

언론 인식 제고를 위한 양질의 뉴스 기사 판별 AI 모델

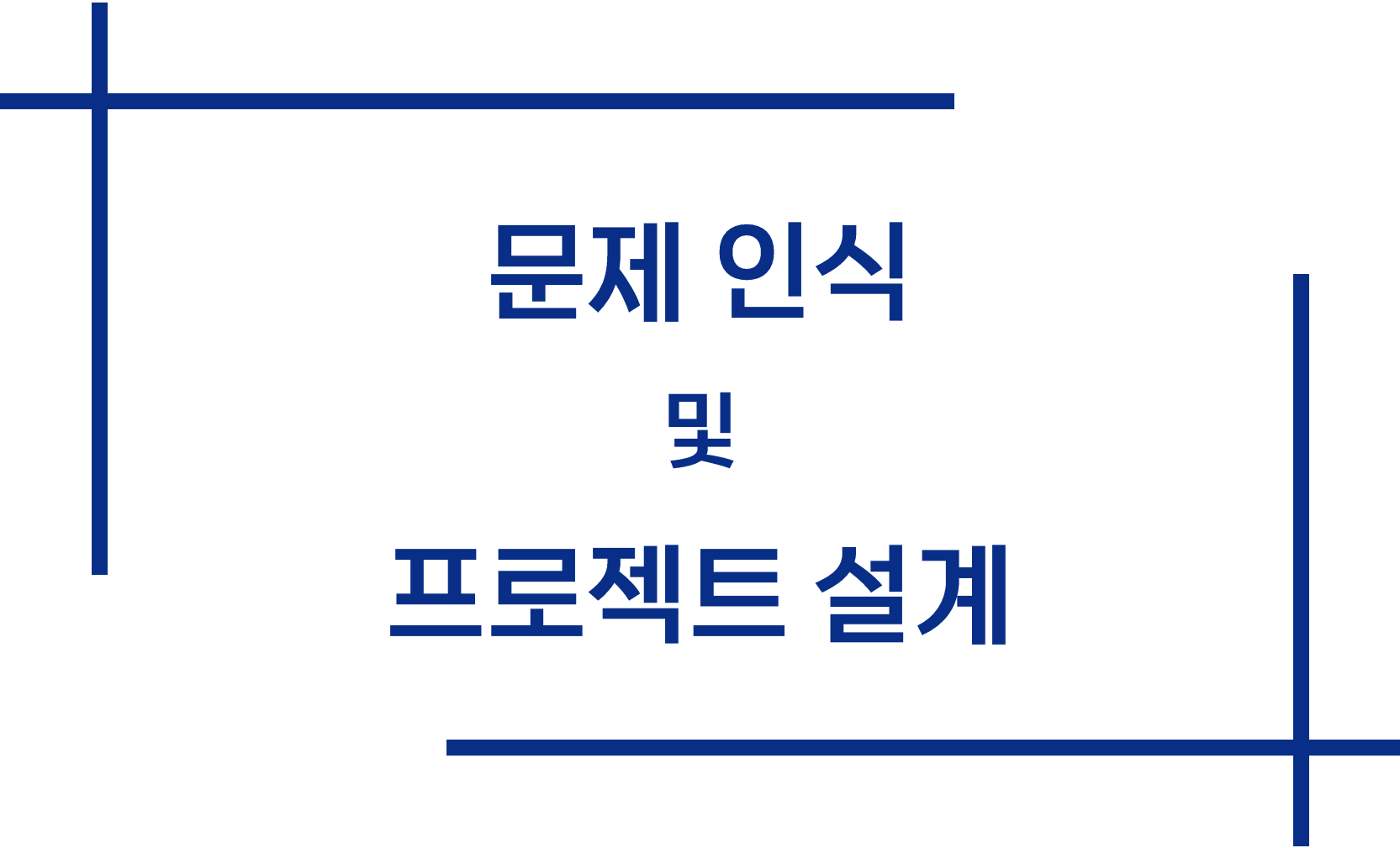


Catholic University of Korea
2023 Deep Learning

202121638 안희수
202121604 김래이
201620565 박수빈
202121633 송지윤
202121656 장서연

Contents

- 01 문제 인식 및 프로젝트 설계
- 02 데이터 전처리
- 03 모델 학습 및 결과
- 04 셀프 피드백



문제 인식 및 프로젝트 설계

01

문제 인식 및 프로젝트 설계

- Task 1 : 카피 기사 파악
 - 보도자료의 무분별한 복제

보도자료
 보건복지부
간호법안 재의요구하기로 제20회 국무회의 의결
· 등록일 : 2023-05-16 [최종수정일 : 2023-05-18] · 조회수 : 1124 · 담당자 : 장태영 · 담당부서 : 보건 의료정책과
간호법안 재의요구하기로 제20회 국무회의 의결 - 보건복지부 장관, 의료·요양·돌봄 통합 시스템 구축 등 정책방향 설명 -

 메디포뉴스 政, 간호법안 재의 요구 결정...국무회의 의결
 빛가람뉴스 간호법안 재의요구하기로 제20회 국무회의 의결
 서울Pn 간호법안 재의요구하기로 제20회 국무회의 의결
 탐사일보 간호법안 재의요구하기로 제20회 국무회의 의결

그림1) 보도자료의 무분별한 복제 예시

하나의 보도자료를 다수의 신문들이 카피하여 다수의 기사를 작성



반복적인 기사에 대한 피로감 유발

문제 인식 및 프로젝트 설계

- Task 2 : 취재 포함 기사 파악
 - 단순 부정적 사실 나열로 끝나는 뉴스 기사

의 국회법 절차를 통해 심의·의결됐다. 지난 2월 민주당이 국회 법제사법위원회를 건너뛰어 간호법을 본회의에 직회부하고 지난달 27일 본회의에서 통과시키면서 직역 간 갈등은 더욱 고조됐다.

그런데 야당과 대통령이 대화나 타협 없이 입법권과 거부권을 쓰면서도 정작 간호법에 영향을 받는 **당사자들의 목소리는 없었다. 여야는 정치적 타협과 대화 대신, 각자 편가르기에 바빴다.**

법안심사 과정에서 주무부처인 보건복지부는 간호사, 의사, 간호조무사 등 당사자들의 이해관계 조정에는 **손을 놓고 갈등을 키웠다.**

그림2) 단순 부정적 사실 나열 기사 예시

취재가 포함된 기사의 경우, 결론에서 이를 한 번 더 언급하는 방식으로 마무리

➡ 기사의 **결론** 부분을 통해 **취재 여부 판별**

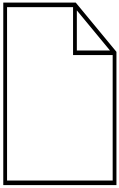


데이터 전처리

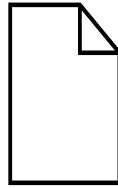
02

데이터 전처리

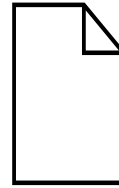
- Task 1 : 카피 기사 파악
 - 크롤링 기사 데이터셋

 { 해당 키워드
뉴스 기사
174개
보도자료

간호법 재의

 { 해당 키워드
뉴스 기사
313개
보도자료

고교학점제

 { 해당 키워드
뉴스 기사
123개
보도자료

후쿠시마 원전

➔ 크롤링을 통해 약 **600개**의 데이터 수집

경향신문, 동아일보, 조선일보, 중앙일보, 한겨레, KBS, SBS, MBC

데이터 전처리

- Task 1 : 카피 기사 파악
 - 크롤링 기사 데이터셋

0//간호법안 재의요구하기로 제20회 국무회의 의결//정부는 제20회 국무회의에서 간호법안에 대한 의료요양돌봄시스템을 구축한다. ④수요자 중심의 통합적인 돌봄체계 구축, ⑤지역 간의 합리적 있는 기반을 마련할 계획이다. 둘째, 거주지역에서 골든타임 내에 중증응급, 분만, 소아진료를 받도 셋째, 간호사의 처우개선은 국가가 책임진다. '간호인력지원 종합대책(23.4월)'을 충실히 이행하며, 탕으로 보건의료인력들이 각자의 역량을 발휘하고 국민 건강에 이바지하도록「보건의료인력 종합계 장을 직접 찾아가 어려움을 함께 느끼며 필요한 정책을 추진해 나갈 계획이다. 마지막으로, 정부는 : 겠다고 밝혔다. 이날 정부는 제20회 국무회의에서 간호법안에 대해 헌법 제53조 제2항에 따라 무회의에서 지난 5월 4일 정부로 이송된 간호법안에 대해 「헌법」제53조 제2항에 의거하여 국회 시스템을 구축하겠습니다. 고령화에 따라 의료와 돌봄 수요는 변화하고 있습니다. 정부는 ④수요자 중 곳에서 돌봄 서비스를 받으실 수 있는 기반을 마련하겠습니다. 본질적인 서비스를 효율화하고, 산재 로운 모델을 마련하며, 의료법, 건강보험법, 장기요양보험법, 노인복지법 등 관련 법령을 정비해 제5 민과 현장인력, 보건복지 분야 전문가의 다양한 의견을 듣고 사회적 논의가 바탕이 된 협업체계를 도 게 개선대책', 3월 '응급의료 기본계획'을 순차적으로 발표한 바 있습니다. 국민이 거주하는 지역에서 위해 노력하겠습니다. 셋째, 간호사 처우개선은 국가가 책임지겠습니다. 정부는 지난 4월 간호사의 : 된 과제를 충실히 이행하며, 간호사가 우수한 전문 의료인으로 성장할 수 있도록 지원하겠습니다. 지 과를 발표하였습니다. 이를 바탕으로 보건의료인력들이 각자의 역량을 발휘하고 국민 건강에 이바지 하고 사회적 논의를 거쳐 마련될 것입니다. 다양한 가치와 이해관계가 공존하는 보건의료 제도를 구 사 여러분들은 지난 100년 동안 환자의 곁을 지켜오셨습니다. 앞으로도 환자의 곁을 계속 지켜주실

1//민주 '처럼회', 대통령거부권 제한법까지 됐다... 당내서도 "위헌"//김용민 더불어민주당 의원./뉴스 고 있다. 국회법을 개정해 대통령의 거부권을 제한하겠다는 것이어서 당장 민주당 안에서도 "위헌(違 가법고 정략적"이라는 비판도 나온다.변호사 출신으로 처럼회 소속인 김 의원은 지난 17일 대통령이 거부권 행사를 막기 외하 건 아니냐는 해석이 나왔다. 화우하 민형배 처간호 이원 득 다른 처럼회

그림3) 간호법 재의 보도자료 및 크롤링 기사 데이터 전처리 예시

➡ '번호//기사제목//기사내용' 형태의 txt파일로 저장

단, 0번은 보도 자료

02

데이터 전처리

- Task 2 : 취재 포함 기사 파악
 - Train dataset : (국내 논문) 문장 의미 태깅 데이터셋 中 '연구 결과'
크롤링 기사 데이터셋 (KBS, MBC, 조선일보, 동아일보)

초기) 문제정의, 가설설정, 기술정의, 대상데이터, 분석방법, 제안방법, 이론/모형, 성능/효과
후속연구/제안

현재) 연구 목적 : 문제정의, 가설설정, 기술정의
연구 방법 : 대상데이터, 분석방법, 제안방법, 이론/모형
연구 결과 : 성능/효과, 후속연구/제안

- '후속연구/제안' 데이터가 개수의 문제로 편향성 존재
특정 단어를 포함한 예시가 대부분
연구 결과는 하위 개념인 '후속연구/제안'을 포함할 것이라 가정

데이터 전처리

- Task 2 : 취재 포함 기사 파악
 - Test dataset : 크롤링 기사 데이터셋(한겨레, 중앙일보)

1	문장	후속연구
2	협회 회장은 개회사에서 “국민의힘과 보건복지부는 간호법 반대 단체들의 일방적 주장만을 수용해	0
3	쟁을 멈추지 않을 것”이라고 주장했다.간협은 투쟁 방안의 하나로 간호법 제정에 반대했던 정치권 연	0
4	혀 사실이 아니다”라고 말했다.교수들은 학생 참여 여부를 확인하는 것 자체가 학생들의 집회 참여를	0
5	처를 검토할 것”이라고 말했다. 백찬기 간협 홍보국장은 “당장 대규모 준법투쟁 참여가 이뤄지지 않	0
6	구에 얼마나 부합할지, 높아지는 지역사회 돌봄 수요엔 어떻게 대처할지, 지역 간의 모호한 업무 범	1
7	게 이뤄져야 한다. 하지만 윤 대통령은 거부권 행사를 정부 수반으로서의 책임이 아닌, 여소야대 국민	1
8	는 물론, 협상력을 발휘하지 못한 채 대통령에게 거부권 행사만 건의하는 집권 여당의 무능함을 지적	0
9	서도 연말까지 업무 지침과 지원 대책을 마련하기로 했다. 1만여명 정도로 추산되는 피에이 간호사를	1
10	제정안에 거부권을 행사한 윤석열 대통령을 비판했다.윤석열 대통령이 간호법 제정안에 재의요구권	0
11	뒀한다. 국민의힘은 14일 한덕수 국무총리, 조 장관 등이 참여한 당정협의회를 연 뒤 보도자료를 내	1
12	취소법 반대 릴레이 단식투쟁을 하는 의협 회원들. 연합뉴스이런 가운데 여야는 강 대 강 대치를 이	1
13	의료노동자 실태조사(신뢰수준 95%, 표본오차 ±0.32%포인트)를 보면, 간호사 응답자 44.9%가 의사	0
14	의견만 반영한다는 지적을 받아왔다. 양쪽이 첨예하게 맞붙는 사안에서 일방의 이해만 대변하는 것	1
15	이 간호법 공포를 위해 적극적인 단체 행동이 필요하다”고 답했다. 간호협회 관계자는 “대통령이 거부	1

그림4) 기사 후속연구 라벨링 예시

추가적으로 일부 기사를 직접 라벨링해 사용(총 104개)

0: 후속연구 제안하지 않음 / 1: 후속연구 제안



모델 학습

및

결과

모델 학습 및 결과

- Task 1 : 카피 기사 파악
 - Doc2Vec & ko-sentence-transformers

Doc2Vec

```
from gensim.models.doc2vec import Doc2Vec, TaggedDocument
from tqdm import tqdm

model = Doc2Vec(tagged_corpus_list)
print(f"Tag Size: {len(model.docvecs)}, end=' / ')

# Doc2Vec 학습
model.train(tagged_corpus_list, total_examples=len(tagged_corpus_list), epochs=10)

# 유사한 문장 찾기
similar_docs = model.docvecs.most_similar(0, topn=10)
print(similar_docs)
```

```
similar_doc = model.docvecs.most_similar('간호법안 재의요구하기로 제20회 국무회의 의결')
print(similar_doc)
```

모델 학습 및 결과

- Task 1 : 카피 기사 파악
 - Doc2Vec & ko-sentence-transformers

Doc2Vec 코드 결과

```
similar_doc = model.docvecs.most_similar('간호법안 재의요구하기로 제20회 국무회의 의결')
print(similar_doc)
```

```
('당정 "尹대통령에 간호법 재의요구 건의키로" ', 0.7000829577445984),
('[속보] 당정 "尹대통령에 간호법 재의 요구 건의키로"', 0.6186033487319946),
('조규홍 "간호법 국민건강에 부정적...내일 국무회의서 거부권 건의" ', 0.6138016581535339),
('당정 "간호법은 新카스트법"尹대통령에 재의요구 건의키로', 0.6126188039779663),
('윤 대통령, 간호법 거부권 행사... "국민 건강 불안감 초래" ', 0.6073388457298279),
('간호법은 거부하고...정부 "범죄 저지른 의사 면허취소 과해" ', 0.6004450917243958),
('간호협회 "간호사 면허증 반납운동...의사 불법지시 거부할 것" ', 0.5854790210723877),
('복지부 장관 "간호법 거부권 행사, 혼란 최소화 고민하며 결정" ', 0.5697782635688782),
('尹 "간호법, 직역 간 과도한 갈등...국민 건강 불안감 초래" ', 0.5455094575881958),
('당정, 간호법 거부권 공식 건의키로... "생명 불모 입법독주법" ', 0.5407320857048035)
```

similar_doc에 **보도자료**를 입력

Cosine similarity를 계산해 보도자료와 유사도가 높은 순서대로 **10개의 기사**를 유사도와 함께 추출

모델 학습 및 결과

- Task 1 : 카피 기사 파악
 - Doc2Vec & ko-sentence-transformers

ko-sentence-transformers

```
from sentence_transformers import SentenceTransformer, util
import numpy as np

embedder = SentenceTransformer("jhgan/ko-sbert-sts")

# Read query from file
query_file = '/content/보도자료_간호법 재의.txt'
query = ''
with open(query_file, 'r', encoding='utf-8') as f:
    query = f.read().strip()

# Read corpus from file (multiple articles)
corpus_file = '/content/ks_ver 간호법 재의 전처리 완료 2.txt'
corpus = ''
with open(corpus_file, 'r', encoding='utf-8') as f:
    corpus = f.read().strip()

# Calculate cosine similarity
cos_scores = util.pytorch_cos_sim(query_embedding, corpus_embeddings)[0]
cos_scores = cos_scores.cpu()

# Sort the scores in descending order
top_results = np.argsort(-cos_scores)

print("\n\n===== \n\n")
print("Query:", query)
print("\nTop 5 most similar sentences in corpus:")

# Print top 5 similar sentences from corpus
top_k = 5
for idx in top_results[:top_k]:
    print(corpus_articles[idx].strip(), "(Score: %.4f)" % (cos_scores[idx].item()))
```

query_file과 **corpus_file**에 각각 **보도자료**와 **크롤링 기사 데이터셋**을 입력

Cosine similarity를 계산해 보도자료와 유사도가 높은 순서대로 **5개의 기사**를 유사도와 함께 추출

모델 학습 및 결과

- Task 1 : 카피 기사 파악
 - Doc2Vec & ko-sentence-transformers

ko-sentence-transformers 코드 결과

Top 5 most similar sentences in corpus:

15//조규홍 복지부장관 “의료현장 혼란 최소화” //간호법 제정안을 두고 의사·간호사
 76//간호법은 거부하고…정부 “범죄 저지른 의사 면허취소 과해” //조규홍 보건복지부
 8//조규홍 복지장관 “간호조무사 고졸 제한은 직업선택 자유 제한” //조규홍 보건복지
 81//간호사들 첫 집단행동 예고…의료현장 지키는 ‘준법투쟁’ 유력//국제 간호사의 날
 173//양곡법엔 거부권, 간호법엔 합구…巨野독주에 與 이종жат대, 왜?//주호영 국민의힘

(Score: 0.6721)

(Score: 0.6572)

(Score: 0.6561)

(Score: 0.6474)

(Score: 0.6431)

Cosine similarity를 계산해 보도자료와 유사도가 높은 순서대로 **5개의 기사**를 유사도와 함께 추출

모델 학습 및 결과

- Task 1 : 카피 기사 파악
 - Doc2Vec & ko-sentence-transformers

Doc2Vec & ko-sentence-transformers 코드 결과

('당정 "尹대통령에 간호법 재의요구 건의키로" ', 0.7000829577445984),	(Score: 0.6721)
('[속보] 당정 "尹대통령에 간호법 재의 요구 건의키로"', 0.6186033487319946),	(Score: 0.6572)
('조규홍 "간호법 국민건강에 부정적...내일 국무회의서 거부권 건의" ', 0.6138016581535339),	(Score: 0.6561)
('당정 "간호법은 新카스트법"尹대통령에 재의요구 건의키로', 0.6126188039779663),	(Score: 0.6474)
('윤 대통령, 간호법 거부권 행사... "국민 건강 불안감 초래" ', 0.6073388457298279),	(Score: 0.6431)
('간호법은 거부하고...정부 "범죄 저지른 의사 면허취소 과해" ', 0.6004450917243958),	
('간호협회 "간호사 면허증 반납운동...의사 불법지시 거부할 것" ', 0.5854790210723877),	
('복지부 장관 "간호법 거부권 행사, 혼란 최소화 고민하며 결정" ', 0.5697782635688782),	
('尹 "간호법, 직역 간 과도한 갈등...국민 건강 불안감 초래" ', 0.5455094575881958),	
('당정, 간호법 거부권 공식 건의키로... "생명 불모 입법독주법" ', 0.5407320857048035)	

ko-sentence-transformers는 한국어 문장에 특화된 모델로, Doc2Vec보다 문맥 정보를 잘 포착

03

모델 학습 및 결과

- Task 2 : 취재 포함 기사 파악
 - RoBERTa_large

RoBERTa_large 코드

```
dataset=[]
if add_from_excel is True:
    df = pd.read_excel(excel_file)
    df = df.iloc[:,df.shape[0]] #첫줄은 인덱스로 받아들이고 나머지줄은 데이터로 받아들이
    label0 = df[df['후속연구'] == 'N']['문장'].tolist()
    for i in range(len(label0)):
        label0[i]=label0[i]+sep+'0\n'
    label1 = df[df['후속연구'] == 'Y']['문장'].tolist()
    for i in range(len(label1)):
        label1[i]=label1[i]+sep+'1\n'
    dataset=label0[:]+label1[:]+sep
print('엑셀에서 추가된 데이터 개수:', len(dataset))

tags=[] #태그 종류 확인용
for root_dir, sub_dirs, filenames in os.walk(dir_path):
    for filename in filenames:
        if filename[-4:] != '.json': continue #json파일만 걸러내기용
        filepath = os.path.join(root_dir, filename)
        input_file_name = filepath
        with open(input_file_name) as file_in:
            for line in file_in.readlines():
                if "sentence" in line:
                    text=line[14:-3] # ,\n 으로 끝나니까 -3까지슬라이싱
                elif "tag" in line:
                    if line not in tags: tags.append(line) #태그 종류 확인용
                    # if '문제 정의' in line or '가설 설정' in line or '기술 정의' in line:
                    #     topic_idx=0
                    # elif '후속연구' in line or '성능/효과' in line:
                    #     topic_idx=1
                    # else: topic_idx=2 #연구방법에 관한 것들을 학습데이터에서 제외시키기 위함
                    if '후속연구' in line: #위처럼 했더니 결과가 잘 안나왔음
                        topic_idx=1
                    else:
                        topic_idx=0
                # elif line=='.{.}\n' or line=='.{.}\n': #연구방법에 관한 것들을 학습데이터에서 제외시키려면 주석처리, 포함시키려면 주석해제
                elif (line=='.{.}\n' or line=='.{.}\n') and topic_idx!=2: #연구방법에 관한 것들을 학습데이터에서 제외시키려면 주석해제, 포함시키려면 주석처리
                    dataset.append(text+sep+str(topic_idx)+'\n')
```

모델 학습 및 결과

- Task 2 : 취재 포함 기사 파악
 - RoBERTa_large

RoBERTa_large 코드 결과

답레이블	예측레이블	제안한 문장	제안안한 문장	제안한 문장	제안안한 문장	기사 제목	문장1	문장2	문장3	문장4	문장5	문장6	문장7	문장8
0	0	0	10	0	10	세종대로 윤석열 대	19일 서울	참석자들이	김혜윤 기	김영경 대	참석자들이	김혜윤 기	현장의 사	
0	0	2	18	2	8	"간호법 반	앞서 하루	하지만 오	이날 규탄	김영경 간	간협은 투	간협은 "다	간협은 50	대통령이
0	0	10	20	8	2	'간호법 거	오 회장은	대한간호	교육부 쪽	교육부 관	다만 이 곳	교수들은	오의금 회	서울의 한
0	0	16	24	6	4	'의사 불법	또 불법진	현장실사	정부는 간	보건복지	백찬기 간	간협은 19	이날 규탄	더불어 간
1	0	21	29	5	5	갈등 조정	여야는 정	애초부터	대통령실	양곡관리	정부는 간	2년여 정	더 큰 문제	정치권에
1	0	29	31	8	2	[사설] 국	간호사단	간호사단	대통령의	하지만 윤	대화하	타	대통령의	그런데 거
0	0	30	40	1	9	간호법에	윤 대통령	국민의힘	장동혁 원	그러나 간	야권은 일	이재명 민	민주당은	김희서 정
1	0	36	43	6	3	간호법은	조규홍 보	복지부는	다만 언제	현재 대학	복지부는	1만여명	한편, 복지	바뀐 의료
0	0	37	52	1	9	"간호법 파	이날 대한	김영경 대	또한 "코	기자회견	한편, 더	불 윤석열 대	김정호 기	김정호 기
1	0	41	57	4	5	간호사들	이번 조사	단체행동	의료현장	간호사 다	민주노총	1만여명	그러나 병	이주호 보
1	0	44	64	3	7	[사설] 해	법 쟁점은	"5 대한의사	반면 간호	또한 간호	그동안 당	양쪽이 첨	정부·여당	그래놓고
0	0	46	71	2	7	당정 "간호	윤 대통령	다만, 당정	이 법은 의	료계는	대한간호	지난 12일	간호협회	간호법에
0	0	49	77	3	6	당정, '간호	한덕수 국	연합뉴스	당정은 14	윤재옥 원	한덕수 총	대통령실	김대기 대	지난달 27
1	0	53	83	4	6	간호조무	서울 지역	간호조무	의료연대	간협이 이	이날 김영	이들은 또	앞서 의협	한편 이날
0	0	53	90	0	7	간호협회	3일 오전	연합뉴스	대한간호	이날 김영	이들은 또	한편, 의협	앞서 의협	윤석열
0	0	59	94	6	4	간호법 통	이에 간호	장숙량 중	하지만 간	우선 '간호	이번에 국	제1조 제	김윤 서울	의협 등 다

그림5) 기사 하이라이트 예시

03

모델 학습 및 결과

- Task 2 : 취재 포함 기사 파악
 - RoBERTa_large

제안 안 한 기사

	usage_ratio	val_ratio	epochs	batch_size	max_input_length	Accuracy
setting1	0.1	0.2	1	10	200	0.721153846
setting2	0.1	0.2	2	20	100	0.711538462
Setting3	0.1	0.2	3	10	100	0.634615385
setting4	0.1	0.2	3	20	80	0.721153846

03

모델 학습 및 결과

- Task 2 : 취재 포함 기사 파악
 - RoBERTa_large

제안한 기사

	usage_ratio	val_ratio	epochs	batch_size	max_input_length	Accuracy
setting1	0.1	0.2	1	10	200	0.721153846
setting2	0.1	0.2	2	20	100	0.711538462
Setting3	0.1	0.2	3	10	100	0.634615385
setting4	0.1	0.2	3	20	80	0.721153846



셀프 피드백

04

셀프 피드백

- Self-feedback & 기대 효과

- ✓ 모델 학습 시 RAM 용량 초과

- 다양한 Parameter 변경의 어려움
 - > 낮은 Epoch 설정
 - > 낮은 max_input_length 설정
- 낮은 정확도

- ✓ Task에 맞는 코드 조사 및 수정

- 적합한 모델을 찾는 과정의 어려움
- 코드 수정의 어려움

- ✓ 뉴스 기사 현황 분석 가능

- 언론사와 기자들의 취재 상황에 대한 문제 제기
 - > 자정 작용 유발 가능
 - > 발전 가능성 제시

- ✓ 공감 피로 개선 효과

- 비구독자에게 사회 이슈에 대한 관심 유발



참고 문헌

및

출처

참고 문헌 및 출처

[참고 문헌]

<https://wikidocs.net/155356>

<https://wikidocs.net/30682>

<https://pypi.org/project/ko-sentence-transformers/>

[국내 논문 문장 의미 태깅 데이터셋 \(kisti.re.kr\)](http://kisti.re.kr)

[\[Private 29위 / Public : 0.86068 / RoBERTa_large\] – DACON](#)

[보도자료/뉴스 기사 출처]

[알림 > 보도자료 내용보기 - '간호법안' 검색결과 " 간호법안 재의요구하기로 제20회 국무회의 의결 " | 힘이 되는 평생 친구, 보건복지부 \(mohw.go.kr\)](#)

[공약과 거부권 사이 '간호법'... 50만 간호사가 운다 - 여성신문 \(womennews.co.kr\)](#)



Q&A

Thank You