연구 개요

- ## 연구 배경
  - 새로운 혁신을 위해 융합기술을 효과적으로 파악하고 분석하는 것이 중요한 활동으로 간주되어 왔음
  - 융합기술의 효과적 분석을 위해 특허분석이 널리 활용되어 왔으며, 융합연구를 위한 다수의 특허분석 연구가 수행되어 왔음 (전상규. 2021)
- ## 연구 동기
  - 대부분의 특허기반 기술융합 연구는 특허 CPC 분류체계를 바탕으로 융합현상을 분석하는 데 초점을 맞추고 있음
  - 현 CPC 분류체계가 융합기술 분류에 적합한가라는 근본적 질문을 다룬 연구는 거의 없음 (강희종,엄미정,김동명. 2006, 배영임,신혜리. 2017, 정명석,정소희,이주연. 2018, 백서인,이현진,김희태. 2020)
- ## 연구 목적
  - 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

연구 현황

- **IPC, CPC 특허 체계를 이용한 기술융합 분석 관련 연구**
  - CPC를 활용하여 연관규칙분석 기반의 사물인터넷과 웨어러블 기술융합동향을 분석함 (강지호, 김종찬, 이준혁, 박상성, 장동식. 2015)
- **현 CPC 특허 체계의 재정립, 제안 관련 연구**
  - 인공지능 기술을 대상으로 프로세스적 관점에서의 분류체계를 제시하며 특허 분석에 적용함 (김지혜, 김병초. 2017)
  - 사용자 분류체계에 따라 특허문헌을 자동으로 분류하는 분류 모델 및 분류기 아키텍처를 설계함 (김성훈. 2021)
- **특허 자동 분류에 딥러닝 모델을 적용한 연구**
  - 다중 레이블을 가지며 클래스에 따른 분포가 매우 불균형한 특허의 특성을 고려한 분류 모델 생성을 위해 BERT 기반의 향상된 극한 다중 레이블 분류 모델을 제안함 (정구익. 2020)
- **텍스트 분류 모델에 XAI를 적용한 연구**
  - 도박사이트 분류 모델에 XAI 기법을 적용하여 주요 키워드를 탐색하는 방법을 제안함 (이경석, 임규민, 조호묵. 2022)

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

명사 빈도 분석

CPC 예측에 높은 영향을 준 키워드 비교

유사도 기반 분석

4차산업 키워드와 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

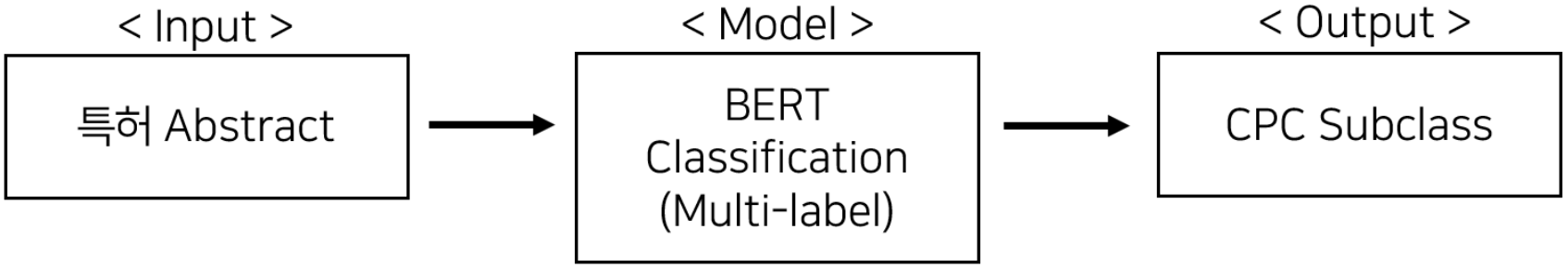
데이터 수집 및 전처리

- ## 수집 데이터 정보
  - 사이트 : USPTO PatentsView
  - 수집기간 : 등록일자 기준 20.01.01 ~ 21.12.31 (2개년)
- ## 데이터 수집
  - 수집 section 선정
    - 목적 : 융합기술을 많이 포함하는 section 선택
    - 방법 : 신기술 분류를 다루는 Y section 특허의 다중 분류된 다른 subclass 분석
    - 선정 section : H section
  - H section 데이터 수집
    - 수집 개수 : 242,844개
    - 수집 요소 : Patent Number, Patent Title, Patent Grant Date, Patent Type, CPC class, subclass, group, subgroup ID, abstract
- ## 데이터 전처리
  - 레이블 one-hot encoding(총 52가지 CPC subclass)
  - Abstract 전처리 : 특수문자 제거, 불용어 제거, 토큰화, 명사 추출, 명사 원형 복원

BERT Classification

BERT 모델을 통해 분류 특허와 미분류 특허를 구분하여 현재 CPC 분류체계로 분류되지 않는 특허를 파악함

- ## BERT Classification Fine Tuning



- 모델 : BERT_base(L=12, H=768, A=12, Total Parameter=110M)
  (L: layer 개수, H: hidden layer 크기, A: self-attention의 head 개수)
- 파라미터 : Raw Data, learning rate : 1e-5, weight decay : 0.005, epoch : 6

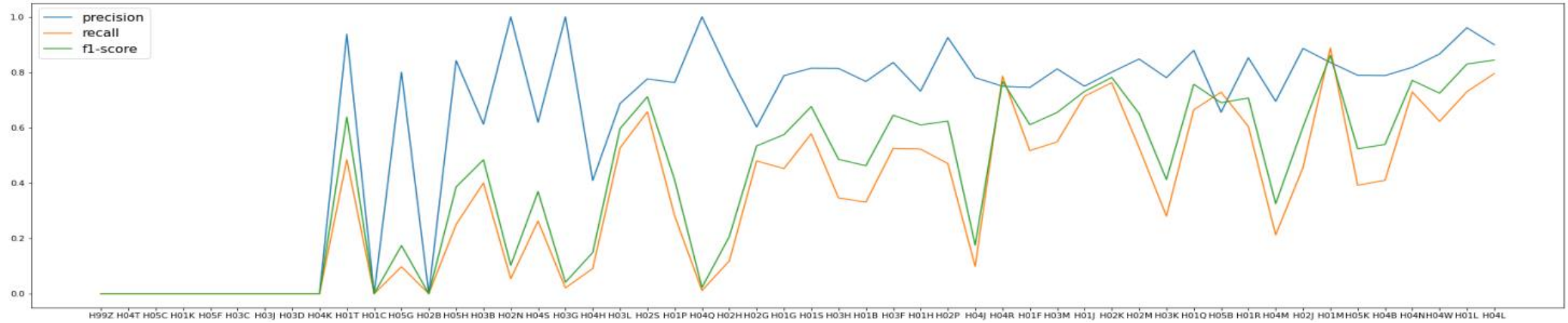
- ## 특허 탐색 기준 설정

- 테스트 데이터 24,285개 적용
- 분류 특허와 미분류 특허 구분
  - 구분 기준 : 특허의 multi-label true값과 pred값의 일치 여부
    - ➢ 분류 특허 : 모델이 subclass 모두 정확히 예측한 특허 13,174개
    - ➢ 미분류 특허 : 모델의 예측이 하나라도 틀린 subclass가 존재하는 특허 11,111개
- 성능에 따른 CPC subclass 구분

BERT Classification

- CPC subclass 별 성능 그래프

X축: 해당 CPC subclass에 포함되는 특허의 개수를 기준으로 오름차순 정렬



- 미분류 CPC subclass 정의

데이터 개수 임계값 : 데이터 개수의 0.01% (24.3개)

- CPC subclass의 데이터 개수가 임계값 미만 : H05C, H04T, H99Z --- (3)
- CPC subclass의 데이터 개수가 임계값 이상
  - 전혀 예측을 하지 못하는 CPC subclass (f1-score와 recall 성능이 0)
    H01K, H05F, H03C, H03J, H03D, H04K, H01C, H02B --- (8)
  - 그래프 주변 CPC subclass에 비해 낮은 성능 (f1-score와 recall 기준)
    H02N, H03G, H04Q, H03H, H01B, H04J, H03K, H04M, H05K, H04B --- (10)

명사 빈도 분석

분류 특허 vs 미분류 특허

분류 특허, 미분류 특허 각각 10%씩 test dataset 레이블 비율대로 랜덤샘플링을 진행하여 MultiLabelClassificationExplainer(XAI) 적용

→ 특허 분류에 영향을 주는 키워드 파악 및 특성 추출

• 미분류 특허 / 분류 특허 키워드 빈도 분석 결과

<미분류 특허>

* 빈도수 기반으로 그룹 세분화

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

<분류 특허>

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

- XAI 키워드에 비해 명사 추출 키워드가 상대적으로 전기 관련 키워드가 적게 등장
- XAI 키워드 비교 결과, 분류 특허에 비해 미분류 특허에 '전기' 관련 키워드 출현 빈도 높음
- 융합기술의 포함 정도를 판단하는 기준이 주관적 → 정량적인 판단 방법 필요

유사도 기반 분석

분류 특허 vs 미분류 특허

4차산업혁명 관련 기술분야 특허분류 체계 키워드를 번역하여 융합기술 키워드로 정의

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

- **Pretrained FastText 결과**
  - keyword to keyword 방식으로 Pretrained FastText 유사도 계산

<분류 특허>

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**

<미분류 특허>

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**

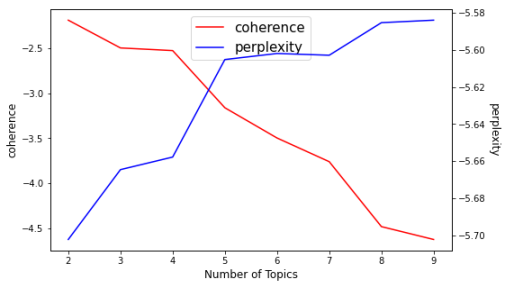
- 명사추출 결과로 유사도를 계산한 결과, 분류 특허와 미분류 특허의 차이가 나타나지 않음
- 분류에 있어 Explainability한 키워드로 구성된 XAI 딕셔너리로 계산한 결과, 미분류 특허 키워드의 융합기술 유사도가 0.02 정도 더 높게 나타남
- 미분류 특허들의 융합기술 관련도가 더 높다고 볼 수 있음

토픽모델링

• 토픽모델링 적용 방법

- CPC subclass에 적합하지 않은 토픽의 존재 여부 검토
- CPC subclass 각각 Perplexity와 Coherence를 고려하여 적정 passes와 토픽 개수 선정

<H03K 예시>



Topic 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

• CPC subclass 별 토픽모델링 결과 표

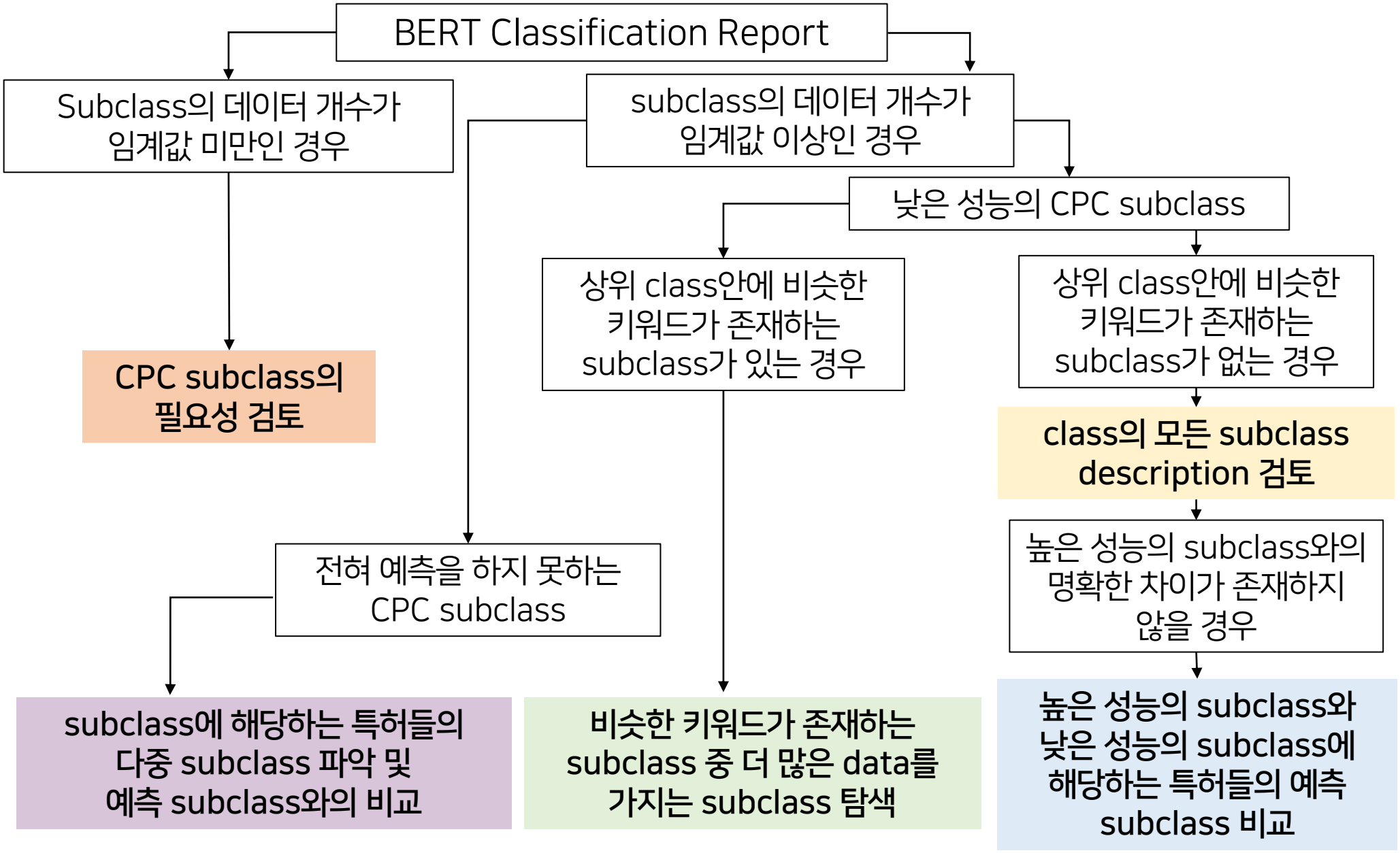
CPC	topic1	topic2	topic3	topic4	topic5	topic6
H03K	Pulse 증폭기술	반도체 스위칭 이용한 pulse기술	센서와 스위치를 이용한 출력 값 제어	컴퓨터 논리회로 적용 방법		
H04M	전화통신 전반	스위칭 센터	전화시스템	전화 통신할 때 네트워크 측면에서의 절차	가정집에 설치하는 전화통신	
H04B	방송국에서 전송시스템 전송 방법	전송시스템 원리	광 전송 시스템	무선 전송 시스템		
H05K	수신기, 수상기와의 조합법	전기장치 구성	인쇄회로 구성	전자장치에서의 인쇄회로	전기장치	인쇄회로 장치 및 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			

- 각 토픽 별 기여도가 높은 키워드 10개를 살펴본 결과, 해당 CPC subclass와 연관되어 있는 단어들이라고 판단

- 해당 CPC subclass와 연관이 없다고 판단되는 토픽은 존재하지 않음

Rule 기반 분석

미분류 CPC subclass 탐색



Rule 기반 분석

- **subclass에 해당하는 특허들의 다중 subclass 파악 및 예측 subclass와의 비교**
  - 대상 Subclass : H02B, H01C, H04K, H03D, H03J, H03C, H05F, H01K
  - 미분류 원인: 연관된 다중 CPC subclass 중 데이터 수가 많은 subclass로 분류되는 경향
- **높은 성능의 subclass와 낮은 성능의 subclass에 해당하는 특허들의 예측 subclass 비교**
  - 대상 Subclass : H05K, H03H, H01B, H04J
  - 미분류 원인 : 모든 전기/전자 관련 제품에 subclass description 내용이 포함되어 있어 상대적으로 다양하게 예측
- **비슷한 키워드가 존재하는 subclass 중 더 많은 data를 가지는 subclass 탐색**
  - 대상 Subclass : H04M, H04B, H03G
  - 미분류 원인 : 비슷한 내용의 데이터 수가 더 많은 subclass로 잘못 분류됨
- **CPC subclass의 필요성 검토**
  - 대상 Subclass : H05C, H04T, H99Z
  - 미분류 원인 : 전체적인 데이터 부족으로 분류 모델에서 학습이 되지 않음
- **class의 모든 subclass description 검토**
  - 대상 Subclass : H03K, H02N, H04Q
  - 미분류 원인 : 명확하지 않은 subclass 분류 정의

결 론

- 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