뉴스미디어 분류 프로젝트 모델링 레포트 (작성일: 2021-03-09)

대신경제연구소
Daishin Economic Research Institute

1.프로젝트 목표

- 미디어를 이용한 ESG 산업, 기업, 주제 자동화를 위한 AI 모델링

2. 진행상황

- **데이터 수집 및 라벨링**

- **데이터 현황**

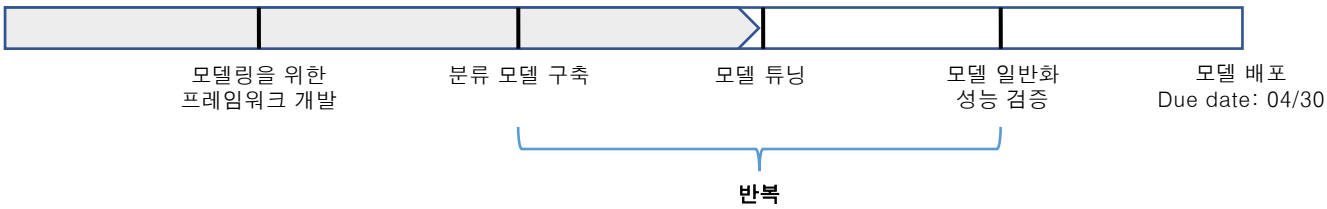
Positive/Negative		
	Negative(0)	Positive(1)
Train	8562	9074
Valid	2233	2177
Test	2763	2749

Category				
기타	38945		자원순환	571
사회공헌	5677		계열회사와의 거래	496
에너지 및 온실가스	5211		지역사회 영향	382
소유구조	3926		감사기구	303
주주가치 보호	1974		이익배분	287
제품/서비스 책임	1599		친환경제품/서비스 개발	214
공급망 관리	1287		환경관리 체계	211
인적자원관리	1278		인권	101
사업장 안전 및 보건	1239		대기오염	99
이사회 구성과 운영	971		공시	78
내부통제/투명성	935		수질오염	42
소비자 보호	806		생물다양성	26
공정거래	610		화학물질 배출	11

Ref. Red letters: 데이터 부족할 시 라벨 통합 고려 중

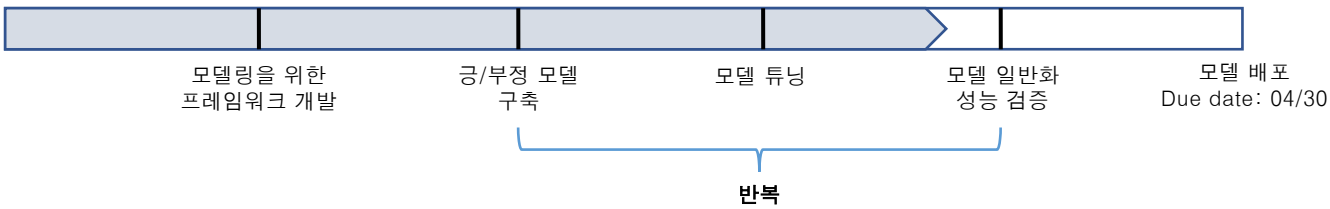
- **분류 모델**

정의: 뉴스 미디어 데이터를 Category에 맞게 분류하는 AI 모델



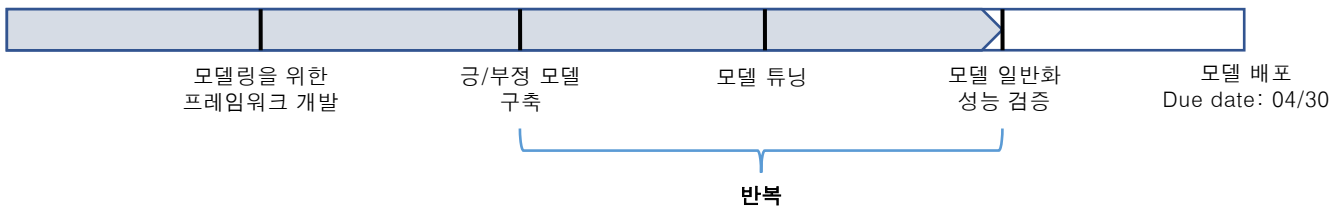
- **긍정/부정 모델**

정의: 뉴스 미디어 데이터의 긍정/부정을 예측하는 AI 모델



- **기업명 추출기**

정의: 뉴스 미디어 데이터에서 Coverage 기업이름을 추출하는 AI 모델



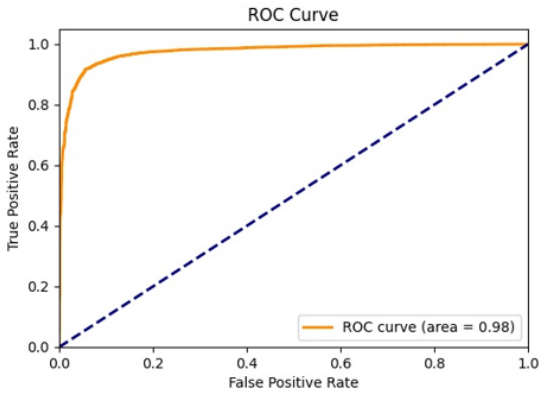
3. 모델 성능

긍정/부정 모델링 결과

Indicator	
f beta	0.960326
AUC	0.975276
F1 score	0.960718
accuracy	0.939449

Confusion Matrix

		모델 예측	
		부정	긍정
실제	부정	2087 (0.87)	323 (0.13)
	긍정	312 (0.04)	7765 (0.96)

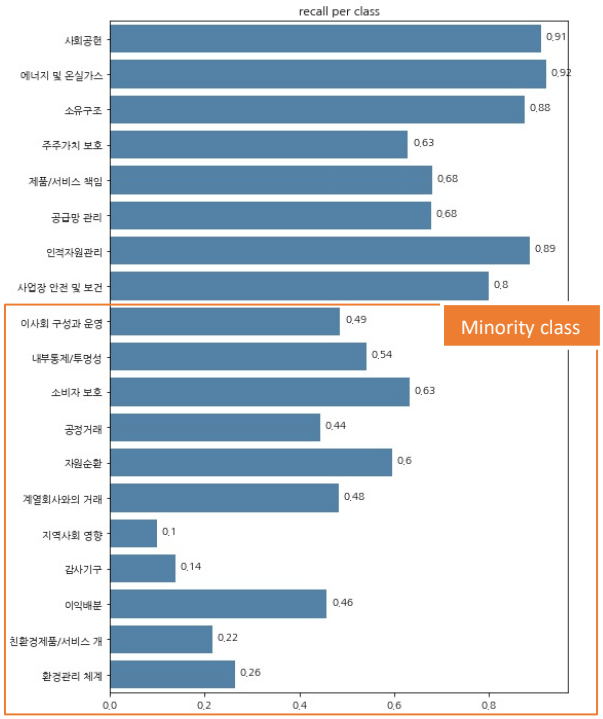


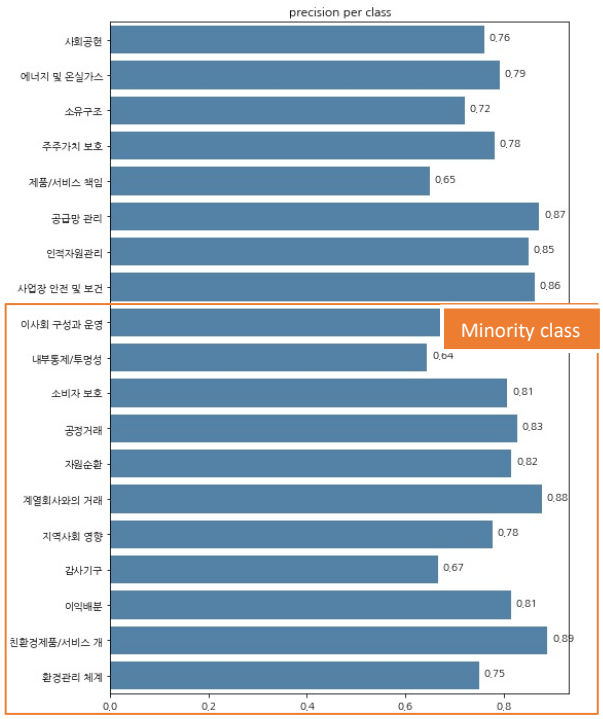
카테고리 모델링 결과

Indicator

	Precision	Recall	F1-score	Sample
사회공헌	0.76	0.91	0.83	1068
에너지 및 온실가스	0.79	0.92	0.85	960
소유구조	0.72	0.88	0.79	748
주주가치 보호	0.78	0.63	0.70	359
제품/서비스 책임	0.65	0.68	0.66	303
공급망관리	0.87	0.68	0.76	242
인적자원관리	0.85	0.89	0.87	238
사업장 안전 및 보건	0.86	0.80	0.83	229
이사회 구성과 운영	0.71	0.49	0.58	185
내부통제/투명성	0.64	0.54	0.59	177
소비자 보호	0.81	0.63	0.71	139
공정거래	0.83	0.44	0.58	108
자원순환	0.82	0.60	0.69	104
계열회사와의 거래	0.88	0.48	0.62	89
지역사회 영향	0.78	0.10	0.18	70
감사기구	0.67	0.14	0.23	58
이익배분	0.81	0.46	0.59	48
친환경제품/서비스개발	0.89	0.22	0.35	37
환경관리체계	0.75	0.26	0.39	34
토양오염	0	0	0	0
ESG거버넌스	0	0	0	0
Accuracy			0.77	
Macro avg			0.62	

Ref. Blue label: Minority class





3. 모델 성능

기업명 추출기 모델링

Change log 작업	전체 데이터 : 105801 언매칭 카운트	날짜
최초버전	29935	02월 17일
한자,특수문자,㈜,공백 제거	28801	02월 25일
한자,특수문자,㈜,공백 제거 + 2차 형태소 분리	20282	02월 25일

컨텐츠	기업명 추출 결과
KB금융그룹(회장 윤종규)은 지난달 29일 오후 여의도 본점에서 윤종규 회장과 대한카누연맹 김용빈 회장 등이 참석한 가운데, 카누 남녀 국가대표팀에 대한 후원 ...	KB금융그룹,대한카누연맹
최선목 한화그룹 커뮤니케이션팀장(62·사진)이 1일 부사장에서 사장으로 승진해 새로 출범하는 그룹 커뮤니케이션위원회 위원장을 맡았다. 한화그룹은 지난 5월 ...	한화그룹,그룹커뮤니케이션위원회
베트남 생명보험 및 손해보험시장 규모는 우리나라의 2.0%, 2.4%에 불과하나, 연 평균 보험료 실질성장률이 15.%, 7.3%로 높은 수준으로 매년 가파른 성장세를 보이고 ...	베트남
손해보험사는 건강연령 연동 보험 상품을 하반기에 내놓을 계획이다. DB손보 관계자는 "보험개발원 산출 모델이 감독당국의 사전 승인을 받으면 적극적으로 관련 상품 ...	DB손보,보험개발원
[한국금융신문 장호성 기자] 리치앤코(대표 한승표)는 통합 보험관리 플랫폼 굿리치 스마트폰 어플리케이션이 지난 1일 업계 최초로 100만 다운로드를 달성했다고 ...	한국금융신문,리치앤코
SK네트웍스는 최신원〈사진〉 회장이 지난달 30일 아시아태평양 국제경영학회(APAIB)와 국제연합(UN)이 공동으로 주관한 '2018 APAIB-UN 조인트 콘퍼런스'에서 ...	UN,국제연합,SK네트웍스는
롯데마트는 오는 31일까지 전점과 온라인몰인 롯데마트몰에서 장마철 대비용품과 먹거리, 패션잡화를 할인 판매하는 '장마철 철벽방어 기획전'을 진행한다고 2일 ...	롯데마트
홍요섭 대표는 에어로바이런먼트(나스닥 AVAV) R&D 센터, (주)세방전지 R&D 센터와 함께 배터리 교체시스템을 대체할 수 있는 스마트 급속충전솔루션을 연구 ...	나스닥AVAV,세방전
신한금융 관계자는 "신한서브는 은행·지주의 자회사나 계열회사가 아니다"라며 "The Bank 신한동우회가 100% 지분을 가진 회사이기 때문에 관련 이사회에서 (왕 전 ...	신한금융,TheBank신한동우회,신한서브

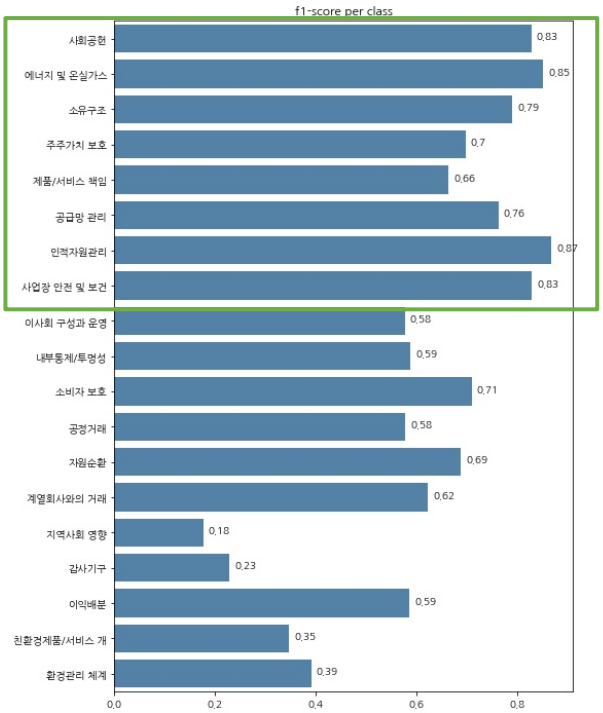
4. 추 후 계획

긍정/부정 모델링

- Precision, Recall 중 Priority가 높은 것을 선정 후 Threshold moving method 적용하여 선정된 Indicator 성능 확보
- 크롤러 인수(From, QESG) 완료 후 일반화 성능 검증 진행
- 추가적인 Fine tunning 진행

분류 모델링

- Precision, recall, f1-score 값은 데이터의 샘플 수, miss-labeling이 가장 큰 영향을 줌
  - Indicator를 구하는 Equation으로 가장 큰 영향을 주는 요소를 파악가능



- 해당 class들은 샘플 수와 성능을 고려 하였을 때 Majority class로 분류 가능하나, 샘플 수 대비 성능의 상관관계가 낮아 ***Miss-labeling의 영향을 파악해야 함.***
- 따라서 현재 바로 작업 가능한 Miss-labeling문제를 해결 한 후, 데이터 샘플 수와 성능의 관계를 고려하여 Majority class와 Minority class를 재정의 후, 모델링을 진행할 계획.

기업명 분류기

- 현재 뉴스의 제목, 요약본에서 나오는 기업명의 대부분을 찾고 있음으로, 기업명 synonym dictionary list를 늘리는 작업 진행
- 분류기는 기업명의 synonym dictionary list와의 매칭 여부로 기업명을 판단하는 구조로 되어있음으로, list가 많아질 수록 찾아내는 기업명도 늘어남.
  - ***즉, synonym dictionary list가 늘어날 수록 Coverage 기업이 늘어남***

5. 대신경제연구소 뉴스미디어 AI VS 지속가능발전소 특허(거절)

뉴스 수집부

지속가능발전연구소	대신경제연구소
기사들 간의 유사도 분석을 통해 유사도가 기준치 이상인 뉴스 기사들에 대한 클러스터링 수행	개발 진행중
TF-IDF로 Vectorizing을 진행하고 Cosine유사도를 이용하는 LexRank 방식 채택 • 희귀 단어에 대응 불가 • Misspelled 상황을 무시함	Byte-pair Encoding(BPE)을 이용하여 Vectorizing 진행 • Character를 기본 subword units로 보고 Looping 과정을 통해 빈도수가 가장 많은 biagram을 찾는 algorithm. • 희귀 단어(자주 등장하지 않은)에 강하다. • Misspelled 상황에서도 Vectorizing이 잘된다.

뉴스 분류부

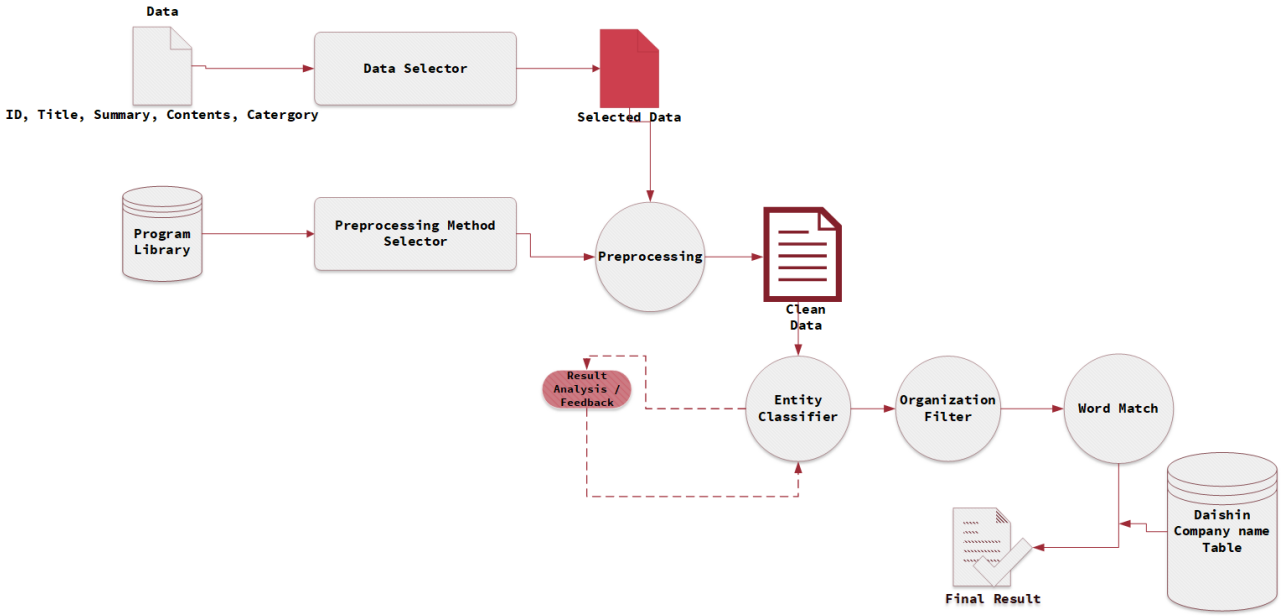
지속가능발전연구소			대신경제연구소		
	설명	Algorithm		설명	Algorithm
이원 분류	수집된 뉴스가 ESG 기업평가에 사용될 수 있는 데이터인지 판단.	Multinomial Bayes, Bernoulli Bayes, Linear SVC (White box model)	이원 분류	수집된 뉴스가 ESG 기업평가에 사용될 수 있는 데이터인지 판단.	Deep Learn-ing model (Black box model)
ESG 분류	환경,사회, 지배구조의 세가지 이슈 중 하나로 분류..		ESG 카테고리 분류	환경, 사회, 지배구조의 27가지 이슈 중 하나로 분류. (대신경제연구소 ESG rating model 미디어 지표).	
카테고리 분류	환경, 사회, 지배구조내에 있는 각각의 항목에 맞게 분류.				

평가 결과 산출 부

지속가능발전연구소		대신경제연구소	
	설명		설명
점수 산출	ESG,관련 리스크, 기업 리스크 및 기타 관련 문제에 대한 리스크를 각각 상이한 방식으로 계산한 후 이를 토대로 ESG 기업 평가점수를 산출 Ref. 결과 Score equation. $\text{Consequence Score} = \max\{5, \frac{1}{2}(ESGrisk + CompanyRisk + Relevance)\}$	ESG Rating model	지배구조연구소의 ESG Rating model.

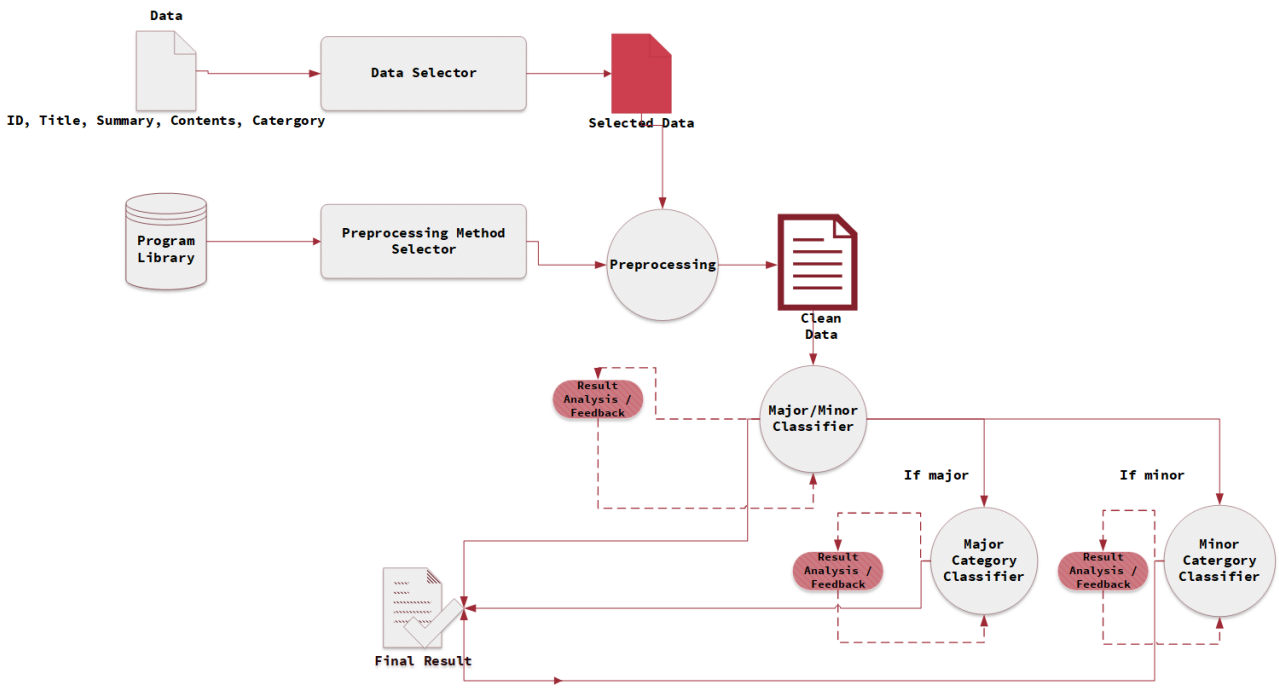
5. Appendix

기업명 분류기 시스템 구성도



1. 조건부 랜덤필드를 이용하여 사전 훈련된 전이확률 및 상태확률로 문장을 분리.
2. 분리된 단어에 대해서, 명사인 단어들만 필터링 후, 남겨진 단어들에 대해서 one hot-encoding 진행
3. One hot-encoding으로 표현된 문장을 Deep learning, Tree 등의 Machine Learning 분류 모델에 입력으로 사용하여 훈련
4. 훈련된 모델을 이용하여 Evaluation value 추론

카테고리 분류기 시스템 구성도



1. 문장을 단어로 나누고 저차원으로 압축 후 복원하는 과정에서 가장 손실이 없는 임베딩 구조를 찾음.
2. 임베딩 구조를 이용하여 양방향 LSTM 모델의 입력 전단계에 적용
3. 각각의 문장의 상태를 Vectorize하고, Vectorized된 정보를 단어의 품사로 복호화
4. 품사 중 기관명인 단어를 예비 기업명 후보로 설정 후, Synonym Dictionary list와의 일치 시 최종 결과로 선정.