

한국어 형태소 분석을 기반으로 한 실시간 뉴스 분류 및 요약

코드스테이츠 인공지능 01기

최중훈

2021. 06. 11.

발표순서

1

기획 배경

| 실시간 뉴스를 간략하게 볼 수 없을까?

2

데이터 분석

| EDA 및 데이터 불균형 처리

3

모델 선정 및 비교

| 사용 패키지, Naive-Bayse, BiLSTM

4

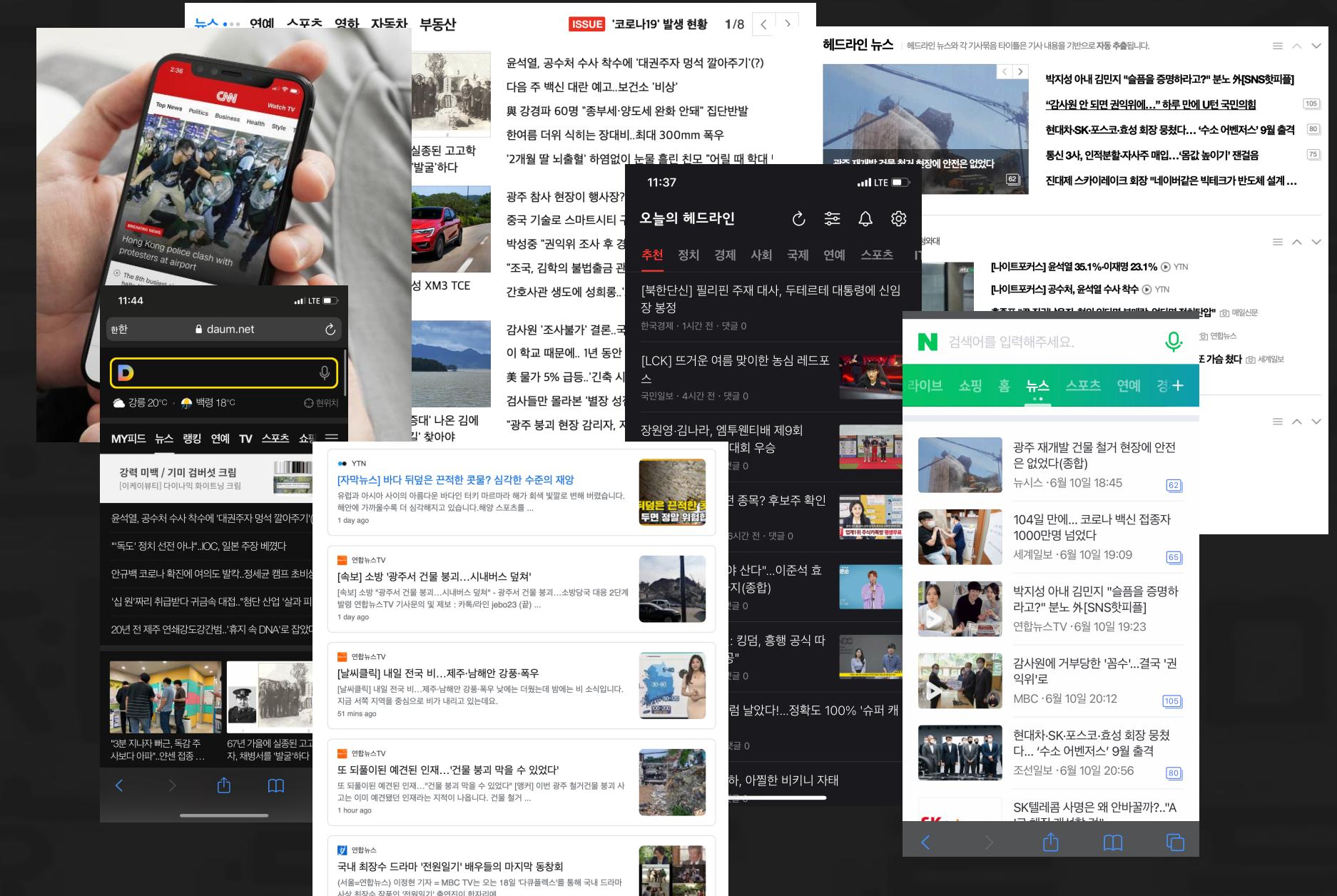
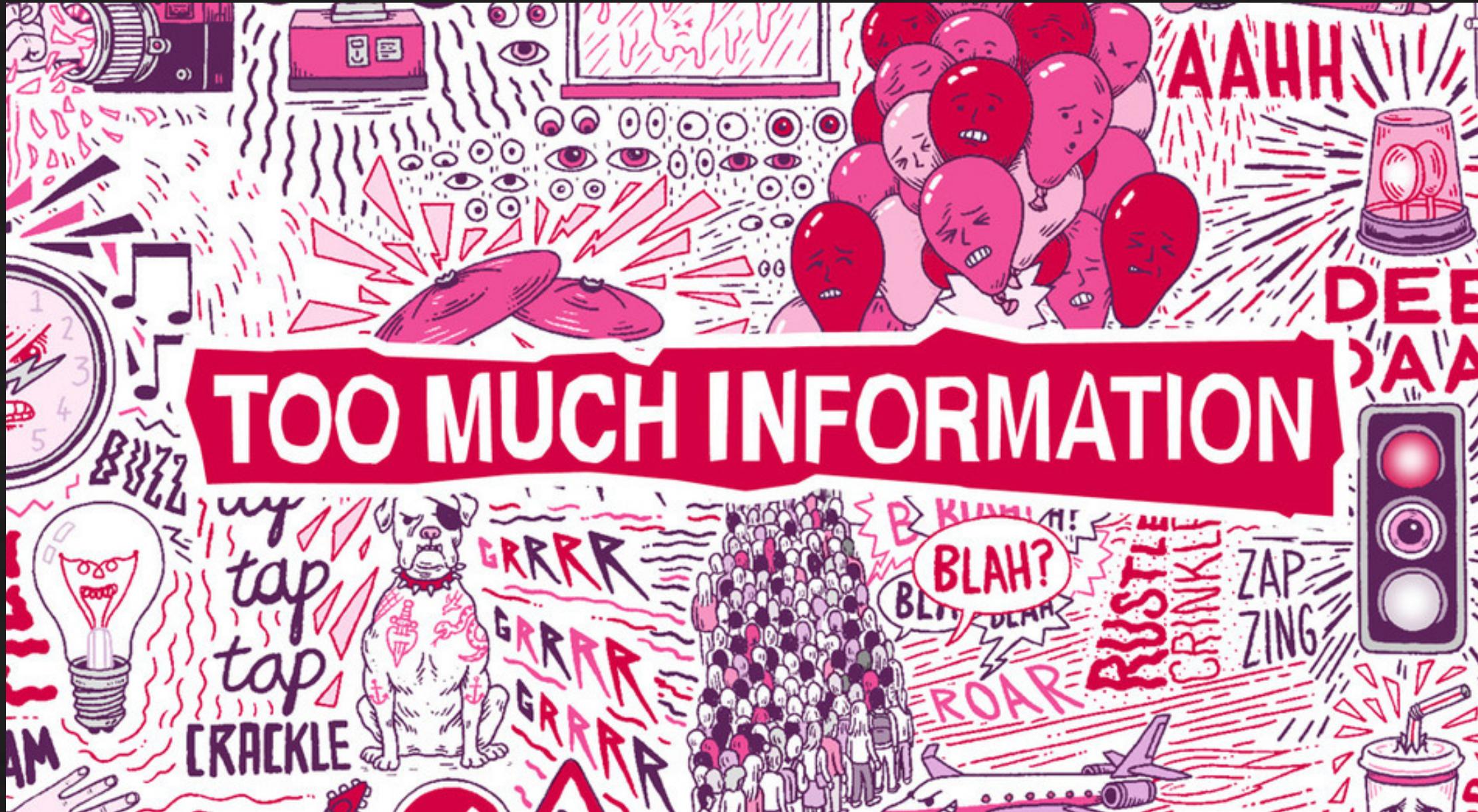
서비스 파이프라인

| 크롤링 -> 전처리 -> 모델 -> 출력

5

개선사항

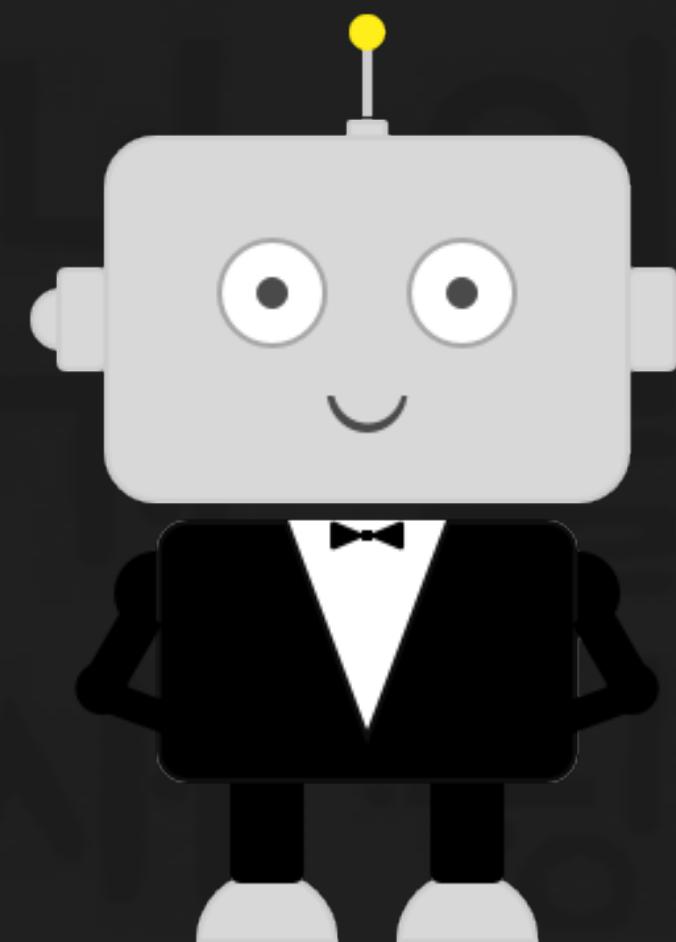
| DL 성능개선, 문서요약 등



- ▶ 국내 신문, 뉴스, 잡지 등 정기간행물 등록사는 23,000여 개 (문화체육관광부 정기간행물 등록시스템 기준)
- ▶ 각종 웹사이트 혹은 어플리케이션을 통해 어디서든 접할 수 있는 정보 매체
- ▶ 기사의 전반 내용을 가늠할 수 없는 자극적인 헤드라인



Q 독자의 편의성을 고려하여 실시간 뉴스를 제공할 수 없을까?



실시간 뉴스의 요약본을 확인하세요.
뉴스봇이 도와드릴게요!



데이터 분석

약 19,000여 개의 인터넷 뉴스 데이터 (한국과학기술정보연구원과 충남대학교가 공동 개발한 정보검색시스템 평가를 위한 한글 테스트 컬렉션)

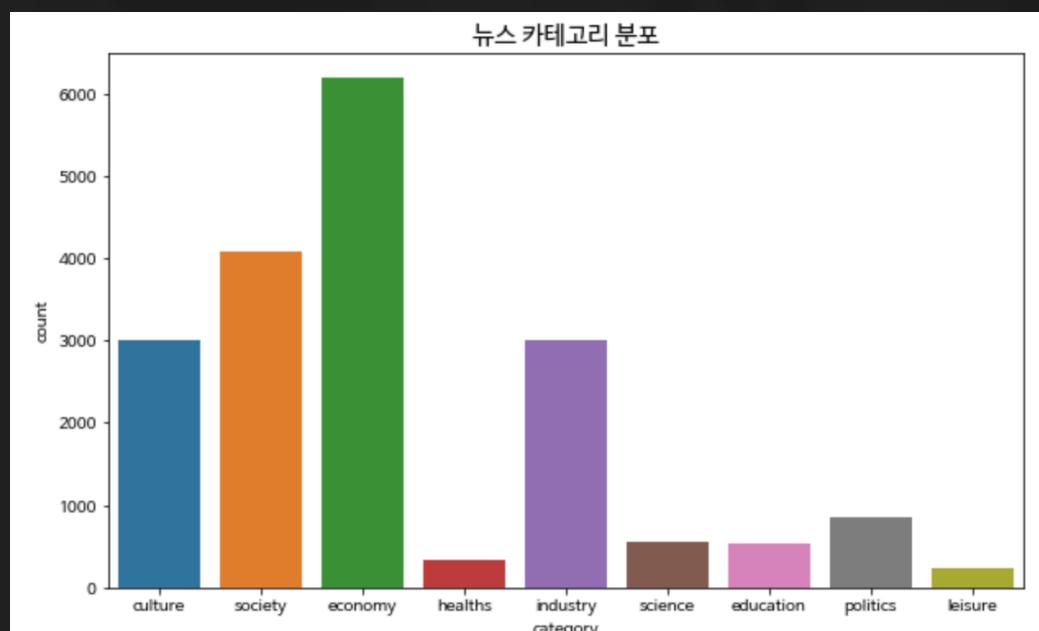
What did

실제 기사를 읽어 보았을 때, 경제/산업 과 같이 사람도 구분을 두기 애매한 경우
혹은 데이터 수가 현저히 적은 카테고리를 병합

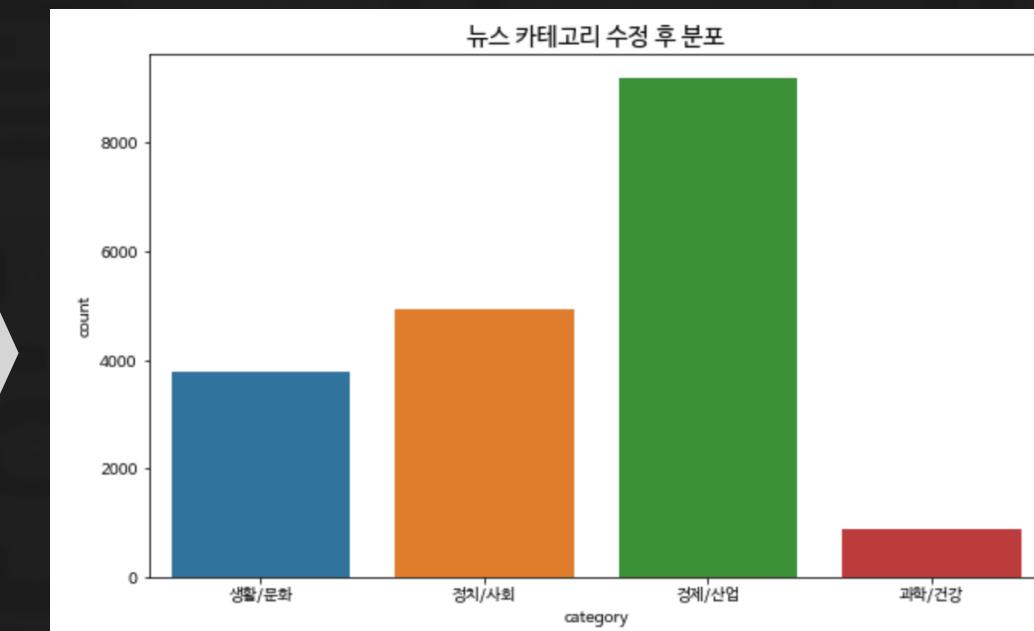
뉴스 기사와 같이 정확한 맞춤법과 어휘로 구성된 텍스트의 경우
이중번역을 통해 데이터 증강 (ex. 한국어 > 영어 > 한국어, 한국어 > 중국어 > 한국어)

자동화 과정에서 지속적인 오류로 인한 증강 작업 중단
정말 적은 데이터를 증강시켰음에도 모델 성능은 0.1 이상 개선되는 것을 확인

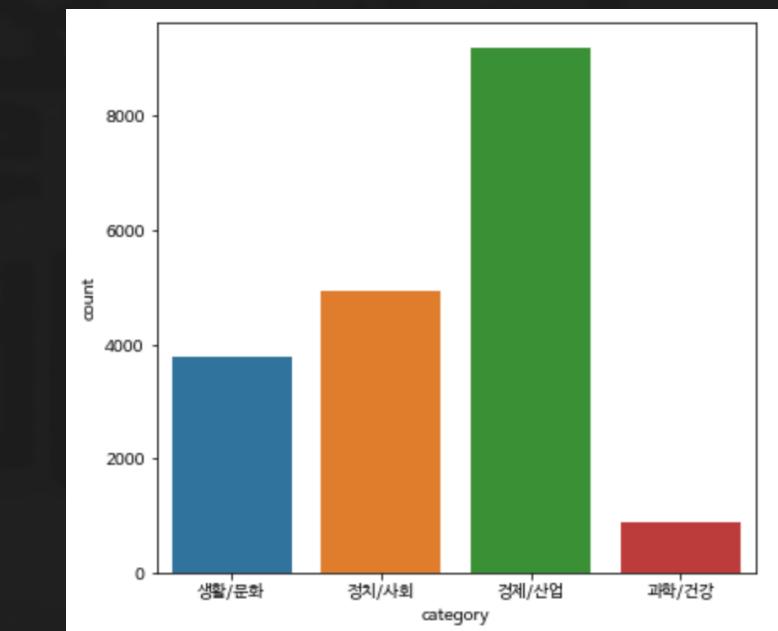
* 카테고리 병합



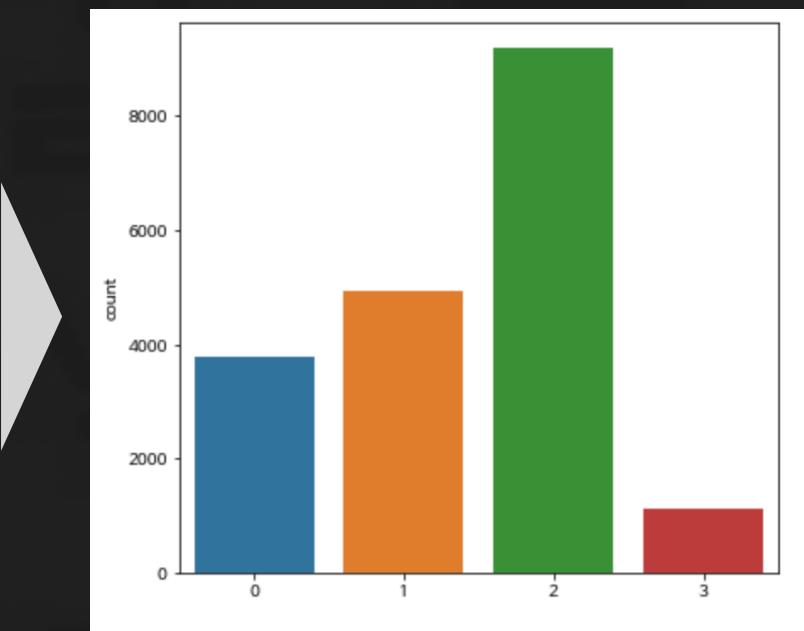
기존 카테고리 9개 >> 4개로 병합
(경제/산업, 사회/정치, 생활/문화, 과학/건강)



* 데이터 증강



매우 불균형한 과학/건강 데이터를 879개에서 1118개로 증강



* 기사 내 어휘 시각화 (명사 기준)

: konlpy MeCab 사전 사용



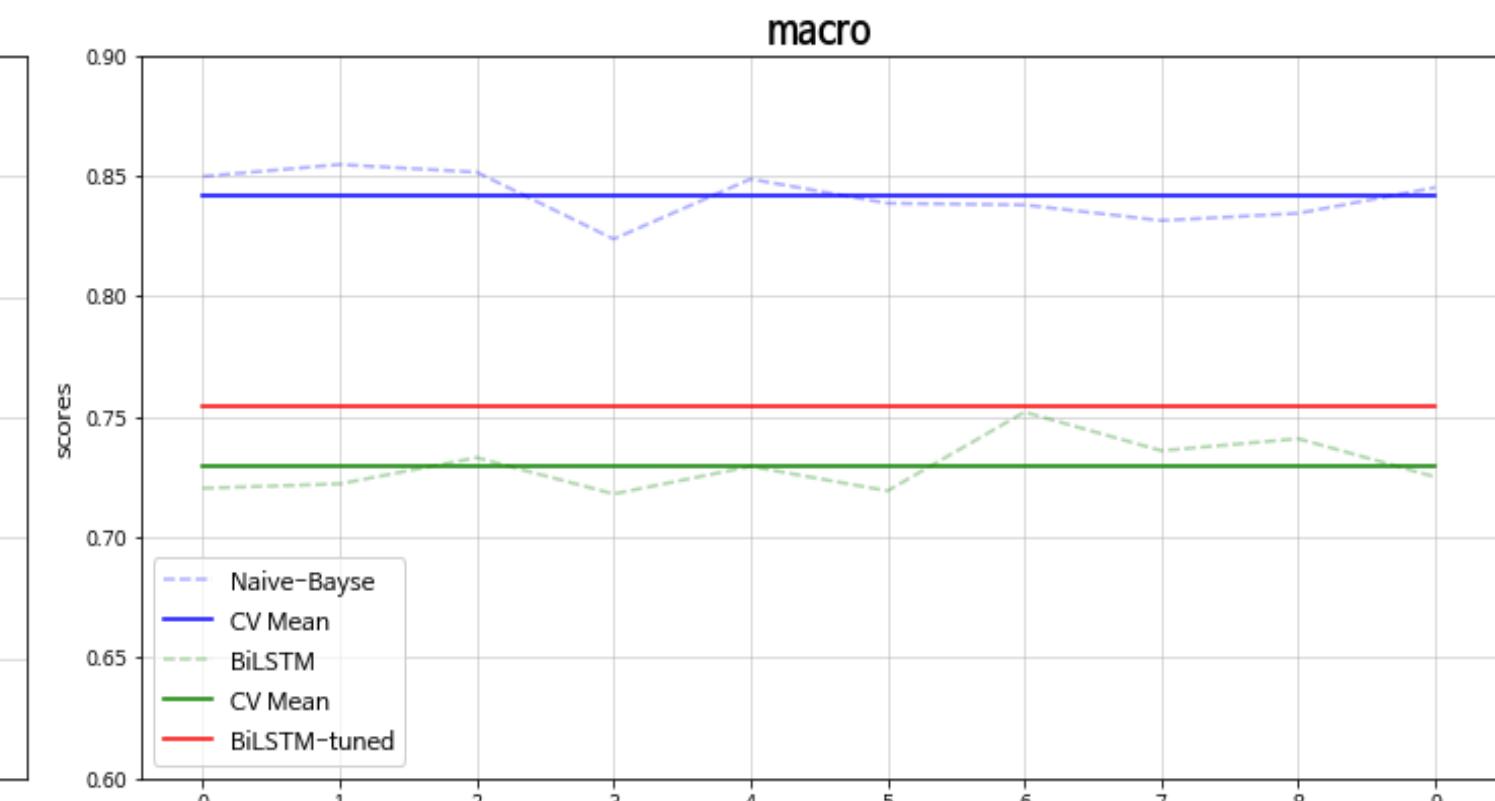
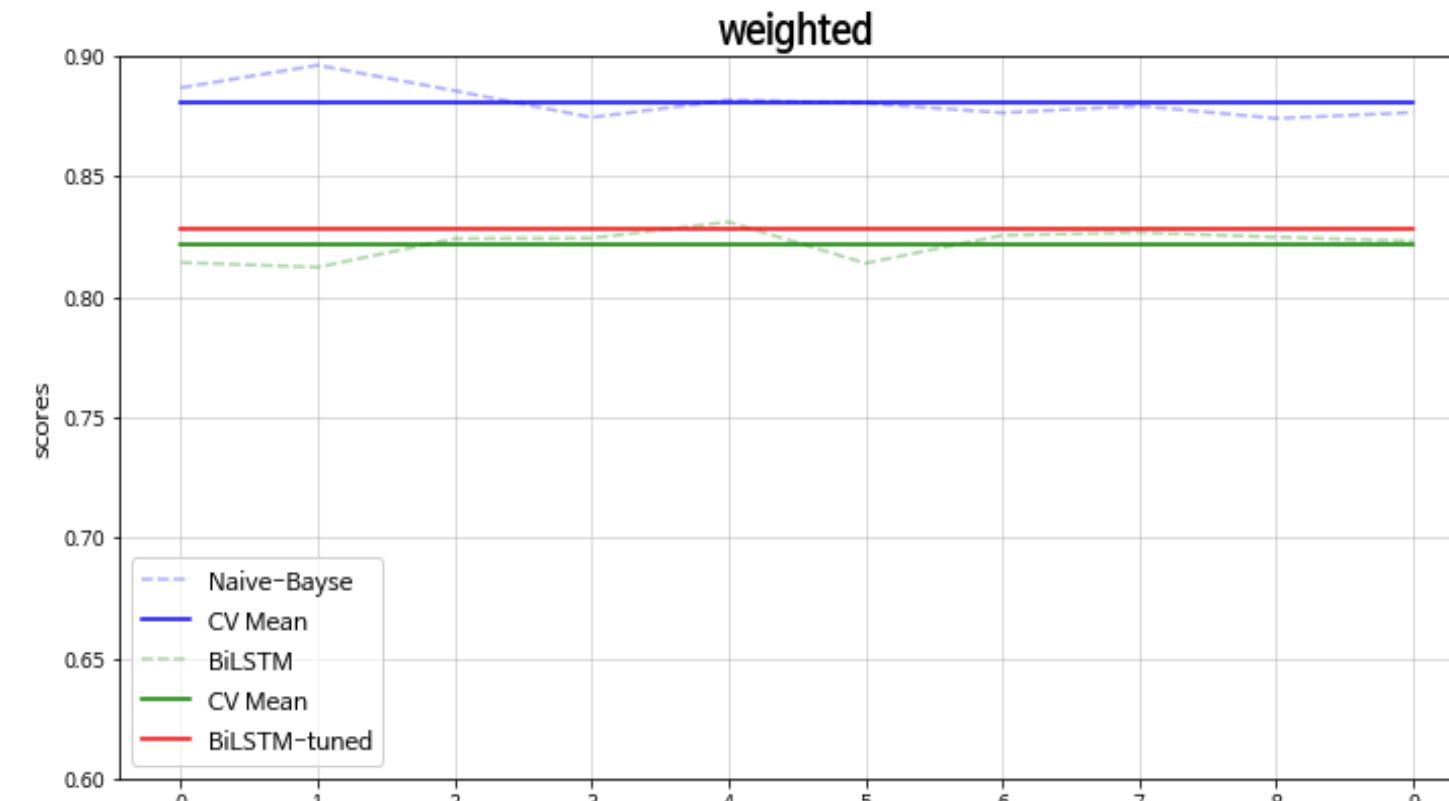
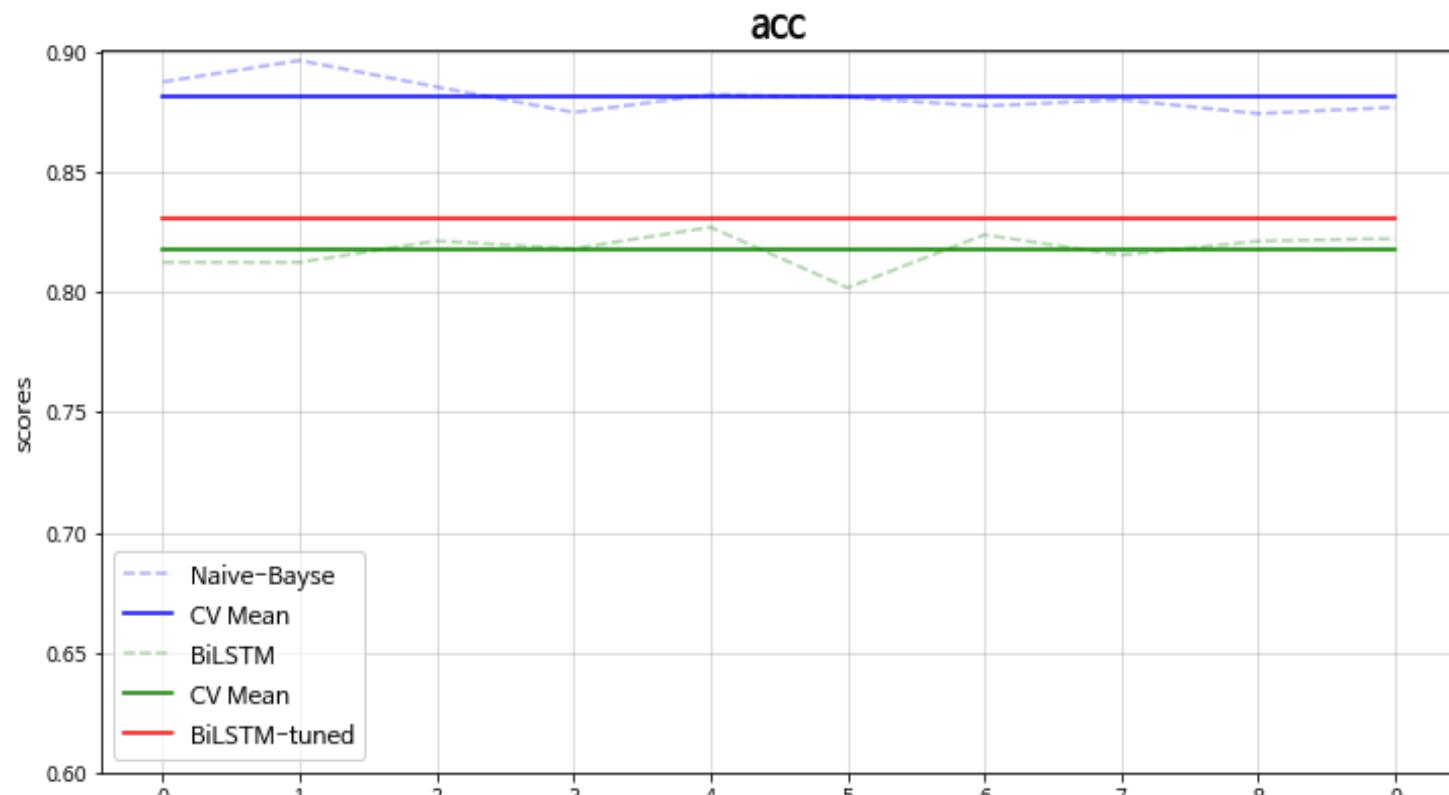


분류모델

- ML 모델: Naive-Bayse
- DL 모델: Bidirectional LSTM

Results

- Naive-Bayse ML 모델과 튜닝 전 후 BiLSTM 세 개의 모델의 성능 비교
- 불균형한 데이터를 고려하여 f1 macro average 를 중점적으로 모델 성능 개선
- 모든 성능 지표에서 기본 ML 모델이 더 뛰어난 성능을 보임



요약모델

- 오픈소스 패키지 사용
- Pororo from Kakaobrain



kakaobrain



1. 실시간 기사 크롤링

- 실시간 네이버 랭킹 뉴스
- BeautifulSoup을 사용한 크롤링 함수 구현

2. 텍스트 전처리

- Konlp Mecab 형태소 분석기 사용
- 각 모델에 맞는 데이터 구축
- ML model: TF-IDF vector
- DL model: Padding Sequence

3. 모델 준비

- 학습된 모델 load
- 가중치 load
- 오픈소스 패키지 load

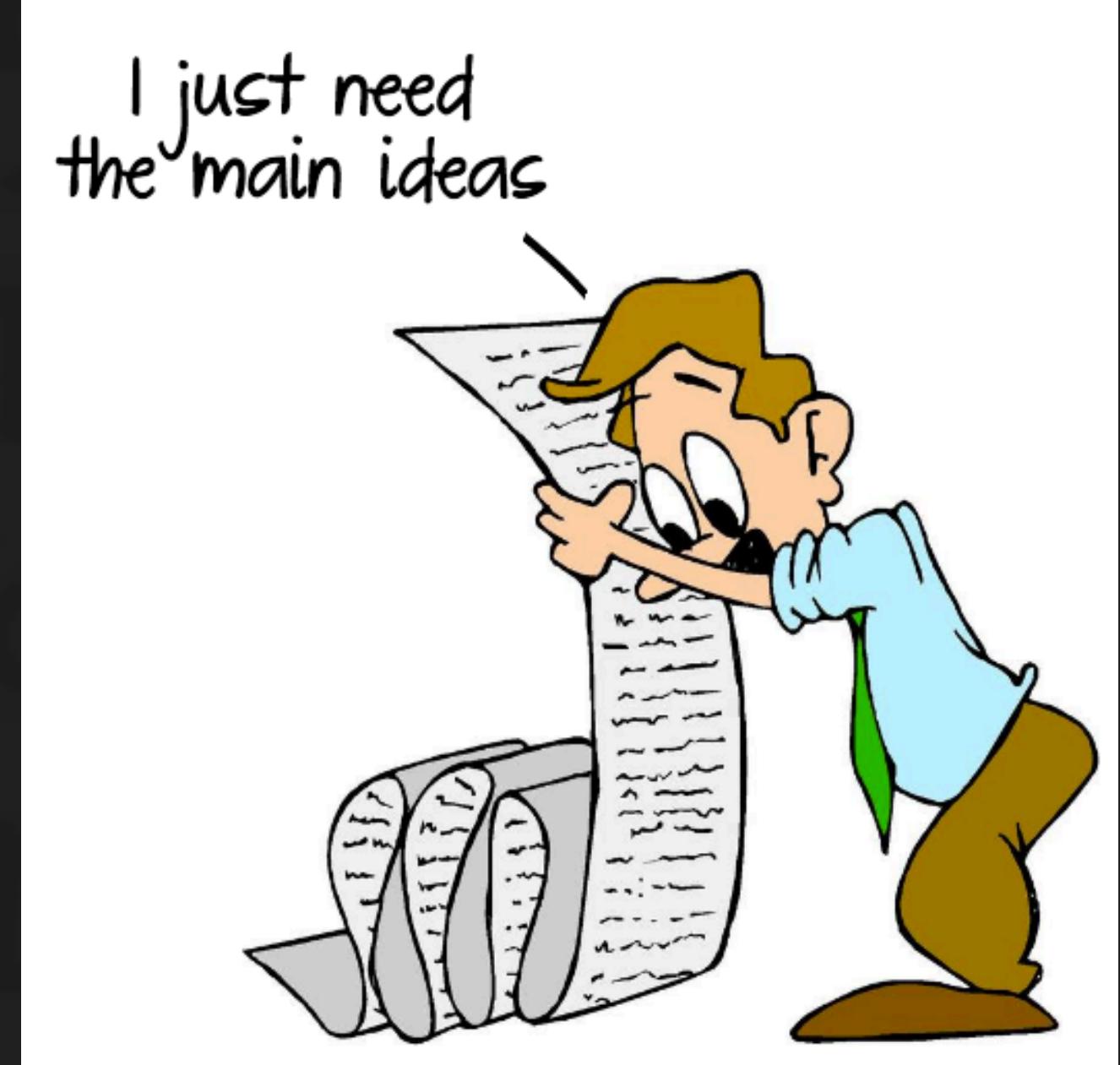
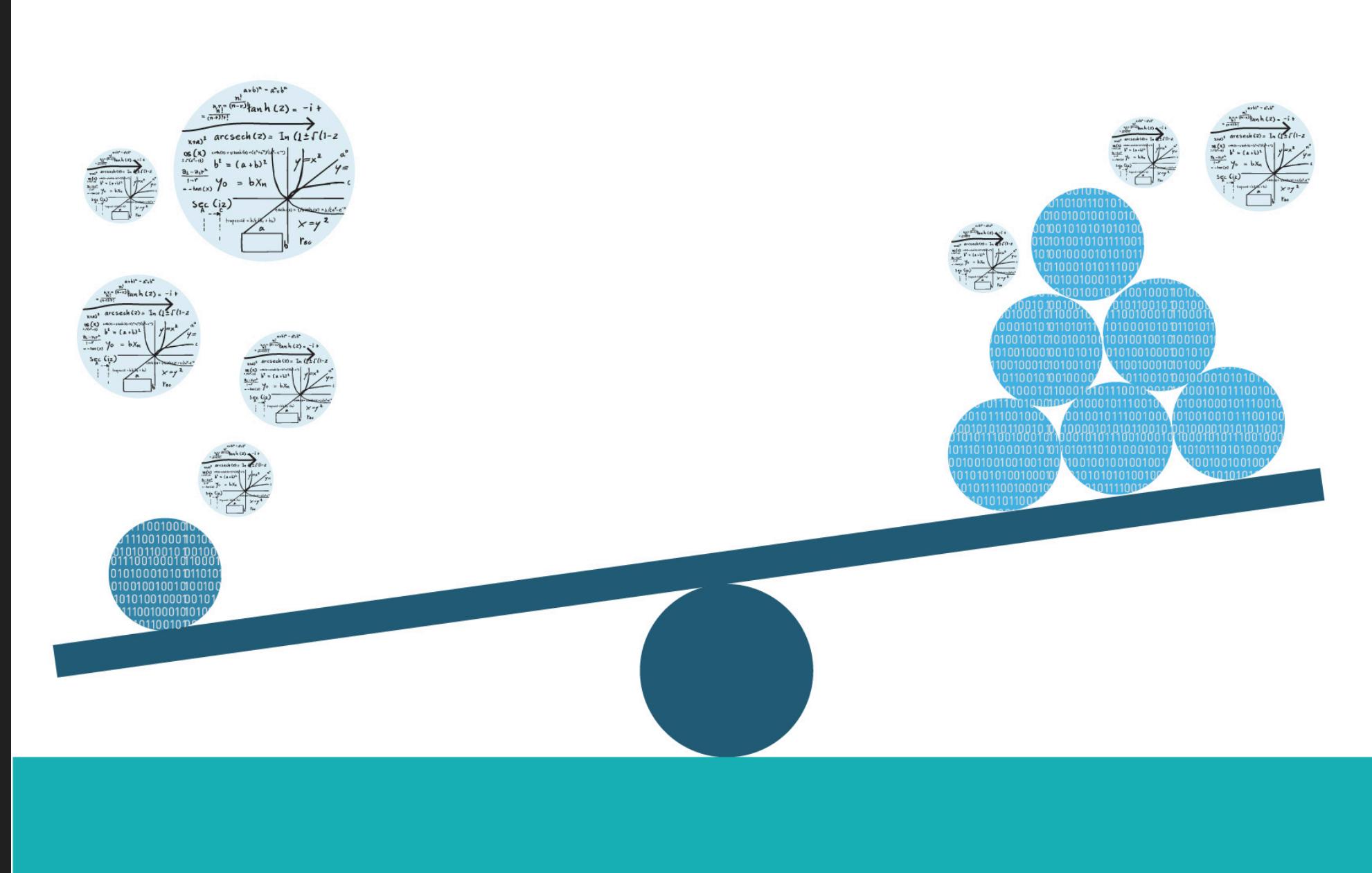
4. 결과 출력

- 기사 원본
- 예측 카테고리
- 기사 요약본

```
41 |     pred_class_lst_nb, pred_proba_lst_nb,
42 |     pred_class_lst_bi, pred_proba_lst_bi,
43 |     abs_summary_lst, ext_summary_lst)
44 |
45 |     def clear_all(b):
46 |         with output:
47 |             clear_output()
48 |             print('='*150)
49 |
50 |
51 |
52 |     display(newsnum, model_select, summarizer_select, getnews, clear, output)
53 |     getnews.on_click(getnews_click)
54 |     clear.on_click(clear_all)
```

1 widget()

✓ 0초 오후 1:24에 완료됨



▶ 데이터 증강을 사용해서 불균형 처리를 완벽히 해줄 수 있지 않을까?

▶ 기사 카테고리를 나누는 분류 문제에서 딥러닝 모델은 비효율적인가?

▶ 오픈소스 패키지가 아닌 직접 요약모델을 구축해 볼 수 있을까?

감사합니다.