

언론사별 정치 편향성 분석



CONTENTS



01 프로젝트 개요

주제 선정 배경 및 목적	4P
연구 방법 및 분석 절차	5P

02 데이터 수집

언론사별 정치 기사 크롤링	8-9P
----------------	------

03 데이터 전처리

데이터 병합	11P
결측값 처리	12P
한글화	13P
토큰화	14-20P
품사 태깅 및 추출	21P
불용어 처리	22P

04 데이터 분석

토픽 모델링	24-32P
감성분석	33-35P

05 분석 결과

언론사별 메인 토픽 비율	37P
언론사별 긍부정 비율	38P
언론사별 정치 편향성	39-40P

06 향후 과제

개선점	41P
-----	-----

07 진행 프로세스

개발 일정	46P
직업 분담	47P

1. 프로젝트 개요

주제 선정 배경, 및 목적

연구 방법 및 분석 절차

01. 프로젝트 개요

주제 선정 배경 및 목적

1. 언론사들의 정치적 편향성 분석

2. 확증편향 현상의 완화

3. 비판적인 정보 소비 기반 마련

?

확증편향

자기 견해와 비슷한 정보만 선택적으로 받아들이고, 반대되는 사실은 무시하면서 기존의 믿음을 강화하는 경향



01. 프로젝트 개요

연구 방법 및 분석 절차

토픽 모델링을 통해, 언론사별 기사들의 토픽 분포 산출(더불어민주당, 국민의힘)



감성 분석을 통해, 언론사별 기사들의 긍/부정 비율 산출



토픽 모델링 결과 및 감성분석 결과를 통해, 언론사별 정치적 편향성 분석

2. 테이터 수집

언론사별 정치 기사 크롤링

02. 데이터 수집

언론사별 정치 기사 크롤링

언론사 리스트

네이버 뉴스 구독 TOP13 언론사

(YTN, JTBC, MBC, SBS, 국민일보, 매일경제, 조선일보,
중앙일보, 아시아경제, 한국경제, KBS, 한겨례, 경향신문)

수집 대상

더불어민주당, 국민의힘 관련 정치 기사

- 기사 제목
- 기사 본문

수집 기간

2020.09.02 ~ 2023.12.31

수집 키워드

- 더불어민주당
- 국민의힘

02. 데이터 수집

언론사별 정치 기사 크롤링

데이터 수집 결과

YTN	JTBC	MBC	SBS	국민일보	매일경제	조선일보
28070	20713	27535	25879	27158	23681	X

중앙일보	아시아경제	한국경제	KBS	한겨례	경향신문
X	30389	26000	25810	X	27282

조선일보, 중앙일보, 한겨례 언론사는 **크롤링 불가능함**

03. 데이터 전처리

언론사별 정치 기사 크롤링

조선일보

```
▼<div class="article-header__headline-cont;
  box--pad-left-md box--pad-right-md">
  <h1 class="article-header__headline" fo;
    <span>[제29회 LG배 조선일보 기왕선] 신종
  </h1>
  <p class="font--primary font--size-md-20
    </div>
</div>
<section class="grid grid__col--lg-12 grid_
```

태그에
언더바(_)가 아닌, 가운데 막대기(-)
→ 웹 크롤링 금지

중앙일보

JO/NS PRIME

조인스프라임 서비스가
2022년 6월 30일 종료되었습니다.

그동안 조인스프라임 서비스를 이용해주신 분들께 감사드립니다.

네이버 뉴스 연동 서비스 종료

한겨레

제22조 (인공지능 학습 및 크롤링)

1. 회사는 회사 홈페이지 및 회사와 제휴를 맺은 사이트 등을
통해 제공하는 모든 한겨레신문(주) 콘텐츠에 대해 로봇(봇),
크롤러, 스파이더, 스크래퍼, 매크로 등 모든 자동화 도구나
수동 프로세스를 활용한 일체의 행위를 허용하지 않습니다.

한겨레는 로봇, 크롤러, 스크래퍼 등
자동화 도구를 통한 수집 행위 허용X

3. 데이터 전처리

데이터 병합

결측값 처리

한글화

토큰화

품사 태깅 및 추출

불용어 처리

03. 데이터 전처리

데이터 병합

언론사별 데이터

AsiaGyeongJe.pkl

GukMinIlBo.pkl

GyeongHyang.pkl

HanGukGyeongJe.pkl

HanGyeoRye.pkl

JoSeonIlBo.pkl

JTBC.pkl

JungAngIlBo.pkl

KBS.pkl

MaellGyeongJe.pkl

MBC.pkl

SBS.pkl

YTN.pkl

병합된 데이터

	Query	Title	Text	News
0	더불어민주당	막대기도 당선될 판 진중권 네거티브 민주당 비판	진중권 전 동양대 교수 이미지출처연합뉴스 AD 원본보기 아이콘아시아경제 황수미 기자...	asiae
1	더불어민주당	1년만에 뒤바뀐 공수민주당 사죄 국민의힘 여론조사가 민심	이낙연 분노와 실망 아프도록 잘 안다 반성하고 혁신주호영 여론조사 2030 차이 민...	asiae
2	더불어민주당	종합민주당 오세훈 내곡동 특혜 의혹에 거짓말이 거짓말을 낳아 맹공	박영선 TV 토론회 오세훈 내곡동 특혜 의혹 두고 공세47 서울시장 보궐선거 더불어...	asiae
3	더불어민주당	박형준 안민석 더불어민주당 의원진보 유튜버 등 4명 부산지검 고발	국민의힘 캠프 부동산 투기 전혀 없다 안 의원 후보부인 부동산 복부인 발언에 반박...	asiae
4	더불어민주당	포토 더불어민주당 원내대책회의	가장 많이 읽힌 뉴스를 제공합니다 집계 기준에 따라 최대 3일 전 기사까지 제공될 ...	asiae
...
262996	더불어민주당	민주 송영길 구속영장 청구 검찰 독재정권의 총선 전략	AD더불어민주당은 검찰이 전당대회 돈봉투 의혹에 연루된 송영길 전 대표에게 구속영장...	YTN
262997	더불어민주당	여야 총선 체제 전환 속도 서울 6석 파장	AD 진행 이승희 앵커 출연 김용남 전 국민의힘 의원 김종욱 전 청와대 행정관...	YTN
262998	더불어민주당	민주 이동관 사표 수리해선 안 돼 기각 시 책임져야	AD앵커더불어민주당은 오늘1일 본회의에서 탄핵안 처리가 예고됐던 이동관 방송통신위원...	YTN
262999	더불어민주당	모기외투 증언 신빙성 인정 사법리스크 재부상	김용 1심 유죄 판결 유통규 진술이 결정적 역할재판부 경험 없이는 알 수 없을 정도...	YTN
263000	더불어민주당	정기국회 내 예산안 무산 쌍특검 20일에 추진	정기국회 9일 종료 예산안 처리 사실상 무산여야 원내대표 예결위 간사 22 협의체 ...	YTN
263001 rows × 4 columns				

03. 데이터 전처리

결측값 처리



03. 데이터 전처리

한글화

Title, Text 한글화

```
# title과 Text 열에서 영문 대소문자, 한글, 숫자, 공백 문자를 제외한 모든 문자 삭제  
total['Title'] = total['Title'].apply(lambda x: re.sub("[^0-9a-zA-Zㄱ-ㅎㅏ-ㅣ가-힣]", '', str(x)))  
total['Text'] = total['Text'].apply(lambda x: re.sub("[^0-9a-zA-Zㄱ-ㅎㅏ-ㅣ가-힣]", '', str(x)))
```

성능

apply 함수가 for 반복문보다 빠름.
데이터셋 크기가 클 경우 차이가 많이 남.

간결함

apply 함수, lambda 함수 사용으로 간단하게 데이터 전처리.
가독성 향상, 유지 관리 용이

03. 데이터 전처리

토큰화

Okt, Komoran, Hannanum, Kkma, Mecab 토크나이저 비교

Okt

오픈소스 한글 형태소 분석기

Kkma

세종 말뭉치를 이용해 생성된 사전

Komoran

Java로 쓰여진 오픈소스 한글 형태소 분석기

Hannanum

KAIST 말뭉치를 이용해 생성된 사전

Mecab

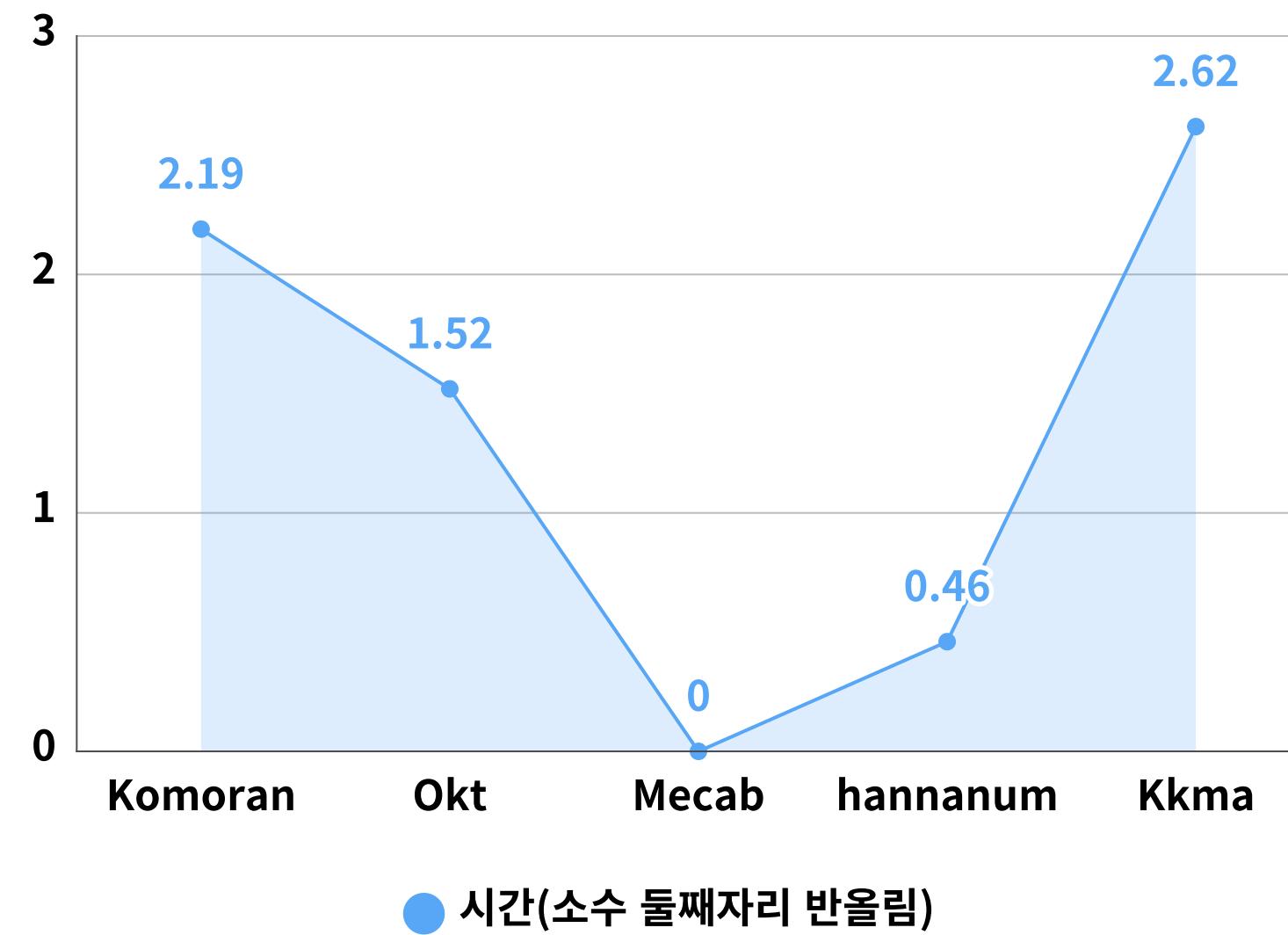
세종 말뭉치로 만들어진 csv 형태의 사전

03. 데이터 전처리

토큰화

Okt, Komoran, Hannanum, Kkma, Mecab 토크나이저 비교

문장: 아버지가방에들어가신다



토크나이저	시간	결과
Komoran	2.188122	[아버지, 가방, 에, 들어가, 시, ㄴ다]
Okt	1.524668	[아버지, 가방, 에, 들어가신다]
Mecab	0.004333	[아버지, 가, 방, 에, 들어가, 신다]
Hannanum	0.461358	[아버지가방에들어가, 이, 시느다]
Kkma	2.622740	[아버지, 가방, 에, 들어가, 시, ㄴ다]

03. 데이터 전처리

토큰화

Okt, Komoran, Hannanum, Kkma, Mecab 토크나이저 비교

문장: 국민의힘

토크나이저	시간	결과
Komoran	1.376308	[국민, 의, 힘]
Okt	0.001000	[국민, 의, 힘]
Mecab	0.003020	[국민, 의, 힘]
Hannanum	0.274814	[국민의힘]
Kkma	0.004229	[국민, 의, 힘]

문장: 더불어민주당

토크나이저	시간	결과
Komoran	1.064523	[더불어민주당]
Okt	0.003000	[더불어, 민주당]
Mecab	0.003205	[더불, 어, 민주당]
Hannanum	0.160915	[더불어민주당]
Kkma	0.006144	[더불, 어, 민주당]

03. 데이터 전처리

토큰화

토크나이저(형태소 분석기) 비교 (Okt, Komoran, Hannanum, Kkma, Mecab)

문문: 아버지가방에들어가신다			
Tokenizer	Time	Result	
0 komoran	0.800281	[아버지, 가방, 에, 들어가, 시, ㄴ다]	
1 okt	0.001000	[아버지, 가방, 에, 들어가신다]	
2 Mecab	0.001000	[아버지, 가, 방, 에, 들어가, 신다]	
3 hannanum	0.182591	[아버지가방에들어가, 이, 시ㄴ다]	
4 kkma	0.007306	[아버지, 가방, 에, 들어가, 시, ㄴ다]	

문문: 양측 노동개혁 제휴여부 주목			
Tokenizer	Time	Result	
0 komoran	0.873130	[양측, 노동, 개혁, 제휴, 여부, 주목]	
1 okt	0.001000	[양, 측, 노동, 개혁, 제휴, 여부, 주목]	
2 Mecab	0.001000	[양측, 노동, 개혁, 제휴, 여부, 주목]	
3 hannanum	0.210864	[양측, 노동개혁, 제휴여부, 주목]	
4 kkma	0.101826	[양측, 노동, 개혁, 제휴, 여부, 주목]	

실제 사용할 데이터셋에서 랜덤 추출 후,
여러 번의 실험



비교 결과

- 소요 시간: **Mecab** > Okt > kkma > hannanum > komoran
- 분석 품질: **Mecab** > Okt > Komoran > Kkma > hannanum

03. 데이터 전처리

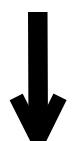
토큰화

Mecab 사용 결정 후 사용자 사전 추가 (더불어민주당, 국민의힘)

```
1 # user-dic:// 단어 추가
2
3 def addUserDic(word: str, has_coda: bool):
4     path = "C:/mecab/user-dic/nnp.csv"
5
6     if has_coda:
7         has_coda = 'T'
8     else:
9         has_coda = 'F'
10
11    with open(path, 'r', encoding='utf-8') as f:
12        file_data = f.readlines()
13        file_data.append(f'{word},,,NNP,*,{has_coda},{word},*,*,*,*\n')
14
15    with open(path, 'w', encoding='utf-8') as f:
16        for line in file_data:
17            f.write(line)
18
19    print(f"단어 '{word}'(이)가 {path}에 추가되었습니다.")
```

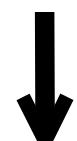
```
1 # 우선순위 변경
2
3 def changePriority(word: str, priority: int):
4     path = "C:/mecab/mecab-ko-dic/user-nnp.csv"
5
6     with open(path, 'r', encoding='utf-8') as f:
7         file_data = f.readlines()
8         for file in file_data:
9             if word in file:
10                 idx = file_data.index(file)
11                 file = file.split(',')
12                 file[3] = str(priority)
13                 file = ','.join(file)
14                 break
15
16         file_data[idx] = file
17
18     with open(path, 'w', encoding='utf-8') as f:
19         for line in file_data:
20             f.write(line)
21
22     print(f"단어 '{word}'의 우선순위가 {priority}(으)로 변경되었습니다.")
23     checkAddedUserDic()
24
25
26 changePriority("더불어민주당", 0)
```

- addUserDic 함수 정의하고 사용자가 지정한 단어를 종성 유무에 따라 'T', 'F'로 표시하고 파일에 추가



A	B	C	D	E	F	G	H	I	J	K	L	M
1 대우				NNP	*	F	대우	*	*	*	*	*
2 구글				NNP	*	T	구글	*	*	*	*	*
3 국민의힘				NNP	*	T	국민의힘	*	*	*	*	*
4 더불어민주당				NNP	*	T	더불어민주*	*	*	*	*	*

- changePriority 함수를 정의하여 주어진 단어의 우선순위를 파일에서 변경하고 변경 내용 출력



A	B	C	D	E	F	G	H	I	J	K	L	M
1 대우	1786	3545	3821	NNP	*	F	대우	*	*	*	*	*
2 구글	1786	3546	2953	NNP	*	T	구글	*	*	*	*	*
3 국민의힘	1786	3546	0	NNP	*	T	국민의힘	*	*	*	*	*
4 더불어민주	1786	3546	0	NNP	*	T	더불어민주*	*	*	*	*	*

03. 데이터 전처리

토큰화

사용자 사전 추가 결과 (더불어민주당, 국민의힘)

```
1 from konlpy.tag import Mecab
2
3 sentence = "아버지가방에들어가신다"
4 test1 = "국민의힘"
5 test2 = "더불어민주당"
6 tokenizer = Mecab(dicpath=r"C:\mecab\mecab-ko-dic")
7
8 print(tokenizer.morphs(sentence))
9 print(tokenizer.morphs(test1))
10 print(tokenizer.morphs(test2))
```

```
['아버지', '가', '방', '에', '들어가', '신다']
['국민의힘']
['더불어민주당']
```

사용자 사전 추가 후,
'국민의힘', '더불어민주당'
두 키워드가 쪼개지지 않는 것을 확인

03. 데이터 전처리

품사 태깅 및 추출

Mecab을 이용한 품사태깅 및 추출

1. 원하는 품사들의 단어만 추출하기 위해 품사 태깅

2. 명사, 형용사, 동사, 부사, 수사, 관형사 추출

	Query	Title	Text	News	Tokenized	Pos_Tagged
0	더불어민주당 막대기도 당선될 판 진중권 네거티브 민주당 비판	진중권 전 동양대 교수 이미지출처연합뉴스 AD 원본보기 아이콘아시아경제 활수미 기자...	asiae	[진중권, 전, 동양, 대, 교수, 이미지, 출처, 연합뉴스, AD, 원본, 보, ...]	[(진중권, NNP), (전, MM), (동양, NNG), (대, XPN), (교수, ...)	
1	더불어민주당 1년만에 뒤바뀐 공수민주당 사죄 국민의힘 여론조사가 민심	이낙연 분노와 실망 아프도록 잘 안다 반성하고 혁신주호영 여론조사 2030 차이 만...	asiae	[이낙연, 분노, 와, 실망, 아프, 도록, 잘, 앤, 다, 반성, 하, 고, 혁신...	[(이낙연, NNP), (분노, NNG), (와, JC), (실망, NNG), (아...	
2	더불어민주당 종합민주당 오세훈 내곡동 특혜 의혹에 거짓말이 거짓말을 날아 맹공	박영선 TV 토론회 오세훈 내곡동 특혜 의혹 두고 공세 47 서울시장 보궐선거 더불어...	asiae	[박영선, TV, 토론회, 서, 오세훈, 내곡동, 특혜, 의혹, 두, 고, 공세, 4...	[(박영선, NNP), (TV, SL), (토론회, NNG), (서, JKB), (오...	
3	더불어민주당 박형준 안민석 더불어민주당 의원 진보 유튜버 등 4명 부산지검 고발	국민의힘 캠프 부동산 투기 전혀 없다 안의원 후보부인 부동산 복부인 발언에 반박...	asiae	[국민의힘, 캠프, 부동산, 투기, 전혀, 없, 다, 안, 의원, 후보, 부인, 부...	[(국민, NNG), (의, JKG), (힘, NNG), (캠프, NNG), (부동...	
4	더불어민주당 포토 더불어민주당 원내대책회의	가장 많이 읽힌 뉴스를 제공합니다 집계 기준에 따라 최대 3일 전 기사까지 제공될...	asiae	[가장, 많이, 읽힌, 뉴스, 를, 제공, 합니다, 집계, 기준, 에, 따라, 최대...	[(가장, MAG), (많이, MAG), (읽힌, VV+ETM), (뉴스, NNG)...]	

03. 데이터 전처리

불용어 처리

불용어 처리

```
stopwords.txt
1 가
2 가까스로
3 가령
4 각
5 각각
6 각자
7 각종
8 갖고말하지만
9 같다
10 같아
11 개의치않고
12 거니와
13 거바
14 거의
15 것
16 것과같이
17 것들
18 게다가
19 게우다
20 겨우
21 견지에서
22 결과에 이르다
23 결국
24 결론을 낼 수 있다
25 겸사겸사
26 고려하면
27 고로
28 꼳
29 공동으로
30 과
31 과연
32 관계가 있다
33 관계없이
34 관계가 있다

32
33 def remove_stopwords(tokens, stopwords):
34     return [token for token in tokens if (token not in stopwords) and (len(token) > 1)] # 한 글자 초파인 단어만 추출, 불용어 사전으로 제거
35
36
37 def main():
38     df = pd.read_feather("./data/processing/posExtracted_total.feather", use_threads=True)
39     print(df.head())
40
41     url = "https://gist.githubusercontent.com/chulgil/d10b18575a73778da4bc83853385465c/raw/a1a451421097fa9a93179cb1f1f0dc392f1f9da9/stopwords.txt" # 불용어 사전
42     response = requests.get(url) # 불용어 사전 다운로드
43     data = response.content.decode("utf-8") # 불용어 사전을 utf-8로 디코딩
44
45     stopwords = data.split("\n") # 불용어 사전을 줄바꿈을 기준으로 분리
46     stopwords = [word for word in stopwords if word] # 빈 문자열 제거
47
48     print(f'stopwords: {stopwords}')
49     df["Title"] = parmap.map(remove_stopwords, df["Title"], stopwords, pm_pbar=True)
50     df["Text"] = parmap.map(remove_stopwords, df["Text"], stopwords, pm_pbar=True)
51
52     print(df.head())
53
54     df.to_feather("./data/processing/stopWordsRemoved_total.feather")
55
56
57 if __name__ == "__main__":
58     main()
```

한 글자, 불용어 사전 기반 토큰 제거

4. 데이터 분석

토픽 모델링

감성분석

04. 데이터 분석

토픽모델링

각 기사들이 어떤 정당에 대해 쓴 기사인가?

접근 방법 3가지

TextRank

- 문서 내 문장 간의 유사성을 기반으로 그래프를 생성
- PageRank 알고리즘을 적용하여 중요한 단어를 추출 후 빈도수를 통해 분류

Keybert

- BERT-based model
- 문서를 대표할 수 있는 키워드들을 추출 후 빈도수를 통해 분류

bart-large-mnli

- Zero-Shot Text Classification 모델을 통해 텍스트의 클래스를 분류

04. 데이터 분석

토픽모델링

TextRank

	Query	News	Title	Text	Top 20 Keywords
0	더불어 민주당	asiae	[막대기, 당선, 판진, 중권, 네거티브, 민주당, 비판]	[진중권, 전동, 양대, 교수, 이미지, 출처, 연합뉴스, 원본, 아이콘, 아시아,...]	후보, 교수, 국민의힘, 보도, 서울, 앞서, 경제, 진전, 방문, 의원, 오세훈,...
1	더불어 민주당	asiae	[공수, 민주당, 사죄, 국민의힘, 여론, 조사, 민심]	[이낙연, 분노, 실망, 아프, 반성, 혁신, 주호영, 여론, 조사, 차이, 민심,...]	국민의힘, 여론, 기자, 위원장, 서울, 조사, 아프, 대표, 호소, 사죄, 국민,...
2	더불어 민주당	asiae	[종합, 민주당, 오세훈, 내곡동, 특혜, 의혹, 거짓말, 거짓말, 맹공]	[박영선, 토론, 오세훈, 내곡동, 특혜, 의혹, 공세, 서울, 시장, 보궐, 선거...]	후보, 내곡동, 서울, 시장, 오세훈, 기억, 거짓말, 박영선, 관련, 측량, 의혹...

Top 20 Keywords	Main_Topic
후보, 교수, 국민의힘, 보도, 서울, 앞서, 경제, 진전, 방문, 의원, 오세훈,...	1
국민의힘, 여론, 기자, 위원장, 서울, 조사, 아프, 대표, 호소, 사죄, 주호영...	-1
후보, 내곡동, 서울, 시장, 오세훈, 기억, 거짓말, 박영선, 관련, 측량, 의혹...	-1
후보, 부산, 사실, 의원, 부동산, 허위, 시민, 시장, 캠프, 공표, 경제, 발...	-1

결과
Top 20 Keywords열에 20개의 중요한 단어 추출

더불어민주당 관련 단어(현재 정당 관련 인물) 개수: A
국민의힘 관련 단어(현재 정당 관련 인물) 개수: B

A와 B의 차이가 2개 이상일 경우:
A가 더 크면 Main_Topic에 0 할당 (더불어민주당)
B가 더 크면 Main_Topic에 1 할당 (국민의힘)
둘 다 아닐 경우 -1 할당

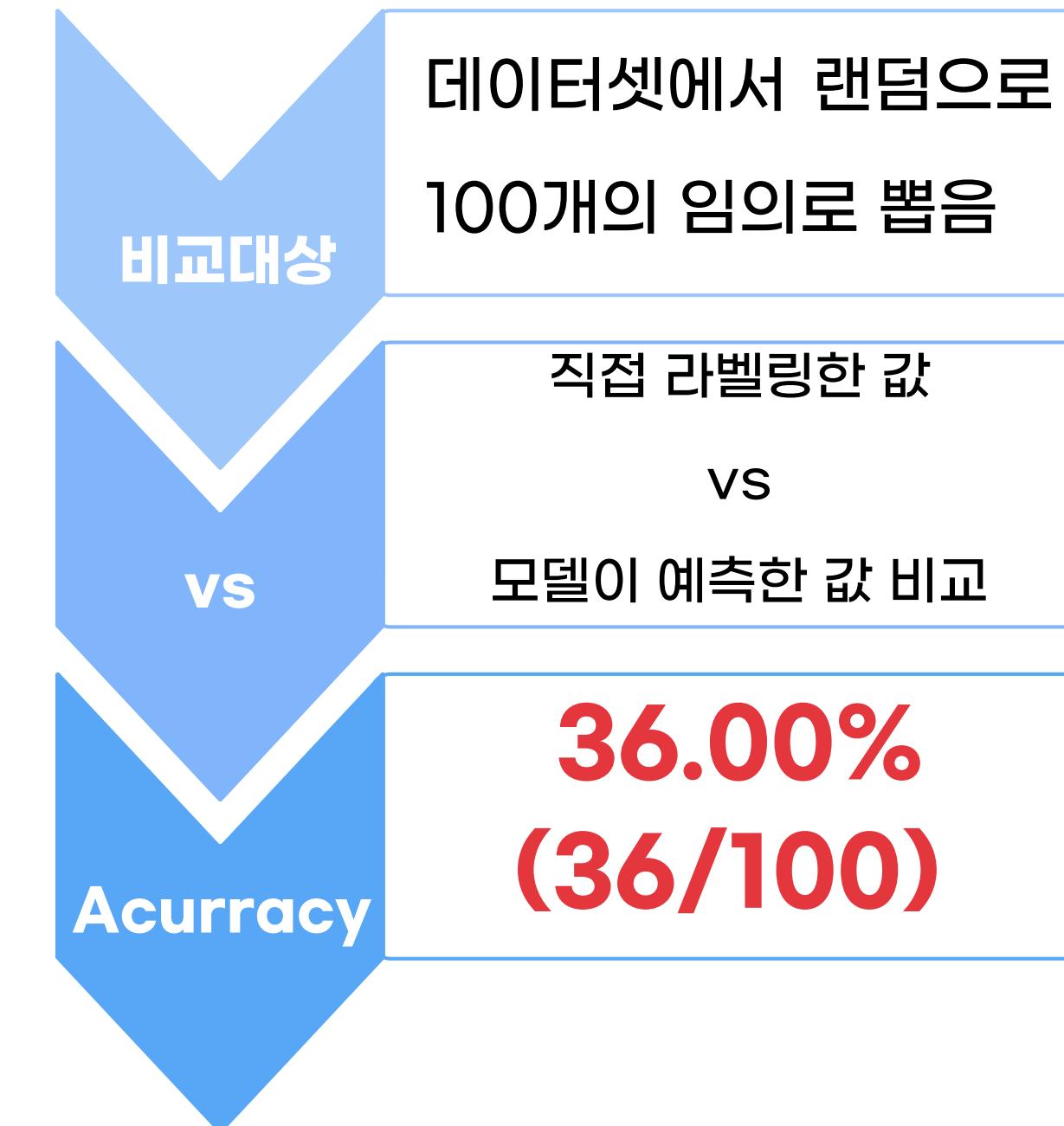
04. 데이터 분석

토픽모델링

TextRank 성능 평가

	label	Main_topic
73379	국민의힘	None
98479	더불어민주당	더불어민주당
31931	더불어민주당	더불어민주당
157359	더불어민주당	더불어민주당
171650	국민의힘	국민의힘
...
69378	더불어민주당	None
46215	국민의힘	국민의힘
20855	중립	None
135924	중립	국민의힘
76387	중립	더불어민주당

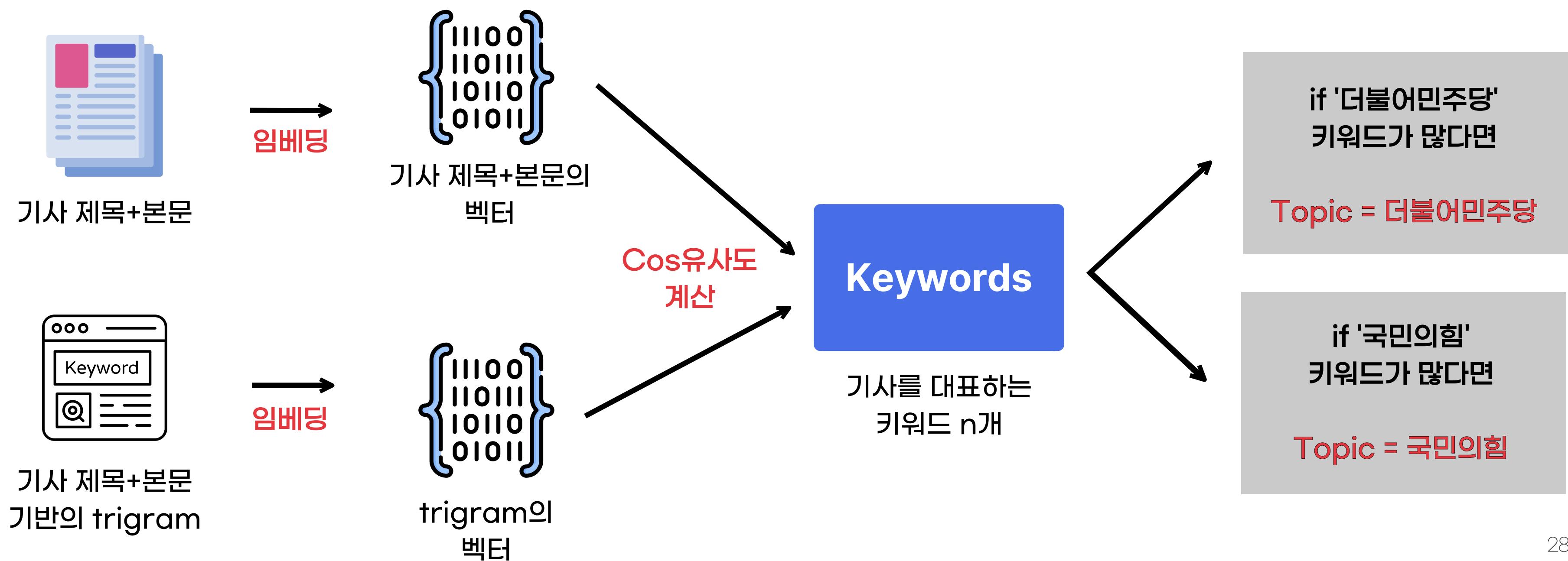
100 rows × 2 columns



04. 데이터 분석

토픽모델링

Keybert



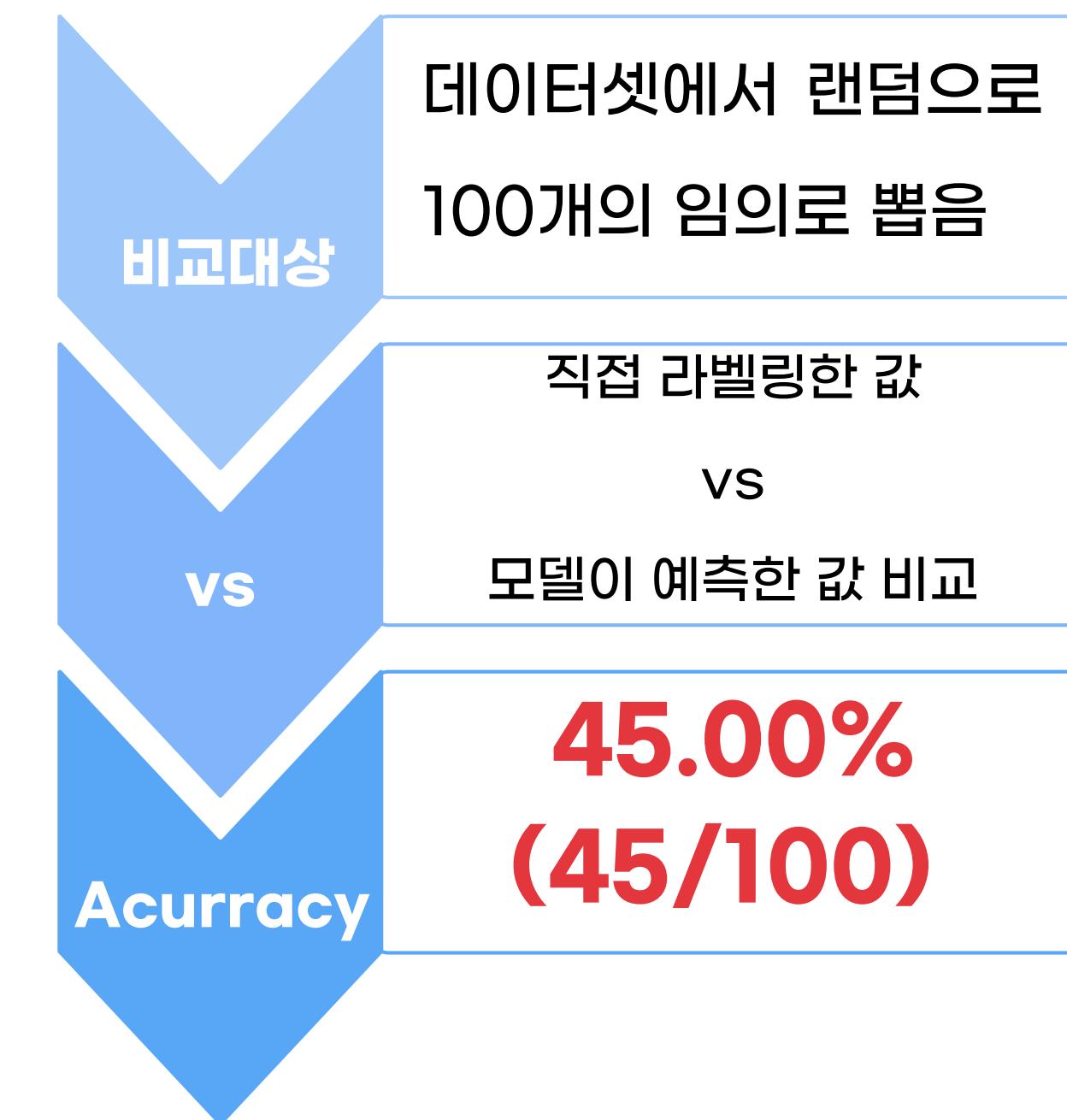
04. 데이터 분석

토픽모델링

Keybert 성능 평가

	label	Main_topic
73379	국민의힘	국민의힘
98479	더불어민주당	None
31931	더불어민주당	더불어민주당
157359	더불어민주당	더불어민주당
171650	국민의힘	국민의힘
...
69378	더불어민주당	국민의힘
46215	국민의힘	None
20855	중립	국민의힘
135924	중립	None
76387	중립	더불어민주당

100 rows × 2 columns



04. 데이터 분석

토픽모델링

bart-large-mnli (zero-shot Text Classification model)

ZSL(Zero-shot Learning): 사전에 해당 범주나 개념의 예를 보지 않고도 객체나 개념을 인식하고 분류하도록 AI 모델을 훈련하는 머신 러닝 시나리오

 Hugging Face

Inference API ⓘ

Zero-Shot Classification Examples

박찬대 “국민의힘 일하라”...17일 본회의 ‘원 구성’ 압박
더불어민주당 박찬대 원내대표는 오늘(14일) 국회 일정 보이콧 중인 국민의힘을 향해 “세비 아깝다는 비판이 안 들리나. 고집 그만 부리고 일하러 나오라”며 원 구성을 조속히 마무리할 것을 촉구했습니다.
박 원내대표는 오늘 국회 최고위원회의에서 “국민의힘의 불법 무노동 생떼 쓰기에 국회 반쪽이 멈춰있다”면서 “국민의힘은 국회의장 주재 양당 원내대표 회동도 거부했다”며 이 같이 말했습니다.
그러면서 “국회의장이 이제 결단을 내려줘야 한다. 이만하면 충분히 기다렸고 기회도 낙관적 드렸다”며 “월요일(17일)에 본회의를 열어 7개 상임위를 구성하도록 거듭 요청 드린다”고 했습니다.
해병대원 특검법의 6월 국회 회기 내 처리 의지도 거듭 확인했습니다.
박 원내대표는 “특검법의 이번 임시회 통과가 민주당의 목표”라며 “특검에만 기대지 않고 국정조사를 병행해 진실을 밝히고 잘못이 있는 자는 일벌백계하도록 하겠다”고 말했습니다.

Possible class names (comma-separated)
더불어민주당, 국민의힘

Allow multiple true classes

Compute

Computation time on cpu: 3.700 s

Class	Probability
국민의힘	0.950
더불어민주당	0.942

JSON Output Maximize

Input = 기사 제목+본문
Class 지정 = [더불어민주당, 국민의힘]

 facebook/bart-large-mnli

Output = 지정한 Class들의 확률값

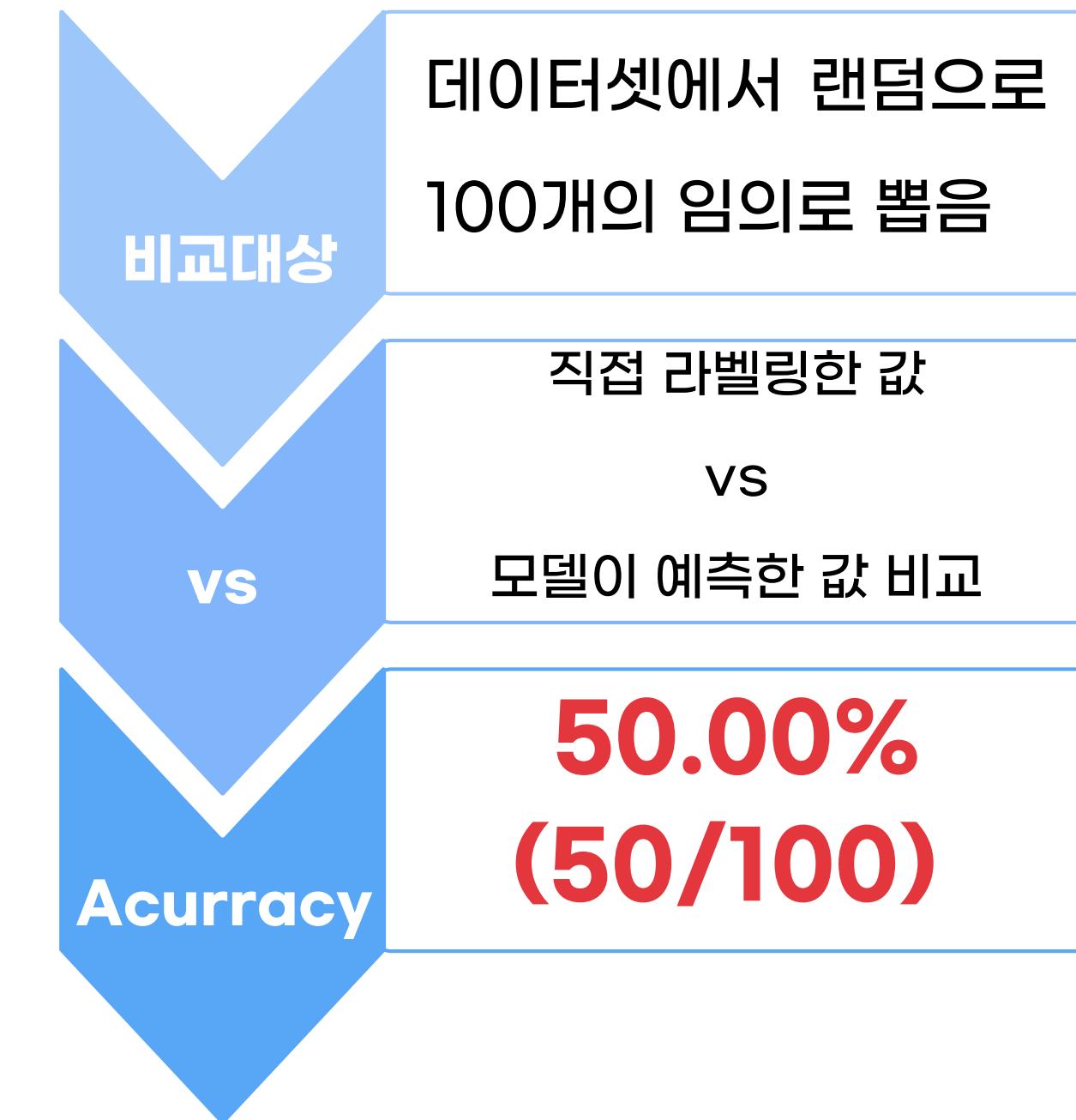
04. 데이터 분석

토픽모델링

bart-large-mnli 성능 평가

label	Main_topic
73379	국민의힘
98479	더불어민주당
31931	더불어민주당
157359	더불어민주당
171650	국민의힘
...	...
69378	더불어민주당
46215	국민의힘
20855	중립
135924	중립
76387	중립

100 rows × 2 columns



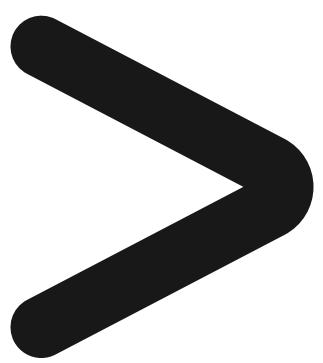
04. 데이터 분석

토픽모델링

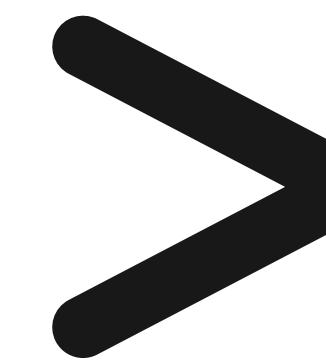
최종 토픽 모델 설정



Accuracy:
50%



Accuracy:
45.00%



Accuracy:
36.00%

04. 데이터 분석

감성분석

Knu, Kosac 감성사전 비교

Knu 감성사전		Kosac 감성사전	
개발	군산대학교	개발	한국과학기술정보연구원
활용분야	감성분석, 여론분석, 고객 반응 분석 등	활용분야	감성분석, 여론분석, 콘텐츠 추천 등
특이점	전문적, 객관적 표현 O 신조어, 축약어 등	특이점	신조어 비율 적음

Knu
→ **고객 반응 분석, 전문적, 객관적 표현O**

04. 데이터 분석

감성분석

감성스코어 비교

$0 < \text{score} \rightarrow \text{긍정}$
 $0 > \text{score} \rightarrow \text{부정}$

정확도 낮음



$0.3 < \text{score} \rightarrow \text{긍정}$
 $0.3 > \text{score} \rightarrow \text{부정}$

정확도 중간



$0.8 \leq \text{score} \rightarrow \text{긍정}$
 $0 > \text{score} \rightarrow \text{부정}$

value 값 1개 이하 중립

정확도 70%

04. 데이터 분석

감성분석

감성분석 결과

	Query	News	Text	sentiword	values	score	sentiment_category	News_Title
0	더불어 민주당	asiae	[진중권, 전등, 양대, 교수, 이 미지, 출처, 연합뉴스, 원본, 아이콘, 아시아,...	[갈등, 미인, 분노, 비판, 악의, 월등히, 의혹, 적극, 지지]	[-1, 2, -2, -1, -2, 2, -1, 1, -1]	-0.333333	부정	막대기도 당선될 판 진중권 네 거티브 민주당 비판
1	더불어 민주당	asiae	[이낙연, 분노, 실망, 아프, 반 성, 혁신, 주호영, 여론, 조사, 차이, 민심,...	[분노, 비판, 성공, 승리, 실망, 안전, 역전패, 의혹, 지지, 혁신, 희망]	[-2, -1, 2, 2, -2, 2, -1, -1, -1, 2, 1]	0.090909	중립	1년만에 뒤바뀐 공수민주당 사죄 국민의힘 여론조사가 민심
2	더불어 민주당	asiae	[박영선, 토론, 오세훈, 내곡 등, 특혜, 의혹, 공세, 서울, 시장, 보궐, 선거...	[거짓, 거짓말, 비난, 비판, 의혹, 진실]	[-2, -2, -2, -1, -1, 1]	-1.166667	부정	종합민주당 오세훈 내곡동 특혜 의혹에 거짓말이 거짓말을 낳아 맹공
3	더불어 민주당	asiae	[국민의힘, 캠프, 부동산, 투 기, 전혀, 의원, 후보, 부인, 부동산, 복부인, ...	[공연히, 명예, 명예 훼손, 범죄, 비방, 사모, 오도, 지지, 폭력, 훼손]	[-1, 2, -2, -2, -1, 2, -2, -1, -2, -1]	-0.800000	부정	박형준 안민석 더불어민주당 의원진보 유튜버 등 4명 부산지검 고발
4	더불어 민주당	asiae	[가장, 많이, 뉴스, 제공, 집계, 기준, 최대, 전기사, 제공]	[]	[]	0.000000	중립	포토 더불어민주당 원내대책회의

5. 분석 결과

언론사별 메인토pic 비율

언론사별 긍부정 비율

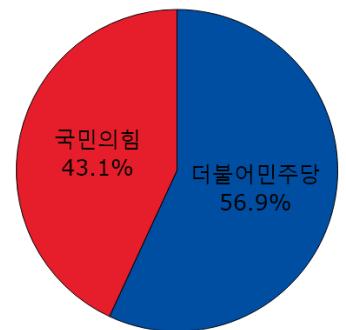
언론사별 정치 편향성

05. 데이터 분석

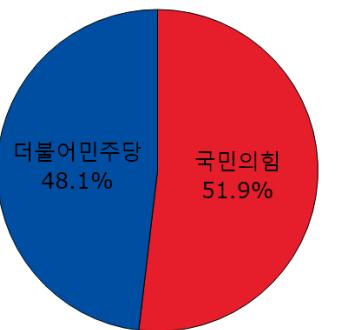
언론사별 메인토픽 비율

언론사별 토픽 분포

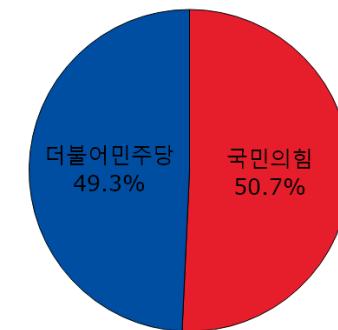
MBC 정치 기사들의 주요 토픽 비율



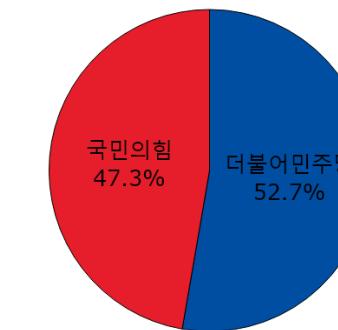
KBS 정치 기사들의 주요 토픽 비율



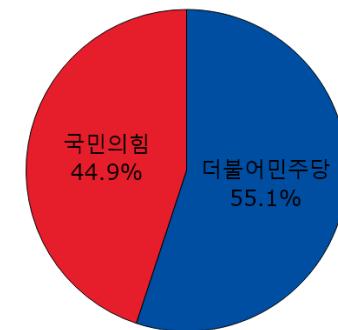
SBS 정치 기사들의 주요 토픽 비율



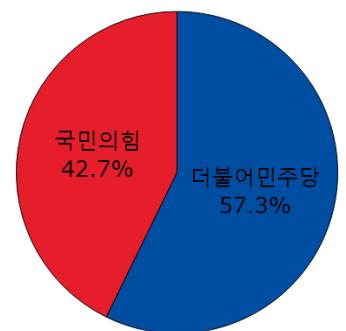
JTBC 정치 기사들의 주요 토픽 비율



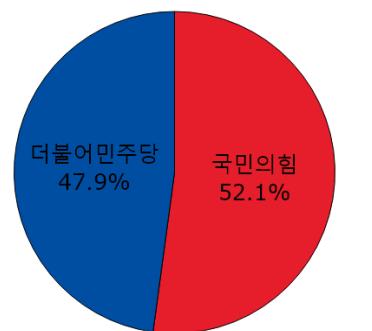
YTN 정치 기사들의 주요 토픽 비율



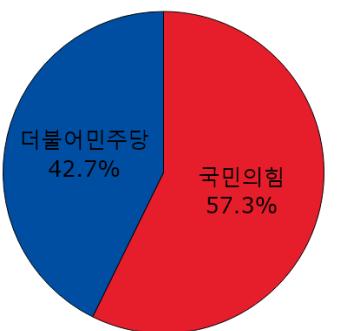
한국경제 정치 기사들의 주요 토픽 비율



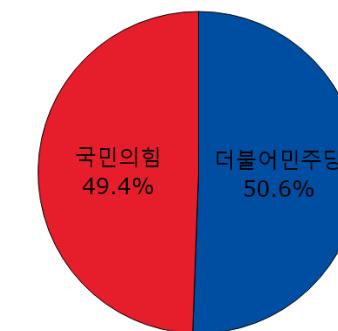
국민일보 정치 기사들의 주요 토픽 비율



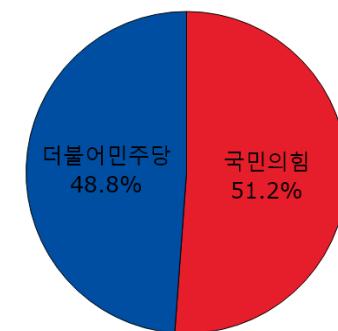
아시아경제 정치 기사들의 주요 토픽 비율



경향신문 정치 기사들의 주요 토픽 비율



매일경제 정치 기사들의 주요 토픽 비율

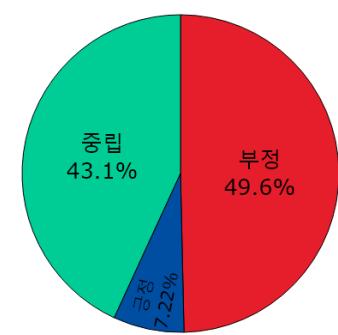


05. 데이터 분석

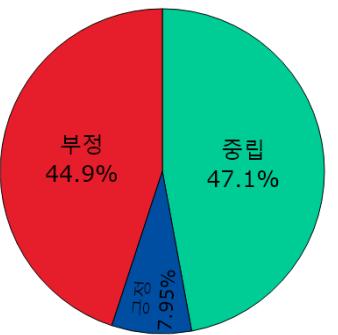
언론사별 긍/부정 비율

언론사별 긍/부정 비율

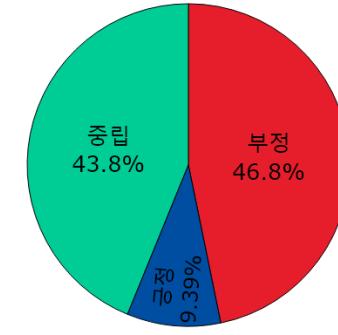
MBC 정치 기사들의 감성 비율



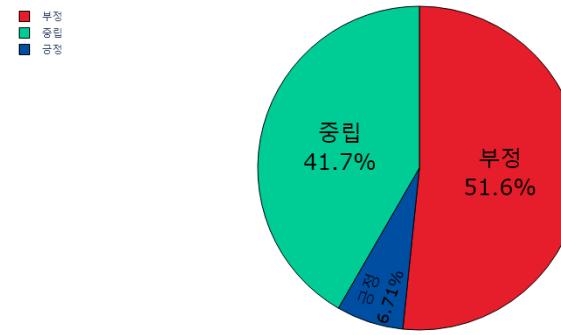
KBS 정치 기사들의 감성 비율



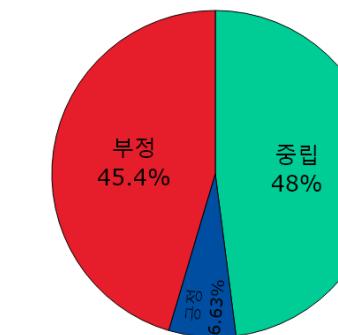
SBS 정치 기사들의 감성 비율



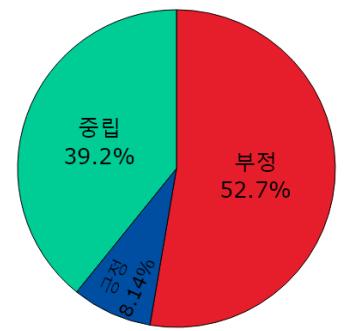
JTBC 정치 기사들의 감성 비율



YTN 정치 기사들의 감성 비율



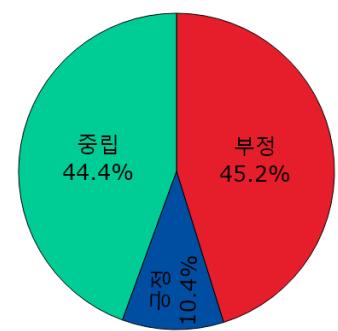
한국경제 정치 기사들의 감성 비율



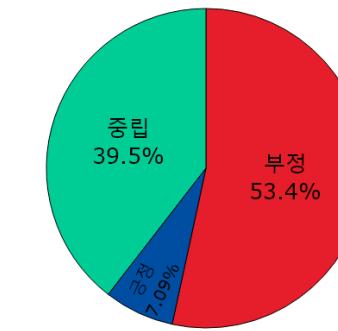
국민일보 정치 기사들의 감성 비율



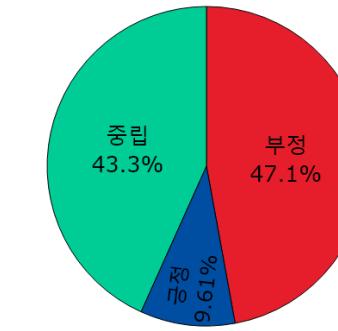
아시아경제 정치 기사들의 감성 비율



경향신문 정치 기사들의 감성 비율



매일경제 정치 기사들의 감성 비율



중립 & 부정의 비율이 압도적으로 큼

05. 데이터 분석

언론사별 정치 편향성 분석

언론사별 정치 편향성 분석

최종 데이터셋

Emotion) 긍정: 1, 중립: 0, 부정: -1

Query	News	Title	Text	Main_topic	Emotion	
0 더불어민주당	아시아경제	[막대기, 당선, 판진, 종권, 네거티브, 민주당, 비판]	[진중권, 전동, 양대, 교수, 이미지, 출처, 연합뉴스, 원본, 아이콘, 아시아,...]	국민의힘	-1	
1 더불어민주당	아시아경제	[공수, 민주당, 사죄, 국민의힘, 여론, 조사, 민심]	[이낙연, 분노, 실망, 아프, 반성, 혁신, 주호영, 여론, 조사, 차이, 민심,...]	국민의힘	0	
2 더불어민주당	아시아경제	[박형준, 안민석, 더불어민주당, 원진, 보유, 튜버, 부산, 지검, 고발]	[국민의힘, 캠프, 부동산, 투기, 전혜, 의원, 후보, 부인, 부동산, 복부인, ...]	국민의힘	-1	
3 더불어민주당	아시아경제	[선거, 지원금, 시작, 민주당, 반전, 계기]	[국회, 본회의, 재난, 지원금, 지급, 규모, 추경, 통과, 윤동주, 기자, 원본...]	더불어민주당	-1	
4 더불어민주당	아시아경제	[민주당, 오세훈, 안철수, 단일, 야합, 스스로, 높이, 욕망, 점철]	[아시아, 경제, 주석, 기자, 더불어민주당, 국민의힘, 국민, 서울, 시장, 후보...]	국민의힘	-1	
...	
88705	더불어민주당	YTN	[주도, 손준성, 이정섭, 검사, 탄핵, 가결, 의회, 폭거]	[앵커, 국회, 오늘, 본회의, 야당, 주도, 손준성, 이정섭, 검사, 탄핵, 추안...]	국민의힘	-1
88706	더불어민주당	YTN	[뉴스, 이슈, 백의종군, 장제원, 불출마, 선언, 증진, 희생]	[진행, 박석원, 앵커, 출연, 김성환, 더불어민주당, 원성, 일종, 국민의힘, 의...]	더불어민주당	0
88707	더불어민주당	YTN	[봉투, 혹정, 송영길, 검찰, 소환, 김건희, 여사, 먼저, 수사]	[재작년, 더불어민주당, 전당, 대회, 봉투, 살포, 의혹, 정점, 지목, 송영길,...]	더불어민주당	-1
88708	더불어민주당	YTN	[여론, 톡톡, 김기현, 책임, 확산, 이낙연, 이준석, 손잡]	[진행, 김영수, 앵커, 출연, 김형준, 배재, 석좌, 교수, 배종찬, 인사이트, ...]	더불어민주당	-1
88709	더불어민주당	YTN	[여야, 총선, 체제, 전환, 속도, 서울, 파장]	[진행, 승휘, 앵커, 출연, 김용남, 국민의힘, 의원, 김종욱, 청와대, 행정관,...]	국민의힘	-1
88710 rows × 6 columns						

05. 데이터 분석

언론사별 정치 편향성 분석

언론사별로 평균 긍/부정 값 산출

언론사별로,
정당별 긍/부정값의 평균

	News	Main_topic	Average_sentiment
0	JTBC	국민의힘	-0.403721
1	JTBC	더불어민주당	-0.407006
2	KBS	국민의힘	-0.340504
3	KBS	더불어민주당	-0.364908
4	MBC	국민의힘	-0.438331
5	MBC	더불어민주당	-0.359028
6	SBS	국민의힘	-0.357143
7	SBS	더불어민주당	-0.327612
8	YTN	국민의힘	-0.387747
9	YTN	더불어민주당	-0.353908
10	경향신문	국민의힘	-0.441344
11	경향신문	더불어민주당	-0.442999
12	국민일보	국민의힘	-0.405405
13	국민일보	더불어민주당	-0.452222
14	매일경제	국민의힘	-0.308397
15	매일경제	더불어민주당	-0.447200
16	아시아경제	국민의힘	-0.296466
17	아시아경제	더불어민주당	-0.361231
18	한국경제	국민의힘	-0.462555
19	한국경제	더불어민주당	-0.474781

05. 데이터 분석

언론사별 정치 편향성 분석

편향성 지표 정의 및 산출

편향성 지표 (Bias, -2~2):
(국민의힘 평균감정) - (더불어민주당 평균감정)

'국민의힘'에 긍정일수록 2에 가까움,
'더불어민주당'에 긍정일수록 -2에 가까움

	News	Main_topic	Average_sentiment
0	JTBC	국민의힘	-0.403721
1	JTBC	더불어민주당	-0.407006
2	KBS	국민의힘	-0.340504
3	KBS	더불어민주당	-0.364908
4	MBC	국민의힘	-0.438331
5	MBC	더불어민주당	-0.359028
6	SBS	국민의힘	-0.357143
7	SBS	더불어민주당	-0.327612
8	YTN	국민의힘	-0.387747
9	YTN	더불어민주당	-0.353908
10	경향신문	국민의힘	-0.441344
11	경향신문	더불어민주당	-0.442999
12	국민일보	국민의힘	-0.405405
13	국민일보	더불어민주당	-0.452222
14	매일경제	국민의힘	-0.308397
15	매일경제	더불어민주당	-0.447200
16	아시아경제	국민의힘	-0.296466
17	아시아경제	더불어민주당	-0.361231
18	한국경제	국민의힘	-0.462555
19	한국경제	더불어민주당	-0.474781

	News	Bias_score
7	매일경제	0.1388
8	아시아경제	0.0648
6	국민일보	0.0468
1	KBS	0.0244
9	한국경제	0.0122
0	JTBC	0.0033
5	경향신문	0.0017
3	SBS	-0.0295
4	YTN	-0.0338
2	MBC	-0.0793

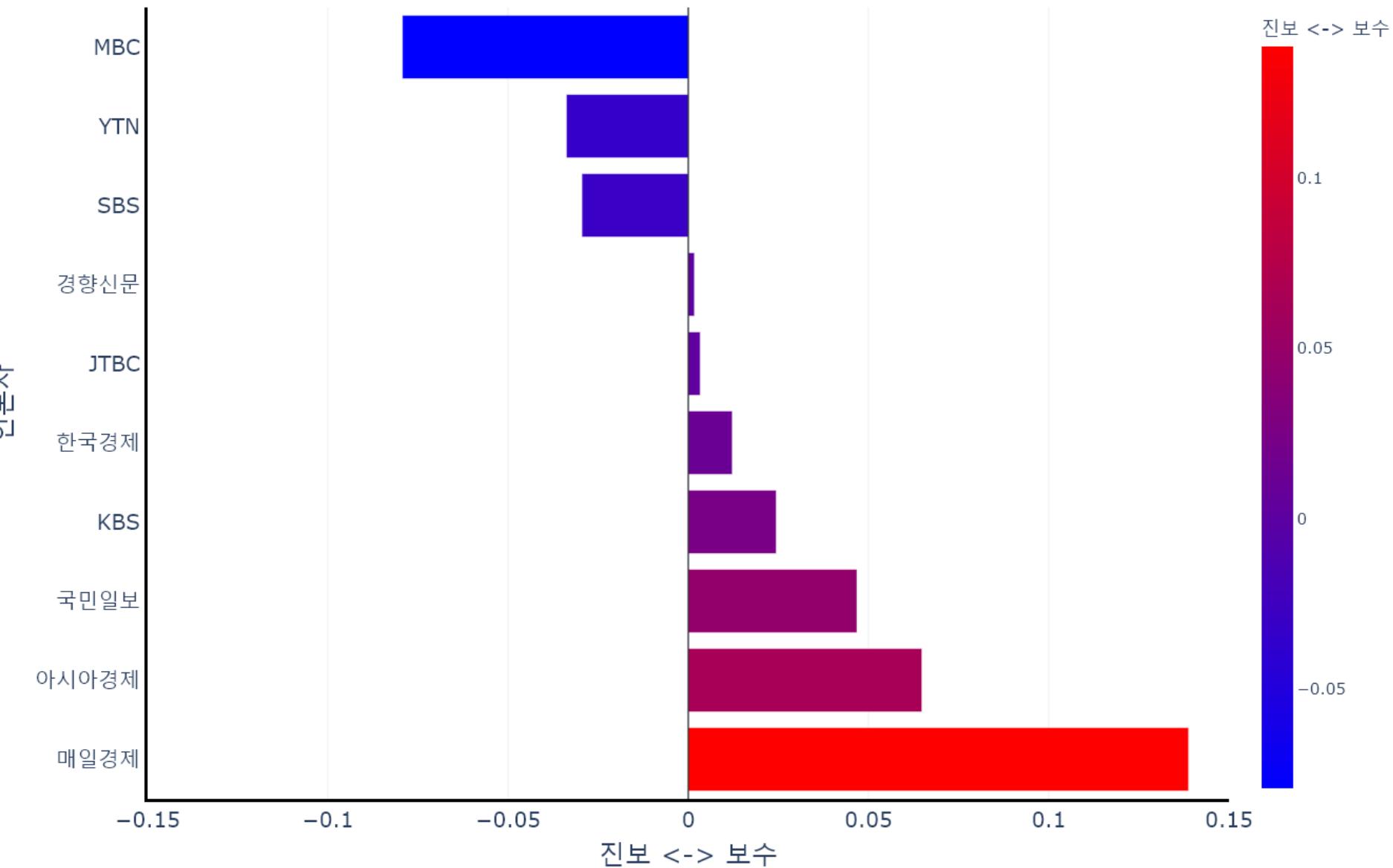


05. 데이터 분석

언론사별 정치 편향성 분석

언론사별 정치 편향성 시각화

언론사별 정치적 편향성 (2020.09.02 ~ 2023.12.31)



6. 향후 과제

개선점

06. 향후 과제

개선점(1). 토픽모델링

토픽 모델링의 정확도

선행한 3가지의 접근
방식 모두
정확도가 높지 않음

사람이 직접 라벨링을
하는 것이 가장 정확하
다고 추측

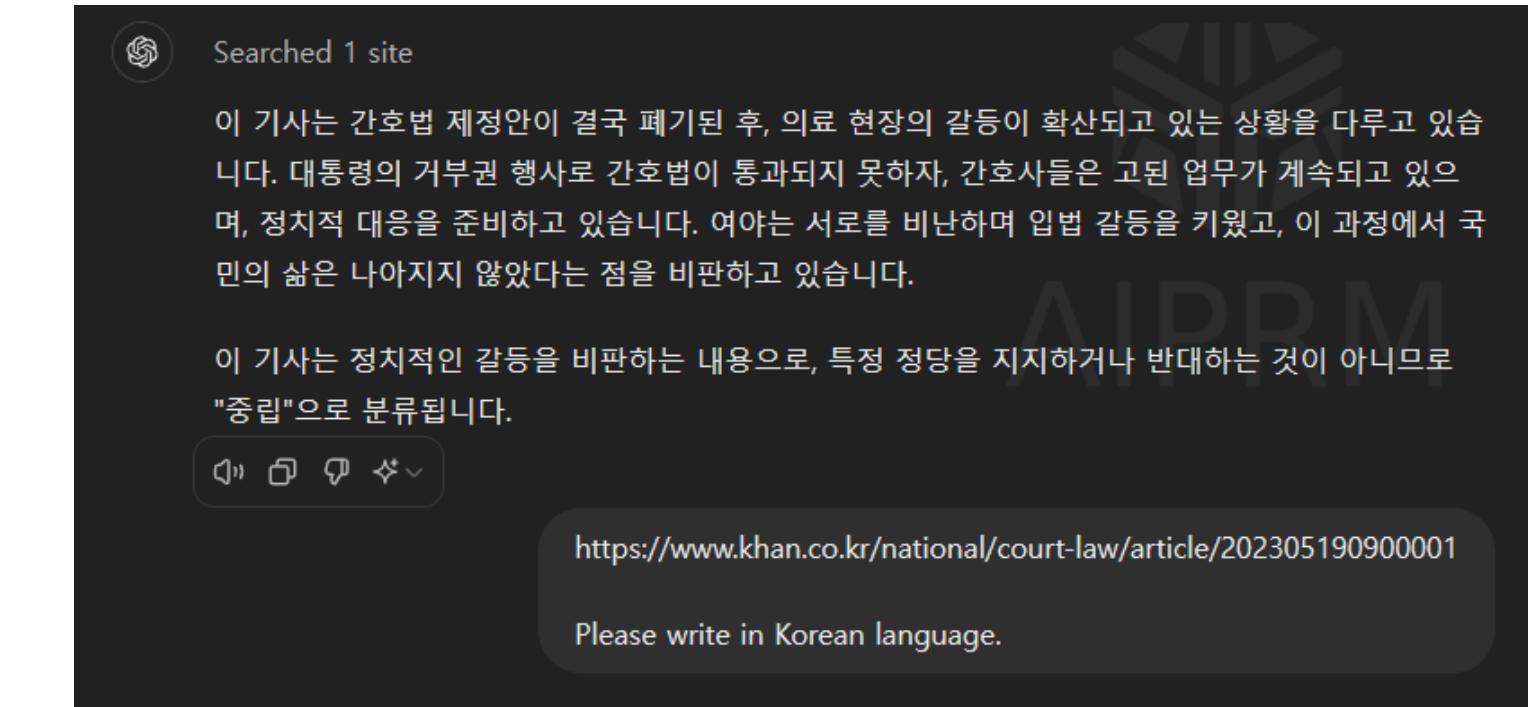
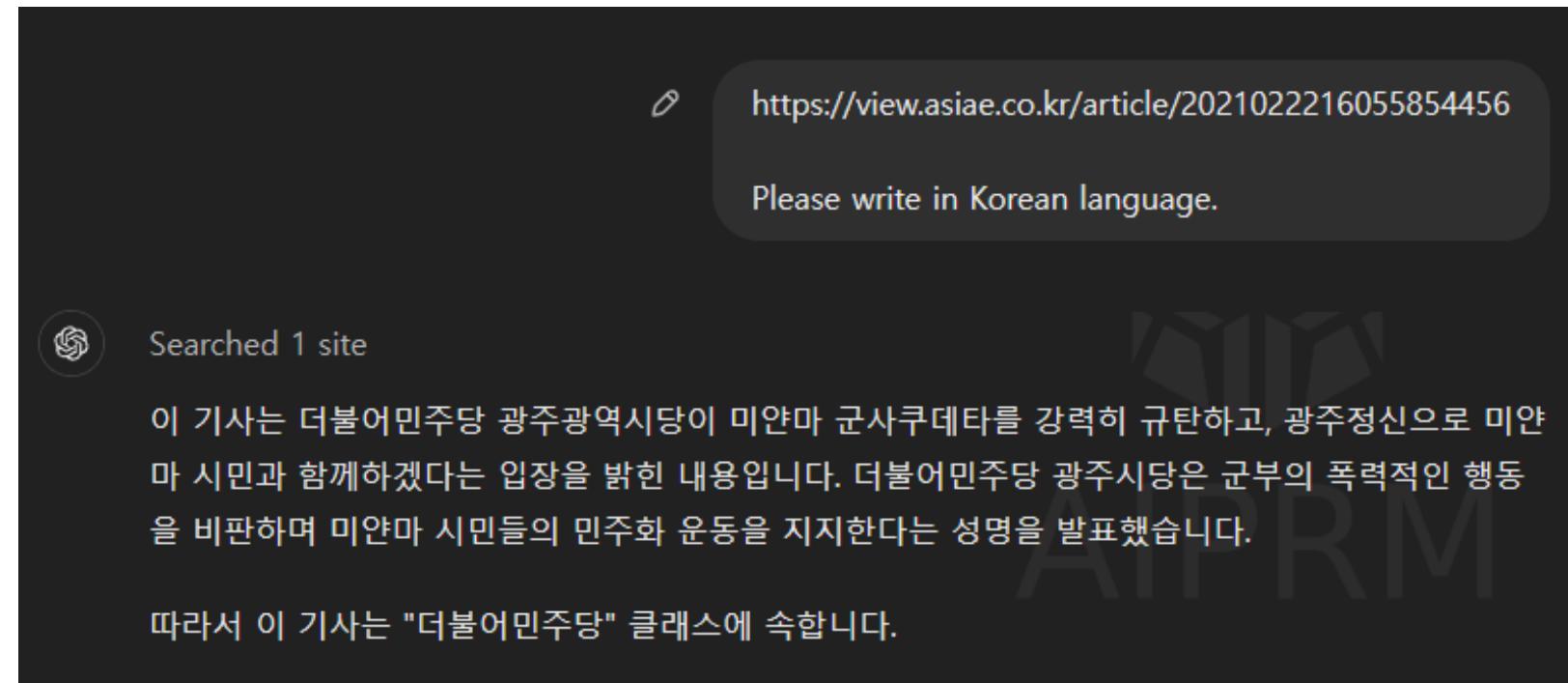
하지만 180,000개의
데이터를 모두 라벨링
하는 것은 현실적으로
불가능

토픽 모델링을 LLM에게 맡기는 것을 어떨까?

06. 향후 과제

개선점(1). 토픽모델링

LLM을 통한 토픽 모델링(chatGPT 4o)



- **사람보다 더 객관적으로 분류한다고 느껴짐**
- GPT API를 사용해 토픽 모델링하는 것도 하나의 방법으로 생각됨
→ 선행연구) TopicGPT: A Prompt-based Topic Modeling Framework, NAACL 2024 전통적인 방법(LDA 등)보다 여러 방면에서 뛰어나다고 함
- 오픈소스의 LLM(ex. Lamma3) 등을 활용해보자!

06. 향후 과제

개선점(2). 토픽 모델링과 감성분석의 연결

토픽 모델링 결과, 감성분석 결과 연결한 해석의 오류?

해석의 오류

토픽 모델링은 주제만을 추출하고,
감성분석은 감정만을 분석하므로
두 분석 결과를 직접적으로 연결하면
잘못된 해석을 유발할 수 있다고 생각됨

모델의 한계

두 분석 모델 모두 완벽하지 않으며,
오류가 발생할 수 있음.
이를 무비판적으로 받아들일 경우
잘못된 결론을 도출할 위험



토픽 모델링에 기반한 감성분석에 대한 연구가 필요

7. 진행 프로세스

개발 일정

작업 분담

07. 진행 프로세스

개발일정

아이디어 공유,
언론사 별 뉴스 기사크롤링



토픽모델링, 감성분석



토픽모델링 비교,
코드 병합, PPT



데이터셋 병합,
전처리



토픽모델링, PPT

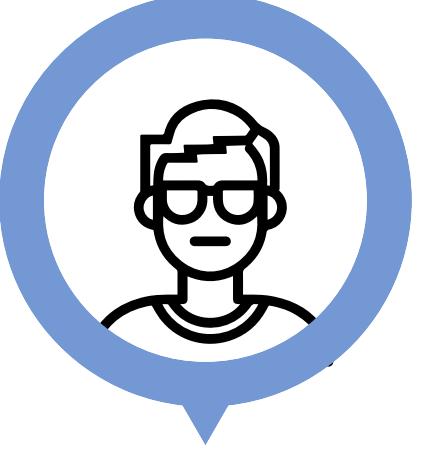
07. 진행 프로세스

작업 분담



정하성

- 팀장
- 회의록 작성
- 크롤링
- 토크나이저 비교
- 중복값 처리
- 토큰화
- 품사태깅
- 불용어 처리
- 빈도분석
- 6품사 추출
- KeyBert
- KeyBert 성능 비교
- bart-large-mnli
- bart-large-mnli 성능 비교
- 파이차트 생성
- 편향성 평가지표 수립
- 편향성 결과 산출
- PPT
- 코드 병합 완성
- 발표



최민규

- 크롤링
- 품사태깅
- 빈도분석
- 감성사전 비교
- 감성분석
- 감성점수 구간평가
- 코드 병합



김민주

- 회의록 작성
- 크롤링
- 데이터 병합
- 결측값 처리
- 한글화
- LDA
- LDA 결과 해석
- TextRank
- TextRank 정확도 비교
- 데이터 라벨링
- TextRank 성능 비교
- 코드 병합
- PPT