BERT로 신문사 분류하기

www.kssah.com Fintech 4th team newspaper Since 2020

언론사별로 성향이나 문체의 차이가 있을까?

발행인: 김보경 송영희 신지민 안재하 허정준

Contents

프로젝트 소개

Q. 신문사마다 정치적 성향이 다르니까, 기사만 보고도 신문사를 구별할 수 있지않을까?

Q. 신문사마다 글을 작성하는 스타일(문체)에 특징이 있을까?

Q. 신문사마다 특별히 사용하는 어휘 표현으로도 신문사를 구별할 수 있을까?

콘텐트 vs 컨텐츠

요즘 시대에 외래어 표준규정집보다 본인 신문사가 먼저 써왔으니 맞다고 쓰는 데 그래서 그 신문사가 어디인지 알다가도 모르겠네. 그걸 버트가 구분하나봅니다.

외래어 표기차이

##일보 김보경 기자

아 무 개 ? ! 그 ? ! 그녀는 어디로..

분명 내가 본 사건의 주인공은 40대 아주머니였지만 기사에서는 [illegible]를 표현해주는 데 이게 [illegible] 차단해주는 [illegible] 남아있는 [illegible] 세상은 밝습니다.

선호하는 인칭대명사

##신문 송영희 기자

코로나 역시 코리아 vs 정부는 코로나 뭐하나

같은 기사여도 잘한점만 써주는 곳, 못한 것만 써주는 곳 [illegible] 골고루 열심히 [illegible] 찾아서 써주면나 펙트가 중요한지 생각의 폭을 넓히게 되는지 생각해보자

진보 VS 보수 성향 차이

##일보 신지민 기자

逗 有 老 漢字

나랏 말ᄊᆞ미 中國듕귁에 달아 文字문쯩와로 서르 ᄉᆞᄆᆞᆺ디 아니ᄒᆞᆯ씨 이런 젼ᄎᆞ로 어린 百姓빅셩이 니르고져 ᄒᆞᇙ 배 이셔도 ᄆᆞᄎᆞᆷ내 제 ᄠᅳ들 시러 펴디 몯ᄒᆞᇙ 노미 하니라 내 이를 ᄆᆡᇰᄀᆞ노니 사ᄅᆞᆷ마다 ᄒᆡ여 수ᄫᅵ 니겨 날로 ᄲᅮ메 便뼌安ᅀᅡᆫ킈 ᄒᆞ고져 ᄒᆞᇙ ᄯᆞᄅᆞ미니라。

한자 사용빈도

##신문 안재하 기자

초벌구이 재벌구이 글에도 예외없다

오늘부터 나도 기자! 열심히 초안 쓰고 사수님께 한번 검토 두번검토 세번검토 국장님께까지 가나 어쩌나 바로 나가려나 포맷,형태 문제는 남으려나 지금 내 주변 사람들 다 비슷하게 쓰려나 알고싶은데 누가 알려주면 좋겠습니다.

문체, 글의 형식

##일보 허정준 기자

1. 데이터 소개

1) BeautifulSoup 패키지를 활용, 네이버 뉴스 크롤링

수집 기간: 2020.02.15 ~ 2020.04.10

검색 키워드: “코로나“

수집 항목: 날짜, 신문사, 기사 제목, 기사 내용, url

	date	press	title	contents	link
0	2020.04.10.	한겨레	“반복적 비말 노출”…콜센터 직원 첫 ‘코로나 산재’	집단감염 구로 콜센터 밀집공간 “업무와 상당한 인과관계 있다” 발병경로 확인돼 3주…	https://news.naver.com/main/read.nhn?mode=LSD&…
1	2020.04.10.	경향신문	코로나19 감염 노동자 국내 첫 ‘산재’ 인정	‧구로 콜센터 직원 1명 신속 판정‧간호사·조무사 3명도 심의 중‧‘업무 연관’ 주…	https://news.naver.com/main/read.nhn?mode=LSD&…
2	2020.04.10.	세계일보	구로 콜센터 상담원 코로나 산재 첫 인정	대구 52일 만에 신규 환자 ‘0’ 기록서울 구로구 콜센터에서 일하다 코로나19 확…	https://news.naver.com/main/read.nhn?mode=LSD&…
3	2020.04.10.	아시아경제	코로나19의 비밀 풀었다..RNA전사체 첫 분석	사스코로나바이러스-2의 생활사[아시아경제 황준호 기자] 국내 연구진이 신종 코로나바…	https://news.naver.com/main/read.nhn?mode=LSD&…
4	2020.04.10.	중앙일보	과연 노벨상 후보…김빛내리 코로나19 치료제 열쇠 찾았다	장혜식 교수와 연구, 세계 첫 성과코로나 RNA 전사체 최초로 분석고해상도 유전자 …	https://news.naver.com/main/read.nhn?mode=LSD&…
5	2020.04.10.	헤럴드경제	김빛내리 교수, 세계 첫 ‘코로나 유전자 지도’ 완성	바이러스 RNA 전사체 분석진단키트 개선·치료제 개발 가속국내 과학계서 노벨상 후보…	https://news.naver.com/main/read.nhn?mode=LSD&…
6	2020.04.10.	세계일보	코로나19 바이러스 유전자 비밀 풀었다	IBS·국립보건연구원 공동연구팀 / 노벨상 유력평가 김빛내리 단장 / RNA 변형 …	https://news.naver.com/main/read.nhn?mode=LSD&…

2) 전처리

중복되는 내용의 기사 제거 (641개 제거)

3) 결과

신문사	기사 수
서울경제	1215
문화일보	1131
한국경제	1018
파이낸셜뉴스	1005
경향신문	994
매일경제	934
전자신문	907
헤럴드경제	737
한국일보	712
동아일보	702
세계일보	674
서울신문	666
국민일보	661
조선일보	661
한겨레	661
이데일리	580
디지털타임스	574
아시아경제	503
머니투데이	503
중앙일보	438
중앙SUNDAY	29

신문사 21개, 기사 15,305개

신문사	기사 수
경향신문	994
동아일보	702
조선일보	661
한겨레	661
중앙일보	438

신문사 5개, 기사 3,456개
80% 훈련, 20% 테스트

신문사	기사 수
국민일보	661
이데일리	702
머니투데이	661

추가적인 테스트에 활용

2. 전처리 과정

1) 신문사 정보 제거

코로나19감염과 업무상 질병이 처음으로 인정됨에 따라 추가 산재인정도 잇따를 것으로 보인다. 현재까지 근로복지공단에는 총 4건의 코로나19 관련 산재신청이 접수됐다. ㄱ씨 사례를 제외한 3건 신청자는 의료기관에 종사하는 간호사, 간호조무사들이다. 근로복지공단은 업무수행 과정에서 코로나19감염자와의 접촉이 확인되고, 생활공간이나 지역사회에서 감염자와의 접촉이 없었을 경우 업무상 질병으로 인정할 수 있다는 입장이다.

강순희 근로복지공단 이사장은 "앞으로도 코로나19 산재신청을 포함해 업무상 재해를 입은 산재노동자가 적기에 적절한 재해보상을 받을 수 있도록 최선을 다하겠다"고 말했다.

이효상 기자 hslee@kyunghyang.com

▶ 장도리 | 그림마당 보기
▶ 경향신문 바로가기 ▶ 경향신문 구독신청하기

©경향신문(www.khan.co.kr), 무단전재 및 재배포 금지

```
def delete_press_info(text):
  press_list = ['조선일보', '조선', '한겨레', '한겨레신문', '경향
  press_web_list = ['hani', 'kyunghyang', 'kahn', 'joongang', 'ch

  for name in press_list:
    text = re.sub(name, '', text)
  text = re.sub('[a-zA-Z0-9+-_.]+[@]+[a-zA-Z0-9-.]+', '', text)
  for web in press_web_list:
    text = re.sub(web, '', text)
  text = re.sub('₩D{5}[기자]₩D{5}', '', text)

  text = text[:-100]
  return text
```

기자의 이름, 이메일, 그리고 신문사 이름이나

연재 기사들이 포함되어 있음.

→ 정규표현식으로 제거

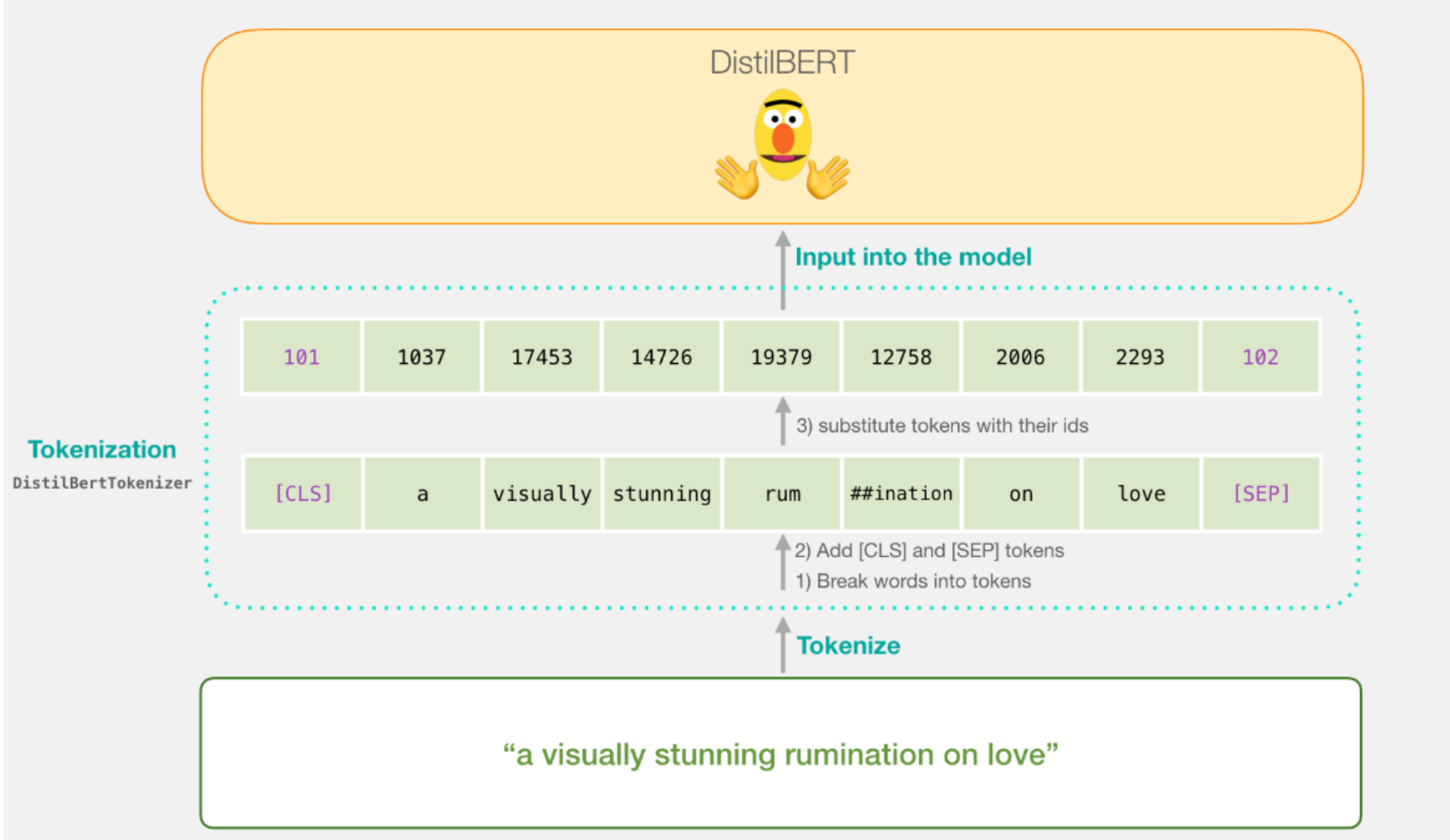
2) 한국어 형태소 분석

Mecab을 사용하여 형태소 분석

불용어 제거(은, 는, 이, 가……)

전처리한 후 문장으로 연결하여 입력자료로 활용

3. 모델링 - DistilBert

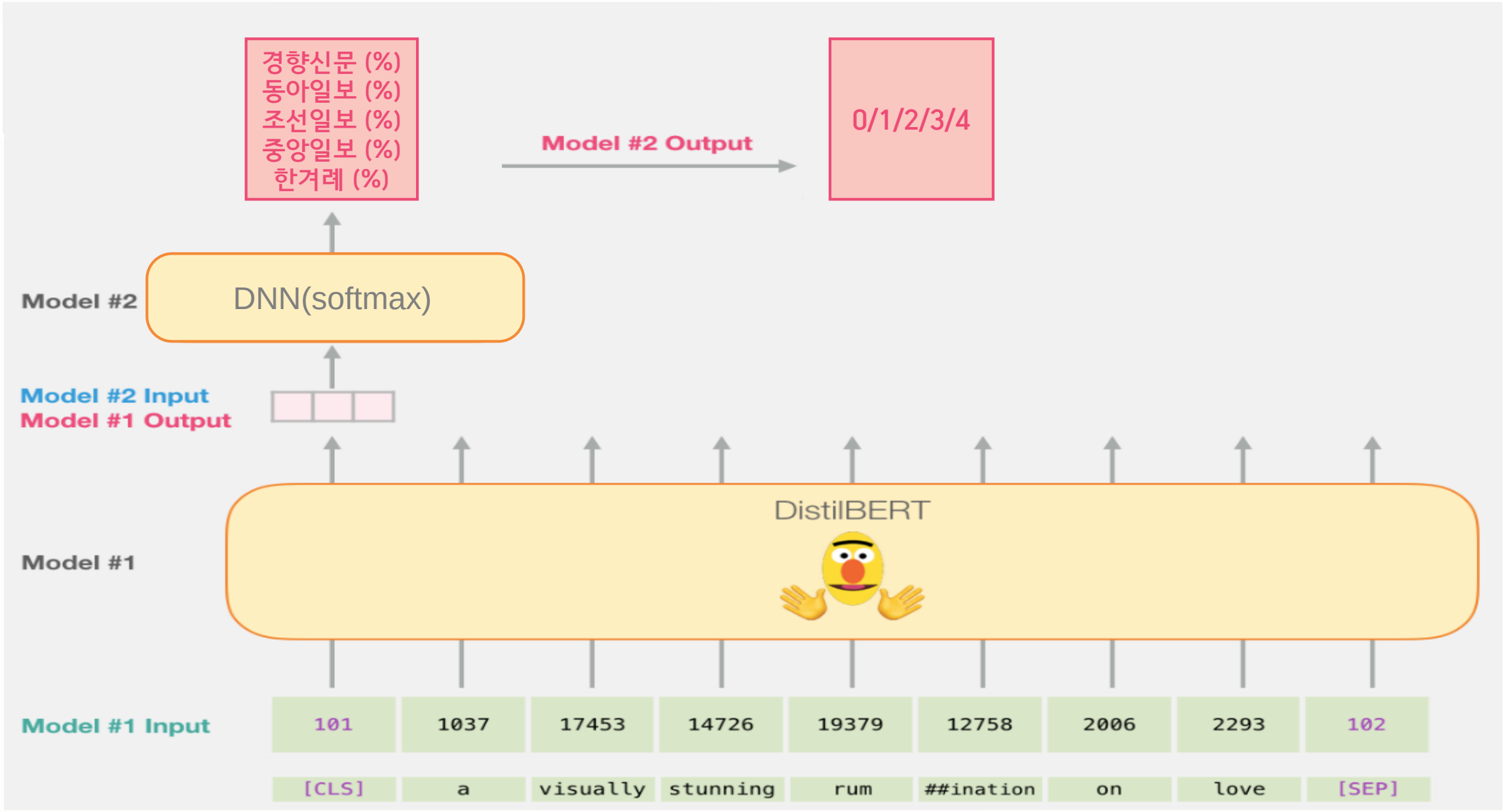


BERT 대비

40% 적은 파라미터

60% 이상 빠른 속도

95% 이상의 성능



http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

4.1. 주요내용 확인

1) 기사 제목

	Precision	Recall	F1 score	Accuracy
경향신문	0.52	0.84	0.64	0.57
동아일보	0.65	0.51	0.57	
조선일보	0.79	0.60	0.68	
중앙일보	0.61	0.41	0.49	
한겨레	0.39	0.30	0.34	

2) 기사 내용

	Precision	Recall	F1 score	Accuracy
경향신문	0.92	0.87	0.89	0.86
동아일보	0.84	0.80	0.82	
조선일보	0.95	0.97	0.96	
중앙일보	0.74	0.80	0.77	
한겨레	0.79	0.83	0.81	

3) 기사 제목+내용

	Precision	Recall	F1 score	Accuracy
경향신문	0.90	0.89	0.90	0.90
동아일보	0.92	0.92	0.82	
조선일보	0.95	0.98	0.97	
중앙일보	0.92	0.80	0.86	
한겨레	0.84	0.89	0.86	

- 기사 제목만 훈련시켰을 때는 분류가 잘 안됨,

  기사 제목과 내용을 함께 넣었을 때 정확도가 가장 높음.

- 조선일보 > 경향신문 > 동아일보> 중앙일보 or 한겨레

  순으로 예측 정확도가 높음

4.2. 정치적 성향 확인

1단계 : 5대 신문사별 유사도 및 성향 확인

신문사	가장 유사한 신문사	정확도
조선일보	중앙일보	보수 85.63%
동아일보	중앙일보	보수 87.75%
중앙일보	동아일보	보수 55.71%
경향신문	한겨레	진보 76.66%
한겨레	경향신문	진보 78.00%

2단계 : 이외 신문사 별 유사도 및 성향 확인

신문사	가장 유사한 신문사	정확도
국민일보	중앙일보	보수 69.92%
이데일리	중앙일보	진보 7.07%
머니투데이	조선일보	진보 22.07%

3단계 : 5개 신문사의 성향을 배제한 정확도

신문사	정확도
조선일보	88.61%
동아일보	
중앙일보	
경향신문	90.63%
한겨레	

- 5개 신문사 내에서는 성향별 유사도가 확인
- 이외의 신문사는 정치 성향 특징이 나타나지 x
- 성향을 배제한 상태에서도 높은 정확도로 분류, 따라서 문체, 어휘 등의 다른 특징에 의해서도 분류가 잘 되는 것을 알 수 있음

4.3. 특정 어휘 표현 확인

표현	예상	결과
그는		조선
그녀는	(경향 제외)	조선
아무개	한겨레	조선
콘텐트	중앙일보	조선
노동자		조선
권리		중앙
最高價(한자)	조선	조선
각 당의 이름		중앙
진보 보수		중앙
[토요판]	한겨레	조선

단어 조합에 따라 분류도 달라져

"심상정" → 중앙

"정의당" → 중앙

"상임선대위원장" → 중앙

"심상정 정의당 상임선대위원장" → 한겨레

4.3. 문장 조합 확인

구로 콜센터 직원 1명 신속 판정 、간호사·조무사 3명도 심의 중 、‘업무 연관’ 추가 사례 이어질 듯.

→ 한겨레

구로 콜센터 직원 1명 신속 판정 、간호사·조무사 3명도 심의 중 、‘업무 연관’ 추가 사례 이어질 듯. 서울 구로구 콜센터에서 일하다 코로나19에 감염된 노동자가 산업재해 인정을 받았다.

→ 동아

구로 콜센터 직원 1명 신속 판정 、간호사·조무사 3명도 심의 중 、‘업무 연관’ 추가 사례 이어질 듯. 서울 구로구 콜센터에서 일하다 코로나19에 감염된 노동자가 산업재해 인정을 받았다. 코로나19 감염과 업무 연관성이 인정된 국내 첫 사례다.근로복지공단은 구로구 콜센터에서 근무 중 코로나19 확진 판정을 받은 ㄱ씨의 산재 신청과 관련해 업무상질병판정위원회 심의를 거쳐 업무상 질병으로 승인했다고 10일 밝혔다.

→ 답: 경향

문장도 조합에 따라 분류가 달라짐

가시적인 패턴 확인이 어려움

특정 단어에 의존하지 않고

문맥에 의해 판단하는 것으로 보임

Contribution 분석 - 중앙일보

+2.373	존슨
+1.714	고
+1.378	총리가
+0.963	영국
+0.837	총리의
+0.821	총리
+0.544	치료를
+0.535	코로나19
+0.500	산소
+0.478	있다
+0.426	이
+0.337	정상적인
+0.335	당신의
+0.296	보리스
+0.292	라브
+0.286	계획을
+0.261	폐렴은
+0.259	<BIAS>
+0.254	그는
+0.253	전
+0.248	한다
+0.242	사실을

+0.234	세인트토머스
+0.226	도널드
+0.222	아니다
+0.209	prayforboris
+0.200	트럼프
+0.199	총리실
+0.196	빈다
+0.193	병원
+0.193	런던
+0.167	트위터에
+0.162	는
+0.152	받았다
+0.150	영국은
+0.150	경우
+0.146	엘리자베스
+0.146	매우
+0.144	상태가
+0.143	가족과
+0.138	메시지를
+0.133	설명했다
+0.126	말했다
+0.121	bbc방송
+0.118	숨겨문
+0.118	지시와
+0.114	못했다
+0.114	입원한

어떤 표현이 가장 크게 영향을 미쳤는지 분석

-0.063	영국인들은 트위터에
-0.063	소식을 듣고
-0.067	7일 브리핑에서도
-0.081	나라가 당신을
-0.081	외무장관에 총리
-0.084	상태 위중해져
-0.086	나가겠다 고
-0.086	화상회의에서 뵈었는
-0.088	트위터에 해시태그
-0.091	병원에 입원다
-0.097	영국 싱크탱크인
-0.097	대변인은 총리는
-0.098	정치 성향은
-0.099	중 한
-0.100	훌륭한 의료진들이
-0.103	명 사망자는
-0.103	있는 재무장관이나
-0.105	총리 직무대행
-0.108	당신을 필요로
-0.111	영국에서 코로나19
-0.112	입원 소식을
-0.116	일생에서 그
-0.119	6월 윈스턴
-0.128	전했다 트럼프
-0.128	인한 국민의
-0.128	며 코로나19와의

-0.129	있나 예상하시
-0.133	데이비드 캐머런
-0.135	뇌졸중으로 쓰러졌을
-0.138	외무장관 afp
-0.141	전했다 하지렉
-0.146	고참 내각
-0.153	수 있다
-0.164	재무장관이나 내무장관
-0.238	런던 세인트토머스
-0.251	필요로 한다
-0.299	보리스 존슨
-0.314	폐렴은 아니다
-0.393	쾌유를 빈다
-0.448	고 밝혔다
-0.492	총리실 대변인은
-0.505	고 말했다
-0.530	세인트토머스 병원
-0.533	존슨 총리의
-0.542	라브 장관은
-0.586	도널드 트럼프
-0.598	도미닉 라브
-0.607	치료를 받았다
-0.709	전 영국
-1.205	고 전했다
-1.295	산소 치료를
-2.740	존슨 총리가

Contribution 분석 - 조선일보

+1.186	층간
+1.085	소음
+0.905	고
+0.704	집에
+0.657	코로나
+0.478	며
+0.462	씨는
+0.448	사는
+0.409	민원이
+0.383	여
+0.359	했다
+0.350	못하게
+0.350	1명
+0.318	안
+0.314	<BIAS>
+0.304	머물고
+0.303	늘었다

-0.178	는 하루하루구성한
-0.180	는 3개월
-0.182	다섯 번이나
-0.185	한계가 있다
-0.187	민원 작년보다
-0.187	사는 고시원에서도
-0.188	사람 1명
-0.189	동작구 한
-0.192	크다 고
-0.199	소음 민원이
-0.199	주거문화개선연구소장은
-0.199	충북 청주의
-0.207	씨는 요즘
-0.219	없었는데 코로나
-0.220	역할이 크다
-0.226	안양에 사는
-0.227	내 단지
-0.236	늘었다 고
-0.271	남모 60
-0.418	집에 머물고
-0.788	고 했다
-1.002	층간 소음

Contribution 분석 - 동아일보

+1.197	24시간
+0.793	코로나19
+0.682	2시간이면
+0.669	2시간
+0.503	고
+0.441	있다
+0.389	파견된
+0.363	인력을
+0.352	2배
+0.337	24시간 환자
+0.281	방호복을
+0.273	양회성
+0.259	병원들이
+0.249	간호사가
+0.232	경력
+0.222	코로나바이러스
+0.211	것도

-0.198	지적했다 향후
-0.200	온 몸이
-0.200	사람도 있다
-0.208	간호 인력이
-0.218	방호복을 입은
-0.219	평소 양더을
-0.221	2배 많은
-0.222	있다는 지적이
-0.224	국립의료원 8층
-0.228	환자
-0.236	고글 때문에
-0.239	<BIAS>
-0.246	대구경북 지역
-0.249	상태 청소까지
-0.251	말했다
-0.252	병원들이 평소
-0.262	평소보다 2배
-0.267	있는 간호사를
-0.296	거의

특별히 의미 있는 단어나, 구절은 없었음

5.1. 결론 및 해석

1) 기사 내용을 바탕으로 신문사를 분류하는 것이 가능

2) 정치적 성향이 비슷한 신문 사이에 어느 정도 내용의 유사성이 있는 것으로 보임, 그러나 주요 판단 기준인지 알 수 없음 (신문사별 특색이 존재)

3) 특정 어휘는 가중치에 큰 영향을 미치지 않음. 문체, 문맥 등 종합적 요소가 영향을 미칠 것으로 예상

5.2. 보완할 점

1) ETRI KoBERT와 같은 한국어를 고려한 모델 활용

2) Contribution의 해석 및 판단 근거, 특징을 도출하는 것이 필요

Thank you ☺