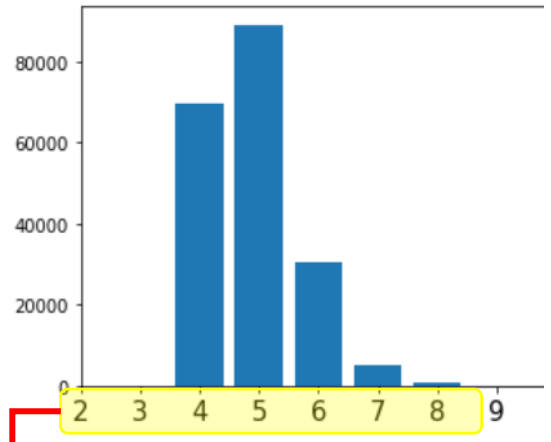
H 섹션 CPC를 포함한 특허가 가장 많음



Y 섹션 내 특허들과 가장 연관된 섹션 파악
→ 융합기술을 많이 포함하고 있는 섹션이라고 판단할 수 있음

- **Y 섹션의 다중 분류 섹션 별 개수 현황**

cpc_section_id	
3	89043
2	69574
4	30380
5	5180
6	762
7	126
9	9
8	8



Y섹션을 포함한 다중 분류 CPC 섹션 개수

융합기술을 많이 포함하는 "H"섹션 선택

Y 섹션은 부가적인 정보로써 단독으로 할당되어 있지 않기 때문에 Y섹션에서 얻은 다중 레이블을 바탕으로
Y섹션과 가장 많이 연관되는 섹션인 **'H섹션'**을 융합기술을 많이 포함하고 있는 섹션으로 선택

04 분석 및 결과 해석 데이터 수집 및 전처리

- **데이터 수집**

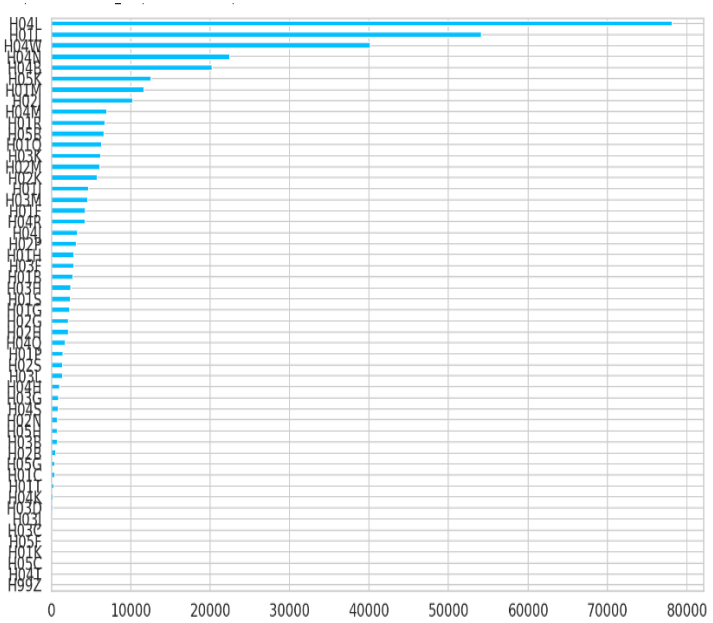
  H 섹션 크롤링 (특허제목, 출원번호, 출원일자, abstract, claim 등)

  - 사이트 : USPTO PatentsView (https://datatool.patentsview.org/query/)
  - 수집기간 : 등록일자 기준 2020.01.01 ~ 2021.12.31 (2개년)
  - 수집대상 : H섹션
  - 수집변수 : Patent Number, Patent Title, Patent Grant Date, Patent Type, CPC class, subclass, group, subgroup ID, abstract
  - 데이터개수 : 242,844개

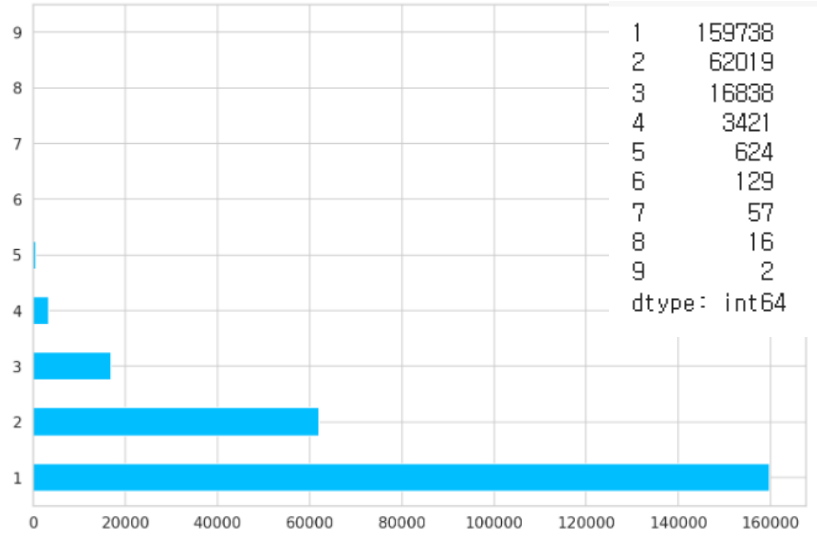
- **데이터 전처리**

  - 52개의 레이블 원핫인코딩
  - 특수문자 제거
  - 불용어 제거
  - 토큰화
  - 명사 추출
  - 명사 원형 복원

[CPC 레이블 분포]



[특허 한 개당 CPC 레이블 개수]



→ 각 클래스가 매우 불균형함을 확인 (sparse)

04 분석 및 결과 해석 BERT

< Input >
특허 Abstract

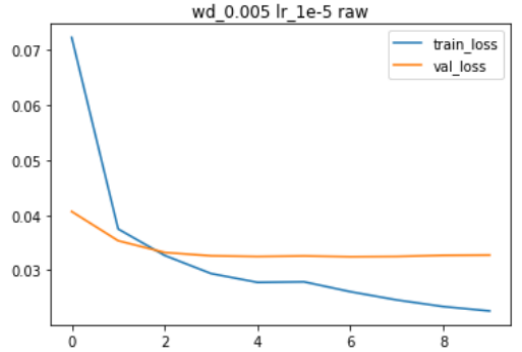
< Model >
BERT Classification (Multi-label)

< Output >
CPC Subclass

전처리한 특허 abstract와 CPC subclass를 바탕으로
BERT multi-label classification 학습 및
파라미터 튜닝을 진행하여
분류 특허와 미분류 특허를 구분하고자 함

- **Finetuning**
  - BERT 모델 종류 선택 (L: layer 개수, H: hidden layer 크기, A: self-attention의 head 개수)
    - BERT_base: L=12, H=768, A=12, Total Parameter=110M
    - BERT_large L=24, H=1024, A=16, Total Parameter=340M
  - 파라미터 튜닝 → **선정 모델 및 파라미터 : BERT_base, Raw Data, learning rate : 1e-5, weight decay : 0.005, epoch : 6**

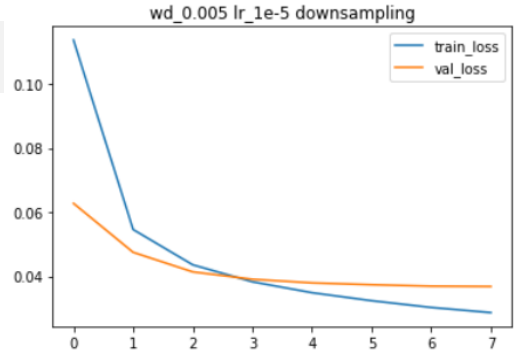
최종 모델 (raw)



	Unnamed: 0	precision	recall	f1-score	support
0	samples avg	0.769302	0.692545	0.708126	35161.0
1	micro avg	0.857416	0.625608	0.723395	35161.0
2	weighted avg	0.848908	0.625608	0.704341	35161.0
3	macro avg	0.628951	0.354296	0.420326	35161.0

후보 모델 (downsampling)

Sparse한 matrix의 비율을
맞추어 downsampling 진행



	Unnamed: 0	precision	recall	f1-score	support
0	samples avg	0.645549	0.559836	0.582550	35068.0
1	weighted avg	0.831423	0.493128	0.595539	35068.0
2	macro avg	0.460577	0.196976	0.249358	35068.0
3	micro avg	0.876039	0.493128	0.631039	35068.0

04 분석 및 결과 해석 특허 탐색 기준 설정

- **특허 탐색 기준 설정**
  - BERT 파라미터 튜닝을 통해 선정한 최종 모델에 테스트 데이터 24,285개를 적용함
  - 1) 분류 특허와 미분류 특허 비교 2) 미분류 CPC subclass 탐색을 통해 CPC 분류체계의 적합성을 검토하고자 함

- **분류 특허와 미분류 특허 구분**
  - 판단 기준 : 특허의 multi-label true값과 pred값의 일치 여부 (하나라도 예측이 틀린 subclass가 존재하면 미분류 특허라고 가정)
  - 분류 특허 (13,174개) / 미분류 특허 (11,111개)
  - 분류 특허와 미분류 특허의 비교를 위해 MultiClassificationExplainer와 t-test를 적용

- **성능에 따른 CPC subclass 구분**
  - 판단 기준 : BERT Classification Report 기반 CPC subclass별 성능 그래프 확인
  - 성능 그래프를 기준으로 미분류 CPC subclass 정의

04 분석 및 결과 해석 XAI

분류 특허와 미분류 특허를 비교하기 위해 분류에 영향을 준 키워드를 추출하고자 함

분류 특허, 미분류 특허 각각 10%씩 test dataset 레이블 비율대로 랜덤샘플링을 진행하여 MultiLabelClassificationExplainer를 적용함

- **키워드 추출 방법**
  - 미분류 특허
    - XAI

      XAI 적용 결과, 각 특허 별 52가지 H섹션 레이블의 분류에 높은 확률 (attribution score)로
      영향을 미친 상위 5개의 명사를 딕셔너리 형태로 저장

      ➢ True(1)인 레이블을 False(0)로 예측한 경우에 해당하는 XAI 딕셔너리

      ➢ False(0)인 레이블을 True(1)로 예측한 경우에 해당하는 XAI 딕셔너리
    - 명사 추출

      ➢ 미분류 특허 전체에 대한 명사 추출 결과
  - 분류 특허
    - XAI

      ➢ 분류 특허 전체에 해당하는 XAI 딕셔너리
    - 명사 추출

      ➢ 분류 특허 전체에 대한 명사 추출 결과

04 분석 및 결과 해석 키워드 빈도 분석

XAI와 명사 추출 방법으로 추출한 키워드의 빈도를 기반으로 미분류 특허의 특성을 파악해보고자 함

- **빈도수 기반으로 그룹 세분화**
  - 상위(0~25%), 중위(25~50% / 50~75%), 하위(75~100%) 그룹
  - 빈도분석 결과의 경우, H 섹션 관련 키워드가 많이 분포하는 상위 그룹에서는 유의미한 차이를 볼 수 없을 것이라 예상하여, 중위 하위 그룹에서의 키워드 분포를 함께 확인하고자 함.
- **미분류 특허**
  - 미분류 특허 키워드 빈도 분석 결과 (그룹별 빈도 상위 단어 8개씩 추출)

분류X	XAI 사용				XAI 사용				명사 추출			
	True label을 False로 예측				True label을 False로 예측				미분류 특허 전체			
	상위 ~25%	중위 ~50%	중위 ~75%	하위 ~100%	상위 ~25%	중위 ~50%	중위 ~75%	하위 ~100%	상위 ~25%	중위 ~50%	중위 ~75%	하위 ~100%
1	wireless	mother	absorb	par	light	die	plug	pump	device	longitudinally	matrix	Object
2	device	ventilation	traction	containers	battery	recording	software	installation	data	digitized	ability	PnP
3	power	images	width	broader	audio	management	generation	quality	system	recognition	gain	capture
4	light	comprising	isolation	stresses	network	bond	header	protector	method	articles	exhibition	humans
5	data	quantum	set	many	device	conduct	invention	housing	plurality	example	captures	storable
6	circuit	resonance	condition	powers	signal	online	lamp	protecting	signal	enter	rotational	bay
7	signal	lc	transit	linear	data	length	vector	umbrella	configured	document	micro	pertains
8	coil	range	request	boost	information	transmitted	adjustment	Imaging	power	dependent	attributes	rest

04 분석 및 결과 해석 키워드 빈도 분석

- **분류 특허**
  - 분류 특허 특허 키워드 빈도 분석 결과 (그룹별 빈도 상위 단어 8개씩 추출)

분류O	XAI 사용				명사 추출			
	상위 ~25%	중위 ~50%	중위 ~75%	하위 ~100%	상위 ~25%	중위 ~50%	중위 ~75%	하위 ~100%
1	layer	visited	socket	training	device	offered	Robotic	beamformers
2	beam	lever	distributed	domains	layer	vocabulary	Candidate	Hermitian
3	power	opera	read	dirt	data	normalizing	Monitors	transposition
4	radio	shipment	zone	law	method	offset	shotgun	turbulators
5	device	satellite	courier	impact	plurality	secrets	damp	departing
6	circuit	select	bundle	variable	semiconductor	level	ratchet	strength
7	base	settings	cut	responding	system	naming	photocells	declining
8	electrode	difference	generates	integral	configured	loader	cassette	space

- 미분류 특허와 분류 특허의 XAI 키워드를 비교했을 때, 분류 특허에 비해 미분류 특허에서 H섹션인 '전기' 관련 키워드의 출현빈도가 높음
- XAI 키워드와 명사 추출 키워드를 비교했을 때, 명사 추출 키워드는 상대적으로 전기 관련 키워드가 적게 등장한 반면, XAI 키워드는 H섹션 subclass title에 관련된 단어들이 더 많이 포함됨
- 융합기술의 포함 정도를 판단하고자 했지만, 판단 기준이 주관적이므로 정량적인 판단 방법인 유사도 계산을 진행함

04 분석 및 결과 해석 Pretrained FastText

4차산업혁명 관련 기술분야 특허분류 체계 키워드를 번역하여 융합기술 키워드로 정의한 후, XAI와 명사추출 방법으로 추출한 분류/미분류 특허의 키워드를 keyword to keyword 방식으로 Pretrained FastText 유사도를 계산하여 융합기술적 특성을 비교함

- **특허청 4차산업혁명 관련 新특허분류체계 키워드**

'ai','data','cloudcomputing','iot','system','uav','blockchain','smartcity','renewableenergy','platform',
'communication','virtual','reality','drone','healthcare','robot','semiconductor','autonomous'

- **미분류 특허 유사도 계산 결과**

분류X	XAI 사용				명사 추출	
	True label을 False로 예측		True label을 False로 예측		미분류 특허 전체	
	합계	평균	합계	평균	합계	평균
상위 25%	1101	0.1805	393	0.1987	7353	0.1947
중위 25~50%	939	0.1544	338	0.1707	7106	0.1882
중위 50~75%	944	0.1547	309	0.1578	6771	0.1793
하위 100%	945	0.1554	305	0.1541	6660	0.1765
전체	3930	**0.1613**	1346	**0.1703**	27891	**0.1847**

- **분류 특허 유사도 계산 결과**

분류O	XAI 사용		명사 추출	
	합계	평균	합계	평균
상위 25%	2600	0.1618	7978	0.1947
중위 25~50%	2384	0.1485	7614	0.1859
중위 50~75%	2234	0.1390	7185	0.1753
하위 100%	2063	0.1285	6832	0.1668
전체	9282	**0.1445**	29611	**0.1806**

명사추출 결과로 유사도를 계산한 결과, 분류 특허와 미분류 특허의 차이가 나타나지 않음

분류에 있어 Explainability한 키워드로 구성된 XAI 딕셔너리로 계산한 결과, 미분류 특허 키워드의 융합기술 유사도가 0.02 정도 더 높게 나타남

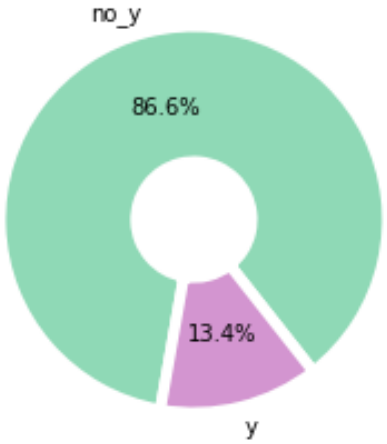
→ **미분류 특허들의 융합기술 관련도가 더 높다고 볼 수 있음**

04 분석 및 결과 해석 T-test

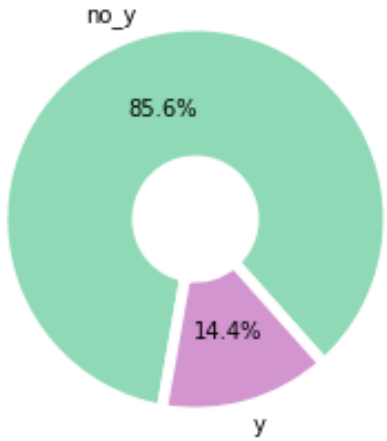
분류 특허와 미분류 특허의 Y섹션 포함 비율의 차이를 검정하고자 함

각 특허의 다중 CPC 섹션 중 Y섹션의 포함 여부에 따라 0과 1로 정의하여 t-test를 진행함

[Y섹션을 포함하는 분류 특허의 비율]



[Y섹션 포함하는 미분류 특허의 비율]



T-test : 두 집단 간의 비율 차이를 확인하기 위한 검정

[가설 설정]

귀무가설: 분류 특허의 Y섹션 포함 비율과 미분류 특허의 Y섹션 포함 비율은 차이가 없다.

대립가설: 분류 특허의 Y섹션 포함 비율과 미분류 특허의 Y섹션 포함 비율은 차이가 있다.

[검정 결과]

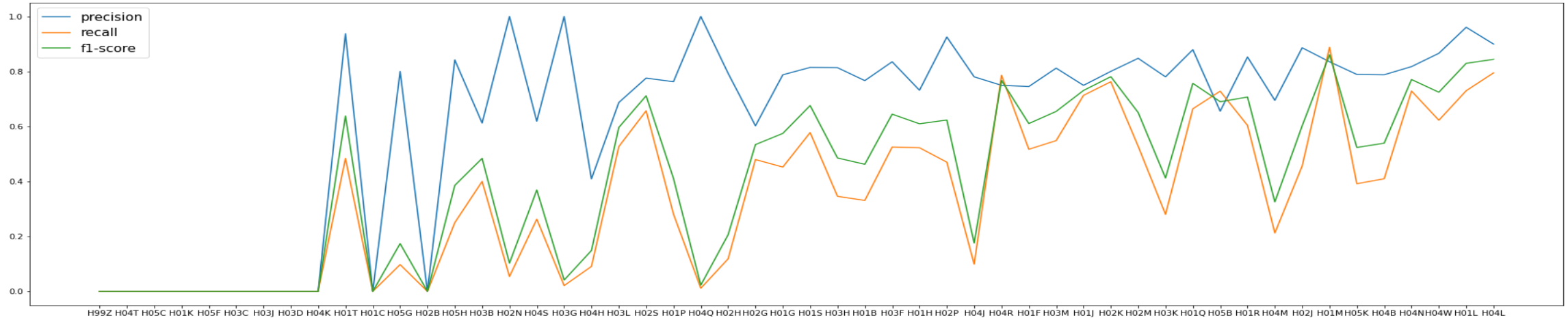
P-value = 0.025 로 0.05보다 작으므로 유의수준 0.05에서 귀무가설을 기각하여 두 집단 간의 비율은 차이가 있다고 볼 수 있음

하지만 특허의 개수가 많고, 유의수준의 설정 범위에 따라 귀무가설 기각 여부가 달라질 수 있다는 한계가 존재함

04 분석 및 결과 해석 BERT Classification Report

미분류 CPC subclass 탐색을 위한 subclass 구분을 정의하고자 함

- **CPC subclass 별 성능 그래프**
  - X축: 해당 CPC subclass에 포함되는 특허의 개수를 기준으로 오름차순 정렬



- **미분류 CPC subclass 정의**

  전체 데이터 개수의 0.01% (24.3개)를 임계값으로 정의

  - **CPC subclass의 데이터 개수가 임계값 미만인 경우**
    - H05C, H04T, H99Z (3가지)
  - **CPC subclass의 데이터 개수가 임계값 이상인 경우**
    - **전혀 예측을 하지 못하는 CPC subclass**
      - ➢ 선정 기준 : f1-score와 recall 성능이 0인 CPC subclass
      - → H01K, H05F, H03C, H03J, H03D, H04K, H01C, H02B --- (8가지)
    - **위 그래프 기준 주변 CPC subclass에 비해 낮은 성능을 가지는 CPC subclass**
      - ➢ 선정 기준 : f1-score와 recall이 주변 CPC subclass에 비해 낮아지는 CPC subclass
      - → H02N, H03G, H04Q, H03H, H01B, H04J, H03K, H04M, H05K, H04B --- (10가지)

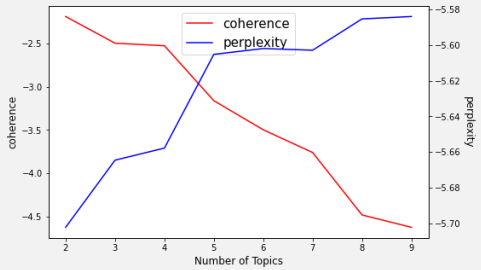
04 분석 및 결과 해석 토픽모델링

토픽모델링을 통해 미분류 CPC subclass를 대표하는 토픽을 정의하고 해당 CPC subclass에 적합하지 않는 토픽이 존재하는지를 파악하고자 함
BERT Classification Report를 통해 미분류 CPC subclass 로 정의한 subclass를 true label로 갖는 특허들을 추출해 토픽모델링 진행
Perplexity와 Coherence를 고려하여 적정 passes와 토픽 개수 선정

H03K

[Test data H03K 개수]
596개

[토픽 개수]
4개 / pass : 20

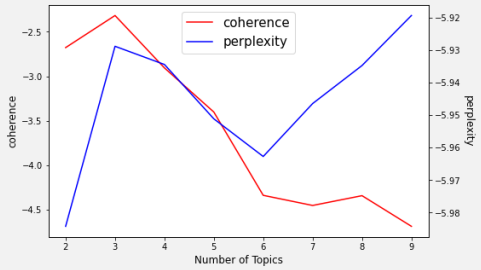


Pulse 증폭 기술	voltage	transistor	cell	source	supply	line	device	method	data	portion
반도체 스위칭 이용한 pulse 기술	circuit	device	data	control	power	voltage	semiconductor	output	element	memory
센서와 스위치를 이용한 출력값 제어	control	unit	switch	power	device	circuit	value	voltage	state	sensor
컴퓨터 논리회로 적용 방법	output	circuit	signal	transistor	input	voltage	clock	gate	reference	level

H04M

[Test data H04M 개수]
697개

[토픽 개수]
5개 / pass : 25

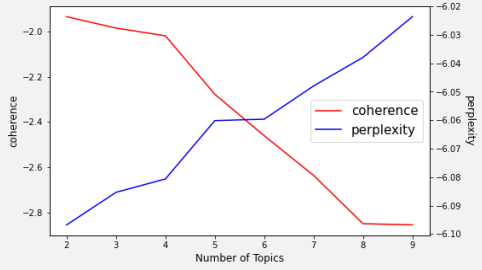


전화통신 전반	call	service	data	method	network	customer	system	time	session	user
스위칭 센터	system	information	layer	notification	element	center	unit	contact	value	customer
전화시스템	device	system	communication	user	information	message	method	network	data	interface
전화 통신할 때 네트워크 측면에서의 절차	communication	data	unit	device	system	information	method	signal	network	terminal
가정집에 설치하는 전화통신	display	portion	device	area	housing	surface	screen	module	member	cover

H04B

[Test data H04B 개수]
1987개

[토픽 개수]
4개 / pass : 25



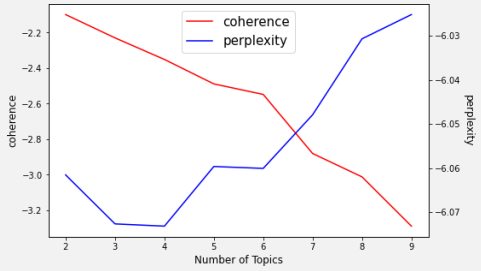
방송국에서 전송시스템 전송 방법	communication	information	transmission	unit	power	station	signal	method	channel	system
전송시스템 원리	signal	circuit	frequency	output	antenna	power	module	input	phase	portion
광 전송 시스템	beam	plurality	channel	antenna	system	signal	port	node	element	reference
무선 전송 시스템	device	communication	data	network	system	wireless	method	information	time	plurality

04 분석 및 결과 해석 토픽모델링

H05K

[Test data H05K 개수]
1261개

[토픽 개수]
6개 / pass : 10

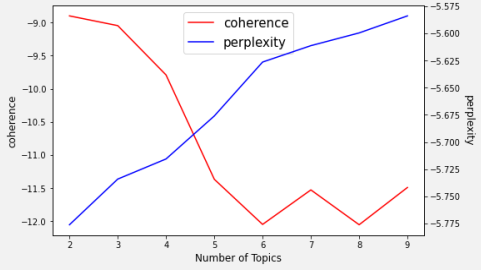


수신기, 수상기와의 조합법	portion	display	member	area	heat	surface	board	device	panel	circuit
전기장치 구성	device	system	component	air	heat	housing	connector	unit	end	rack
인쇄회로 구성	surface	plate	substrate	side	base	conductor	component	line	device	heat
전자장치에서의 인쇄회로	power	module	portion	device	circuit	battery	output	unit	terminal	member
전기장치	part	housing	layer	surface	component	body	region	portion	element	direction
인쇄회로 장치 및 제조법	layer	circuit	board	metal	plurality	surface	device	pad	structure	method

H02N

[Test data H02N 개수]
74개

[토픽 개수]
3개 / pass : 10

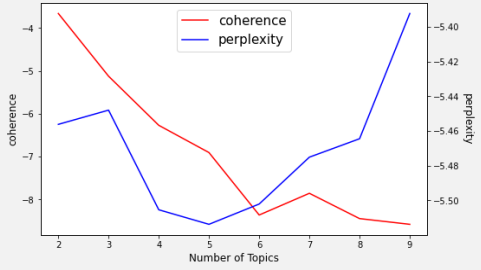


전동기 원리	surface	vibration	direction	portion	member	region	element	body	point	system
열에너지를 이용하는 전동기 원리	electrode	layer	actuator	power	element	part	beam	motion	energy	body
전하 제거에 의한 발전기와 진동기	power	device	plurality	generator	vibrator	member	frequency	surface	charge	message

H03G

[Test data H03G 개수]
95개

[토픽 개수]
4개 / pass : 20

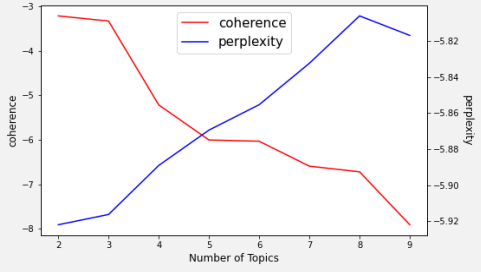


증폭기 전압이득	circuit	gain	output	path	voltage	system	method	signal	source	input
증폭기 회로 설계	power	circuit	voltage	supply	transistor	amplifier	frequency	state	gain	source
증폭기로 만들어진 Audio의 volume 제어	volume	voltage	level	system	control	signal	audio	device	output	group
Audio 재생	signal	device	output	input	playback	control	circuit	gain	audio	plurality

H04Q

[Test data H04Q 개수]
175개

[토픽 개수]
3개 / pass : 10



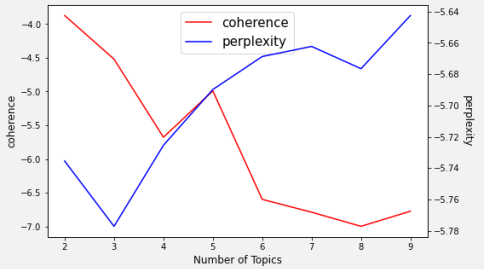
Switch 설계	sensor	plurality	data	channel	output	device	signal	system	node	switch
대화장치 시스템	network	device	data	communication	system	information	method	service	unit	plurality
네트워크 시스템에서의 정보 이동	data	system	power	node	network	unit	sensor	device	information	method

04 분석 및 결과 해석 토픽모델링

H03H

[Test data H03H 개수]
240개

[토픽 개수]
5개 / pass : 15

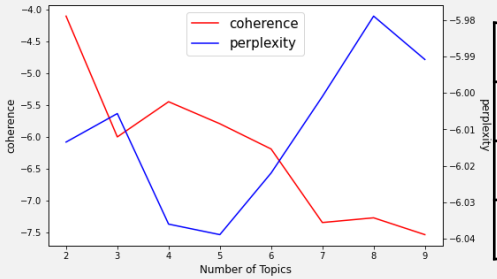


공진 회로 전반	capacitor	member	terminal	layer	resonator	surface	plate	method	filter	wave
공진기의 진동수 필터링	resonator	portion	electrode	arm	frequency	circuit	element	part	connection	filter
공진회로 작동법	circuit	Impedance	resonator	device	control	transistor	plurality	voltage	stage	capacitance
Transistor 원리	surface	layer	electrode	Substrate	film	region	portion	line	side	wave
공진기 output 제어	signal	circuit	output	device	port	input	filter	frequency	switch	voltage

H01B

[Test data H01B 개수]
278개

[토픽 개수]
4개 / pass : 25

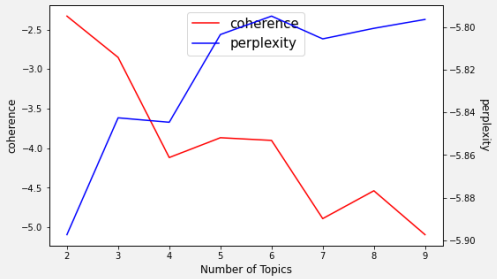


케이블의 구성 요소	layer	particle	body	part	device	group	surface	resin	polymer	conductor
도체 케이블	wire	portion	conductor	plurality	material	power	harness	member	cable	section
탄소 섬유 케이블	material	line	surface	member	particle	method	carbon	direction	step	power
케이블 제조 장치	layer	cable	film	metal	composition	conductor	resin	structure	end	invention

H04J

[Test data H04J 개수]
323개

[토픽 개수]
3개 / pass : 45



광다중화 시스템	network	data	unit	plurality	system	method	packet	information	fiber	beam
통신장치	signal	data	network	plurality	port	channel	device	time	node	transmission
다중 통신	communication	device	system	transmission	station	information	base	sequence	method	cell

04 분석 및 결과 해석 토픽모델링

- **CPC subclass 별 토픽모델링 결과 표**

CPC	topic1	topic2	topic3	topic4	topic5	topic6
H03K	Pulse 증폭기술	반도체 스위칭 이용한 pulse기술	센서와 스위치를 이용한 출력 값 제어	컴퓨터 논리회로 적용 방법		
H04M	전화통신 전반	스위칭 센터	전화시스템	전화 통신할 때 네트워크 측면에서의 절차	가정집에 설치하는 전화통신	
H04B	방송국에서 전송시스템 전송 방법	전송시스템 원리	광 전송 시스템	무선 전송 시스템		
H05K	수신기, 수상기와의 조합법	전기장치 구성	인쇄회로 구성	전자장치에서의 인쇄회로	전기장치	인쇄회로 장치 및 제조법
H02N	전동기 원리	열에너지를 이용하는 전동기 원리	전하 제거에 의한 발전기와 진동기			
H03G	증폭기 전압이득	증폭기 회로 설계	증폭기로 만들어진 Audio의 volume 제어	Audio 재생		
H04Q	Switch 설계	대화정치 시스템	네트워크 시스템에서의 정보 이동			
H03H	공진회로 전반	공진기의 진동수 필터링	공진회로 작동법	Transistor 원리	공진기 output 제어	
H01B	케이블의 구성 요소	도체 케이블	탄소섬유 케이블	케이블 제조장치		
H04J	광다중화 시스템	통신장치	다중통신			

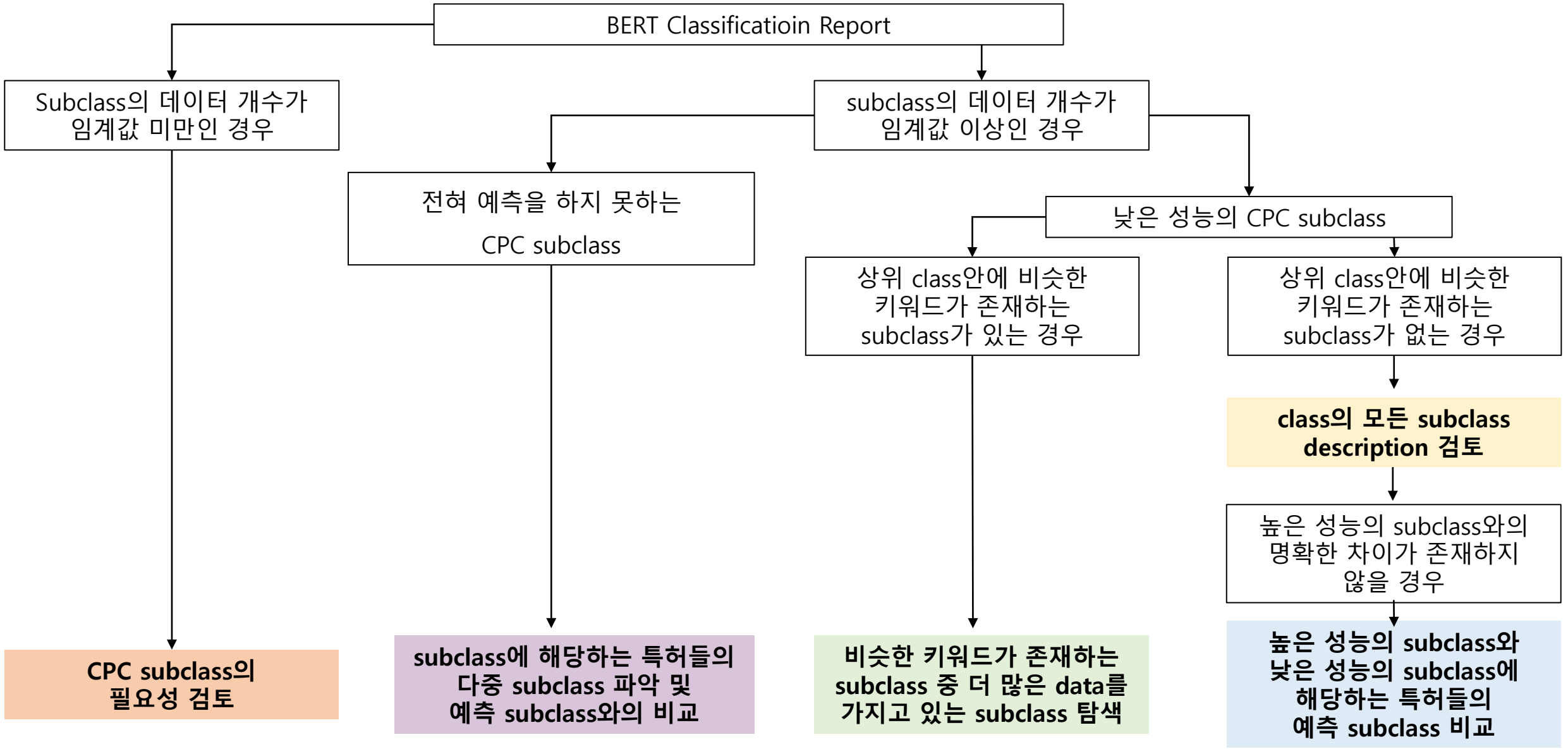
CPC subclass별로 적정 passes와 토픽의 개수를 선정하여 토픽모델링 진행

각 토픽 별 기여도가 높은 키워드 10개를 살펴본 결과, 해당 CPC subclass과 연관되어 있는 단어들이라고 판단함

해당 CPC subclass과 연관이 없다고 판단되는 토픽은 존재하지 않음

04 분석 및 결과 해석 Rule 기반 분석

정량적인 기준과 정성적인 기준을 포함한 프레임워크 설계를 통해 미분류 CPC subclass의 원인을 탐색하고자 함



04 분석 및 결과 해석 Rule 기반 분석

- **탐색 방법 : subclass에 해당하는 특허들의 다중 subclass 파악 및 예측 subclass와의 비교**

레이블	레이블 설명	탐색 결과	분류가 잘되지 않은 원인
H02B	2년간 특허 중 536개 존재 Test 데이터셋: 48개	• 2개의 CPC subclass로 구성된 특허가 가장 많으며, H01R/H02G/H01H/H05K 등이 관련 높음 • 관련 높은 CPC subclass 중 support가 높은 H01H/H01R/H02G로 예측되는 경우 많음	연관된 CPC subclass 중 데이터 수가 많은 subclass로 분류되는 경향
H01C	2년간 특허 중 431개 존재 Test 데이터셋: 47개	• Support가 낮은 CPC subclass이지만, H01C 단일 클래스를 가진 특허가 전체의 32%를 차지하며 가장 많음 → 고유성이 있는 CPC subclass • 어떤 CPC로도 예측하지 못하는 경우 많음	H01C 자체로 예측을 잘 하지 못해서 어떤 CPC로도 예측이 안되는 경우
H04K	2년간 특허 중 245개 존재 Test 데이터셋: 26개	• Support가 낮은 CPC subclass이지만, H04K 단일 클래스를 가진 특허가 전체의 24%를 차지하며 가장 많음 → 고유성이 있는 CPC subclass • Support가 높은 H04L/H04W/H04B 등이 관련 높으며, 이 3가지 subclass로 예측되는 경우가 많음	연관된 CPC subclass 중 데이터 수가 많은 subclass로 분류되는 경향 → H04K의 데이터 수 부족
H03D	2년간 특허 중 219개 존재 Test 데이터셋: 27개 [H03D = 하나의 반송파에서 다른 반송파로의 변조의 변조 또는 전이]	• 3개 다중 CPC로 구성된 특허가 가장 많으며, H04B/H04L 등이 관련이 높음. • H03L로 예측되는 경우가 많음 → '주파수' 관련 공통점이 있음 • H04L은 예측 성능이 높은 CPC인데, 예측 결과 H03D와 같이 구성된 H04L의 예측력이 떨어지는 특징이 나타남.	데이터 수가 많은 subclass로 분류하기 보다는 연관된 CPC 특징을 가진 subclass로 분류하는 경향
H03J	2년간 특허 중 125개 존재 Test 데이터셋: 12개	• H01L + H03F + H03H + H03J + H04B 가 조합된 결과가 가장 많음. • H03J로 분류된 특허들이 대부분 단일 CPC 보다는 다중 CPC로 구성되어 있음 → H03J의 필요성을 확인 할 필요가 있음	연관된 다중 CPC subclass 중 데이터 수가 많은 subclass로 분류되는 경향

04 분석 및 결과 해석 Rule 기반 분석

레이블	레이블 설명	탐색 결과	분류가 잘되지 않은 원인
H03C	2년간 특허 중 99개 존재 Test 데이터셋: 6개	• H03L과 같이 구성된 특허가 가장 많으며, 다중 CPC subclass이 대부분 H03 계열에 속함	연관된 다중 CPC subclass 중 데이터 수가 많은 subclass로 분류되는 경향
H05F	2년간 특허 중 99개 존재 Test 데이터셋: 11개 [H05F = 정적 전기; 자연 발생 전기]	• H05F 단일 클래스로 분류된 특허가 가장 많음 • 다양한 CPC 를 가지고 있음에도 해당 CPC 중 하나라도 예측을 못하는 경우가 많이 보임 • True label에 관련없이 Support 높은 CPC로만 예측하는 경향을 보임	H05 계열이 '달리 분류되지 않는 전기 기술'을 나타내는 클래스 → Description에서 '전기' 를 포함하고 있어서 '전기'를 나타내는 H섹션의 support 높은 일부 CPC subclass로 예측하는 경향
H01K	2년간 특허 중 43개 존재 Test 데이터셋: 11개	• H01L, H05B로 예측되는 경우만 존재	연관된 다중 CPC subclass 중 데이터 수가 많은 subclass로 분류되는 경향

다중분류개수	다중분류 subclass	합계	전체 합계
H02B only			68
H02B + 1	H02B + H01H	73	217
	H02B + H05K	51	
	H02B + H02G	35	
	H02B + H01R	21	
	…	37	
H02B + 2	H02B + H01R + H05K	15	151
	H02B + H01R + H02G	12	
	H02B + H01H + H01R	10	
	…	114	
H02B + 3	H02B + H02J + H02K + H02P	14	70
	H02B + H01R + H02G + H05K	8	
	…	48	
H02B + 4 ~ 7			30
	-		536

H02B 다중 CPC 분류 현황

다중분류개수	다중분류 subclass	합계	전체 합계
H01C only			142
H01C + 1	H01C + H05B	47	141
	H01C + H01L	24	
	H01C + H05K	17	
	H01C + H01H	15	
	…	38	
H01C + 2	H01C + H01G + H05K	11	85
	H01C + H01H + H02H	6	
	…	68	
H01C + 3	H01C + H01F + H01G + H01L	6	43
	H01C + H01H + H01T + H02H	4	
	…	33	
H01C + 4 ~ 7			20
			431

H01C 다중 CPC 분류 현황

04 분석 및 결과 해석 Rule 기반 분석

- **탐색 방법 : CPC subclass의 필요성 검토**

레이블	레이블 설명	탐색 결과	분류가 잘되지 않은 원인
H05C	2년간 특허 중 12개 존재	• 2년 간 특허 중 H05C로 분류된 특허는 12개로 매우 적음 → H05C의 필요성을 확인 할 필요가 있음	전체적인 데이터 부족으로 분류 모델에서 학습이 되지 않음
H04T	2년간 특허 중 3개 존재	• 2년 간 특허 중 H04T로 분류된 특허는 3개로 매우 적음 → H04T의 필요성을 확인 할 필요가 있음	전체적인 데이터 부족으로 분류 모델에서 학습이 되지 않음
H99Z	2년간 특허 중 1개 존재	• 2년 간 특허 중 H99Z로 분류된 특허는 1개로 매우 적음 → H99Z의 필요성을 확인 할 필요가 있음	전체적인 데이터 부족으로 분류 모델에서 학습이 되지 않음

- **탐색 방법 : 높은 성능의 subclass와 낮은 성능의 subclass에 해당하는 특허들의 예측 subclass 비교**

레이블	레이블 설명	탐색 결과	분류가 잘되지 않은 원인
H05K	인쇄회로 - 전기장치/부품 - 제조, 케이싱, 부품, 신뢰성 개선, 공정, 인덱싱, 조합	• H05B, H05K를 제외하고는 매우 적은 test data 의 개수를 가지고 있음 • 잘못 예측되는 subclass의 분포가 고르게 분포 • H01 subclass에도 인쇄회로 관련 subclass 존재	모든 전기/전자 관련 제품에 인쇄회로가 있기 때문에 상대적으로 다양하게 예측
H03H	임피던스 회로망 – 공진 회로/공진기, 증폭/대역폭, 통신, 주파수, 네트워크, 필터	• H03H 다음으로 가장 많이 예측하는 H01L subclass에도 공진기 관련 subclass 존재 • 다른 subclass들에 비하여 잘못 예측되는 subclass의 분포가 고르게 분포	많은 전기/전자 관련 제품에 임피던스 회로망이 존재하기 때문에 상대적으로 다양하게 예측
H01B	케이블; 도체; 절연체; 전도성, 절연 또는 유전체 특성을 위한 재료 선택	• H01B 다음으로 가장 많이 예측하는 H01M subclass에도 재료, 방법 및 수단 관련 subclass 존재 • 잘못 예측되는 subclass의 분포가 고르게 분포	많은 전기/전자 관련 제품에 케이블, 도체, 절연체 등 재료 및 방법, 수단 관련 개념이 많이 존재하기 때문에 분류가 잘되지 않음
H04J	다중 통신 – 전송, 신호, 주파수/시분할, 다중화, 시스템	• H04J 다음으로 가장 많이 예측하는 H04W subclass에도 통신, 네트워크 관련 subclass 존재 • 다른 subclass들에 비하여 H04W, H04B로 월등히 잘못 예측 (분류 키워드 비슷)	많은 전기/전자 관련 제품에 통신, 네트워크, 전송 관련 개념이 존재하기 때문에 상대적으로 다양하게 예측

04 분석 및 결과 해석 Rule 기반 분석

- **탐색 방법 : 비슷한 키워드가 존재하는 subclass 중 더 많은 data를 가지고 있는 subclass 탐색**

레이블	레이블 설명	탐색 결과	분류가 잘되지 않은 원인
H04M	전화통신 - 장치/부품, 자동/반자동 교환기, 통신, 시스템, 제어, 시설/매체	• 장치, 시스템, 제어, 시설 등의 부분에서 비슷한 내용의 subclass 존재 • H04 subclass 중 훈련 데이터 개수가 많은 subclass 존재 • 훈련 데이터의 개수가 많은 H04L, H04M, H04N, H04W로 잘못 예측한 경우가 많음	비슷한 내용의 데이터 수가 더 많은 subclass로 잘못 분류됨
H04B	전송 - 종류별 전송시스템(전파, 전자기파, 음파, ...), 잡음/제한, 감시/시험, 인덱싱	• H04B의 test data 개수는 충분히 많음(1987개) • H04 subclass 중 전송, 무선 키워드 관련 subclass 존재 • 전송, 무선과 연관된 H04L, H04W로 잘못 예측한 경우가 많음	비슷한 내용의 데이터 수가 더 많은 subclass로 잘못 분류됨
H03G	증폭기의 제어 – 임피던스 회로망, 장치, 증폭기/주파수 변환기, 이득/음질/대역폭제어, 음량, 진폭	• H03G의 test data 개수가 토픽의 개수에 비해 적음 • H03 subclass 중 증폭기, 임피던스 회로망 관련 subclass 존재 • 증폭기, 임피던스 회로망과 연관된 H03F로 잘못 예측한 경우가 많음	비슷한 내용의 데이터 수가 더 많은 subclass로 잘못 분류됨

- **탐색 방법 : class의 모든 subclass description 검토**

레이블	레이블 설명	탐색 결과	분류가 잘되지 않은 원인
H03K	펄스 - 정보 전송, 위상차 감지 회로, 자동 제어, 안정화, 동기화, 코딩/디코딩	• 잘 분류된 subclass H03B, H03F, H03L, H03M은 각각 발생, 증폭기, 제어, 암호화로 주제가 명확 • 잘 분류되지 않은 subclass은 주제에 다양한 내용을 포함	명확하지 않은 subclass 분류 정의로 인해 미분류 CPC subclass가 됨
H02N	타류에 속하지 않는 전기 – 발전기/전동기, 열/운동에너지, 흡인력/반발력, 전하	• H02N의 test data 개수가 토픽의 개수에 비해 적음 • H02 subclass 중 잘 분류된 subclass의 주제는 명확함 • 잘 분류되지 않은 subclass은 주제에 다양한 내용을 포함	명확하지 않은 subclass 분류 정의로 인해 미분류 CPC subclass가 됨
H04Q	선택 – 스위치/계전기/셀렉터/무선 통신 네트워크, 선택배치, 다중화, 시스템	• H04Q의 test data 개수가 토픽의 개수에 비해 적음 • H04 subclass 중 잘 분류된 subclass의 주제는 명확함 • 잘 분류되지 않은 subclass은 주제에 다양한 내용을 포함하며 다양한 기기에 대한 포괄적인 분류 주제로 정의됨	포괄적인 subclass 분류 정의로 인해 미분류 CPC subclass가 됨

05 결론

- 기술동향 파악 및 융합패턴의 탐색 중심으로 이루어진 기존의 연구와는 달리 현 CPC 분류체계의 적합성 검증이라는 새로운 측면에서의 접근을 시도함
- 현 CPC 분류 체계가 융합기술을 잘 반영하는지에 대한 적합성을 검증하는 정량적, 정성적 프레임워크를 설계함
- BERT 기반으로 정의한 미분류 특허의 원인 탐색을 통해 CPC subclass 수준에서의 적합성을 입증하고자 함
- 딥러닝 자동 분류 모델에 설명가능한 인공지능을 적용하여 추출한 키워드를 기반으로 다양한 방법론을 이용해 미분류 특허의 특징을 제시함

06 참고 문헌 (1/2)

강희종, 엄미정, 김동명. (2006). 특허분석을 통한 유망융합기술의 예측. 기술혁신연구, 14(3), 93-116.

전상규. (2021). 특허 네트워크 분석을 통한 기술융합 및 융합기술의 확산 연구 —디지털 데이터 처리 기술 중심으로—. 지식재산연구, 16(4), 161-202.

윤민호. (2011) DRAM 산업의지식확산, 기술궤적과산업주도권의이동: 특허인용 네트워크 분석과 신슘페터주의 기술경제학. 지식재산연구, 6(3), 239-270.

권오진, 노경란, 이방래, 고병열, 문영호. (2007). 매개중심성 분석을 통한 기술군간 융합 정도측정. 한국콘텐츠학회 종합학술대회 논문집, 5(1), 1-5.

송영화, 임동현, 김민수.(2019).친환경 융합기술의 기술사업화 전략 : 전기자동차의 특허분석과 비즈니스 생태계 분석을 중심으로.기술혁신학회지,22(5),780-804.

강지호, 김종찬, 이준혁, 박상성, 장동식.(2015).CPC를 이용한 IoT와 Wearables 기술융합 동향분석.한국지능시스템학회 학술발표 논문집,25(1),15-16.

배영임, 신혜리.(2017).인공지능기술의 특허네트워크 분석을 통한 융합패턴 연구.GRI 연구논총,19(1),113-133.

정명석, 정소희, 이주연. (2018). 국내외 특허데이터 기반의 인공지능분야 기술동향 분석. 디지털융복합연구, 16(6), 187-195.

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

김지혜, 김병초. (2017). 특허데이터를 활용한 프로세스관점에서의 인공지능 관련기술 체계 재분류. 한국경영정보학회 2017년 경영정보관련 추계학술대회, 392-400.

김성훈, 김승천 (2021). 클래스 불균형 문제가 있는 특허분류 데이터의 자동분류 성능 개선을 위한 모델 재귀적 오버 샘플링 방법. 전자공학회논문지, 58(4), 43-49.

이관용. (2022). BERT 기반 특허-상품유사군 분류 시스템. 한국교통대학교 일반대학원컴퓨터정보공학과 컴퓨터정보공학 전공 석사학위 논문

정구익. (2020). 클래스 불균형을 극복한 향상된 BERT 기반 특허의 극한 다중 레이블 분류. 숭실대학교 대학원 소프트웨어학과 석사학위 논문

한동희. (2019). 특허 분류의 계층 및 의미 관계를 고려한 다중 분류 기법. 고려대학교 컴퓨터정보통신대학원 빅데이터융합학과 석사학위 논문

김성훈. (2021). 사용자 정의 분류체계에 따른 딥러닝 기반의 특허문서 자동분류. 한성대학교 대학원 스마트융합컨설팅학과 스마트융합제품전공 박사학위 논문

정다운. (2018). 인공지능분야 특허의 기술 파급에 관한 연구. 한성대학교 대학원 스마트융합컨설팅학과 스마트융합컨설팅전공 박사학위 논문

06 참고 문헌 (2/2)

조한슬, 유재흥, 조원영, 신정우.(2021). 산업별 인공지능 융합경쟁력 지수 개발 연구.한국혁신학회지,16(4),243-265.

장청롱, 안현철. (2022). Self-Attention 시각화를 사용한 기계번역 서비스의 번역 오류 요인 설명. 한국IT서비스학회지, 21(2), 85-95.

황상흠, 김도현.(2020).한국어 기술문서 분석을 위한 BERT 기반의 분류모델.한국전자거래학회지,25(1),203-214.

김경외, 이준민, 이창준.(2021).효율적인 기술 정책 제안을 위한 한국 인공지능 지식 구조와 진화 궤적의 탐색적 분석.한국혁신학회지,16(3),139-172.

심우철, 민재옥, 조유정, 고봉수, 노한성.(2020).한국 특허문헌 특성 및 딥러닝 기반 분류모델을 고려한 CPC 자동분류에 관한 연구.한국정보과학회 학술발표논문집,(),406-408.

백현미, 김명숙. (2013). 특허 네트워크 분석을 통한 융합 기술 트렌드 분석: 한국·미국·유럽·일본의 특허데이터를 중심으로. 벤처창업연구, 8(2), 11-19.

유준상, 이희상.(2013).특허에 기반을 둔 기술융합 분석.대한산업공학회 추계학술대회 논문집,(),1105-1133.

심재륜.(2018).전자상거래(G06Q) 분야에서 '사물인터넷' 기술의 CPC 코드 기반 기술 융복합 분석.한국정보전자통신기술학회 논문지,11(6),678-683.

정병기, 김정욱, 윤장혁.(2016).융합기술의 동향분석을 위한 의미론적 특허분석 접근 방법.지식재산연구,11(4),211-240.

김지은, 이성주.(2013).특허정보를 활용한 산업융합성 평가 방법론.대한산업공학회지,39(3),212-221.

이경석, 임규민, 조호묵.(2022).XAI 기반 도박사이트 분류 모델 분석을 통한 주요 키워드 탐색.한국정보과학회 학술발표논문집,(),1291-1293.

Goštautaite, D.; ˙ Sakalauskas, L. (2022).Multi-Label Classification and Explanation Methods for Students' Learning Style Prediction and Interpretation. Appl. Sci. 5396.