

**필터버블을 해소하는 뉴스레터**

# **BUBBLEPOP**

김성환 김소윤 김예찬 최수현

# CONTENT

- |
  - | 팀 소개
  - | 문제 정의 및 원인 분석
  - | 프로젝트 파이프라인
  - | 결론 및 한계점
  - | 시연

# INTRODUCTION

- 1) 팀 소개
- 2) 문제 정의 및 원인 분석

# ABOUT US

김성환

Model, Data

김소윤

Model, FrontEnd



김예찬

DB, BackEnd

최수현

Model, FrontEnd

매주 일요일 정기회의

만나면 6시간은 기본..

## 문제 정의 및 원인 분석

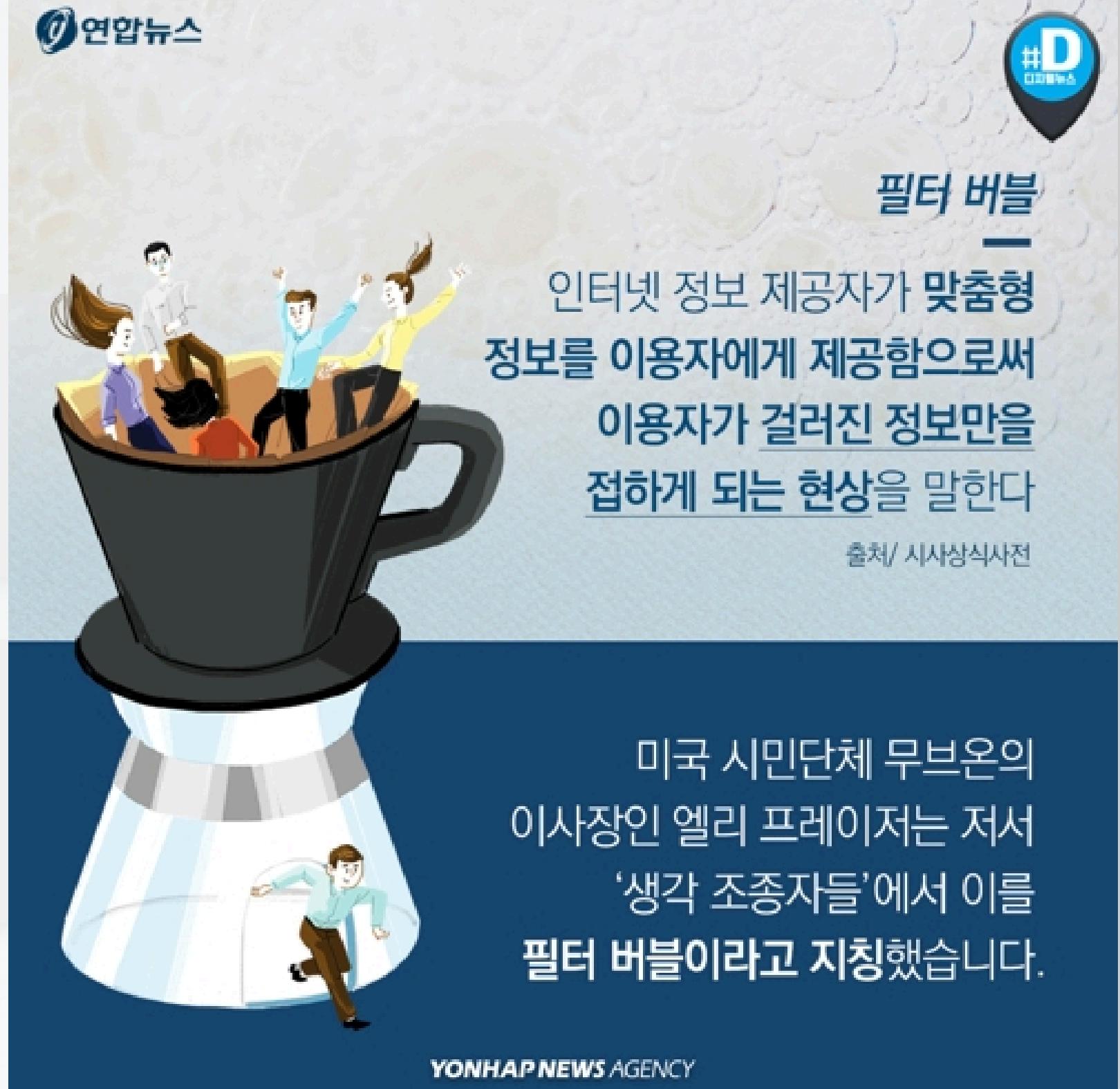
### 뉴스를 볼 시간이 없는 바쁜 현대인 초천 기사나 봐볼까..?



바쁘다 바빠 현대사회에 사는 현대인들은  
뉴스를 직접 고를 시간이 없어요  
뉴스 포털이 상단에 띄워주는 뉴스를 읽어요

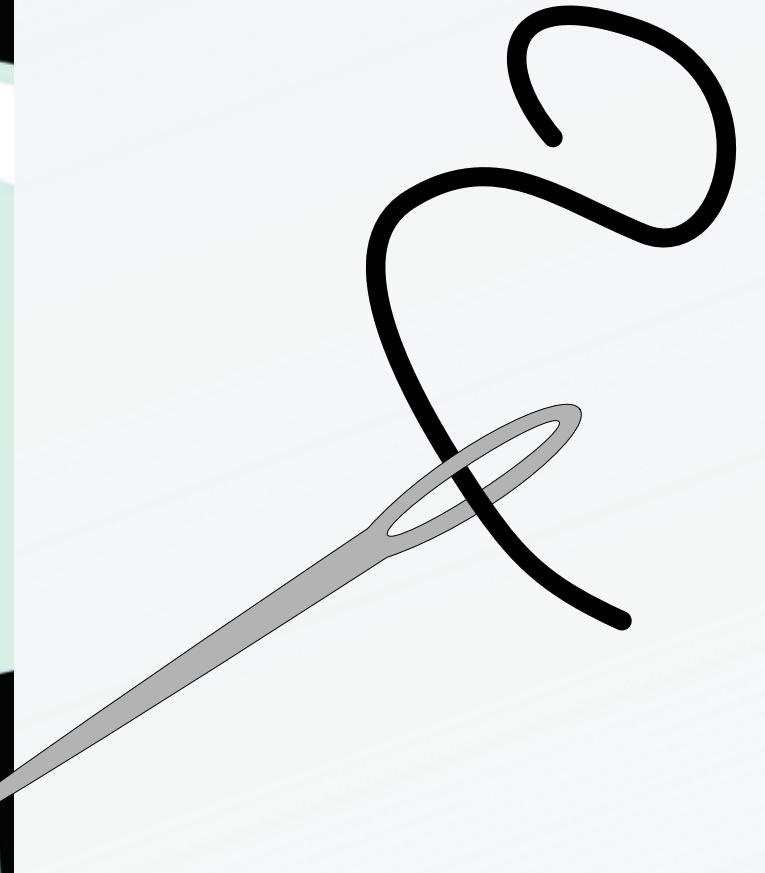


뉴스 포털들은 바쁜 현대인들을 위해  
맞춤형 정보를 제공해요  
이런 방식이 편리하지만, 뉴스 소비자들은  
온연중에 정파성을 고착화하게 될지도 몰라요



미국 시민단체 무브온의  
이사장인 엘리 프레이저는 저서  
'생각 조종자들'에서 이를  
필터 버블이라고 자칭했습니다.

YONHAP NEWS AGENCY



그래서 필터버블 해소 뉴스레터를 만듭니다

필터버블을 터트리는 뉴스레터  
**BUBBLEPOP NEWSLETTER**

# 문제 정의 및 원인 분석

**AS IS**

바쁜 현대인들은 직접 뉴스를  
비교해가며 볼 시간이 없어요

**TO BE**

딥러닝 모델로 비교할 뉴스를 알아서  
제공해주니  
직접 비교해서 볼 필요가 없어요

뉴스 포털 상단에 띄워진 뉴스만 읽게  
되다보니 비슷한 뉴스만 계속 읽게 돼요

사용자들은 평소에는 읽을 일이  
거의 없는 새로운 성향의 뉴스를 읽게 돼요

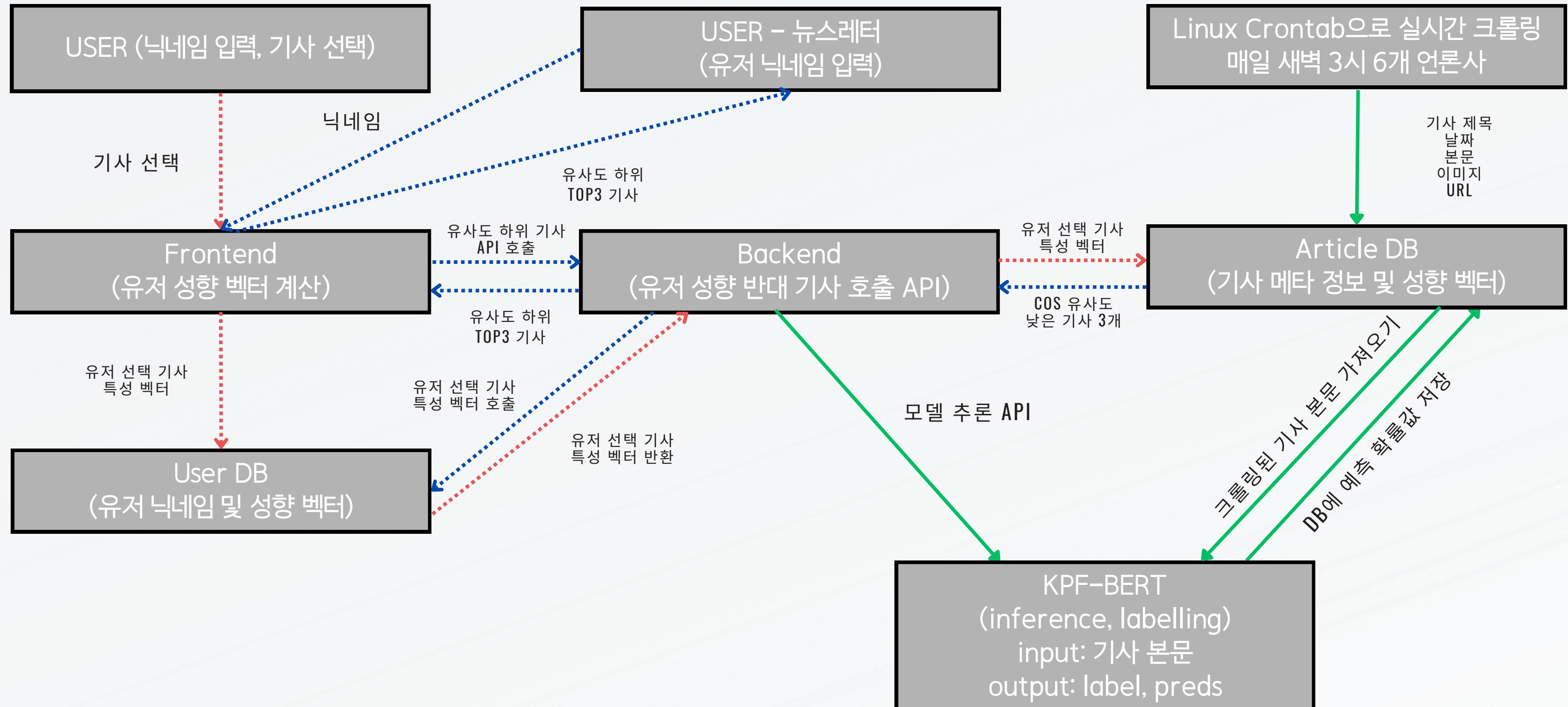
언론사들은 타겟유저의 니즈를 만족시켜야하니  
계속해서 정파성이 짙은 기사를 작성해요

뉴스 유목민들로 유저 페르소나가 확장되면  
언론사들은 건강한 저널리즘을 위해  
다른 면으로 노력해볼 수 있어요

# PROJECT PIPELINE

- 1) Data collection and Preprocessing
- 2) Modeling
- 3) Data Base and Back-End
- 4) Front-End

# PROJECT PIPELINE



# DATA COLLECTION



## Bigkinds



- 뉴스수집시스템, 분석시스템, 저장시스템 등으로 구성돼 있으며, 저장된 뉴스 분석 정보는 국민, 언론사, 학계, 스타트업 등이 활용할 수 있는 **뉴스빅데이터 분석서비스**.

정형화된 데이터

빅데이터화

가치있는 정보

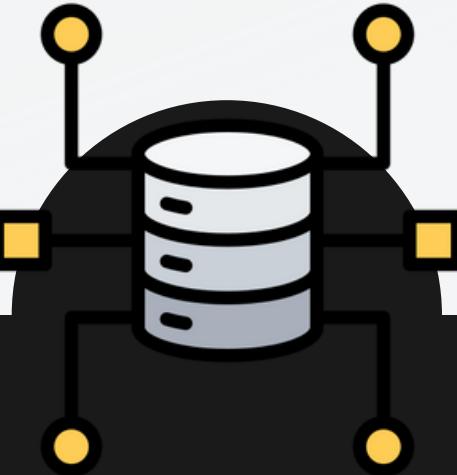


## Selenium



- 웹 애플리케이션을 자동으로 테스트하고 제어할 수 있는 오픈 소스 도구
- 웹 브라우저를 자동화하여 사용자의 행동을 시뮬레이션을 할 수 있다.
- 웹 애플리케이션의 기능 테스트, 크롤링, 스크래핑 등에 사용

# DATA COLLECTION



- 모델 학습을 위한 데이터셋 구축
- 성향이 명확하게 드러나는 사설 뉴스 수집
- 100,000개 뉴스를 크롤링 후 전처리

학습 데이터 크롤링



- 최신 정보를 제공하기 위한 데이터 수집
- 필터버블을 해소하는 동시에 소비자 만족도도 향상 시킨다.

최신 데이터 크롤링

# DATA PREPROCESSING

## DATA LABELING

보수 - 0:

- 조선, 중앙, 동아

진보 - 1:

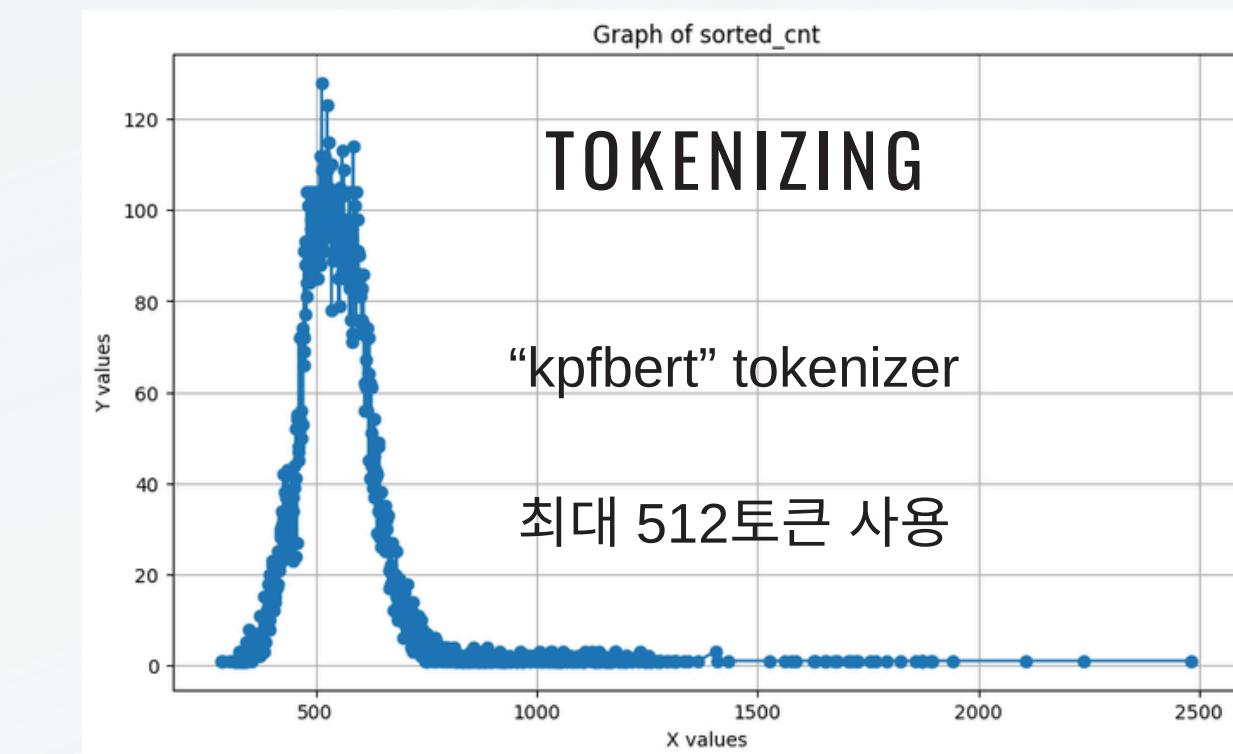
- 한겨레, 경향

STEP 1

## PREPROCESSING

불필요한 정보를  
정규표현식을 통해 제거

STEP 2



STEP 3

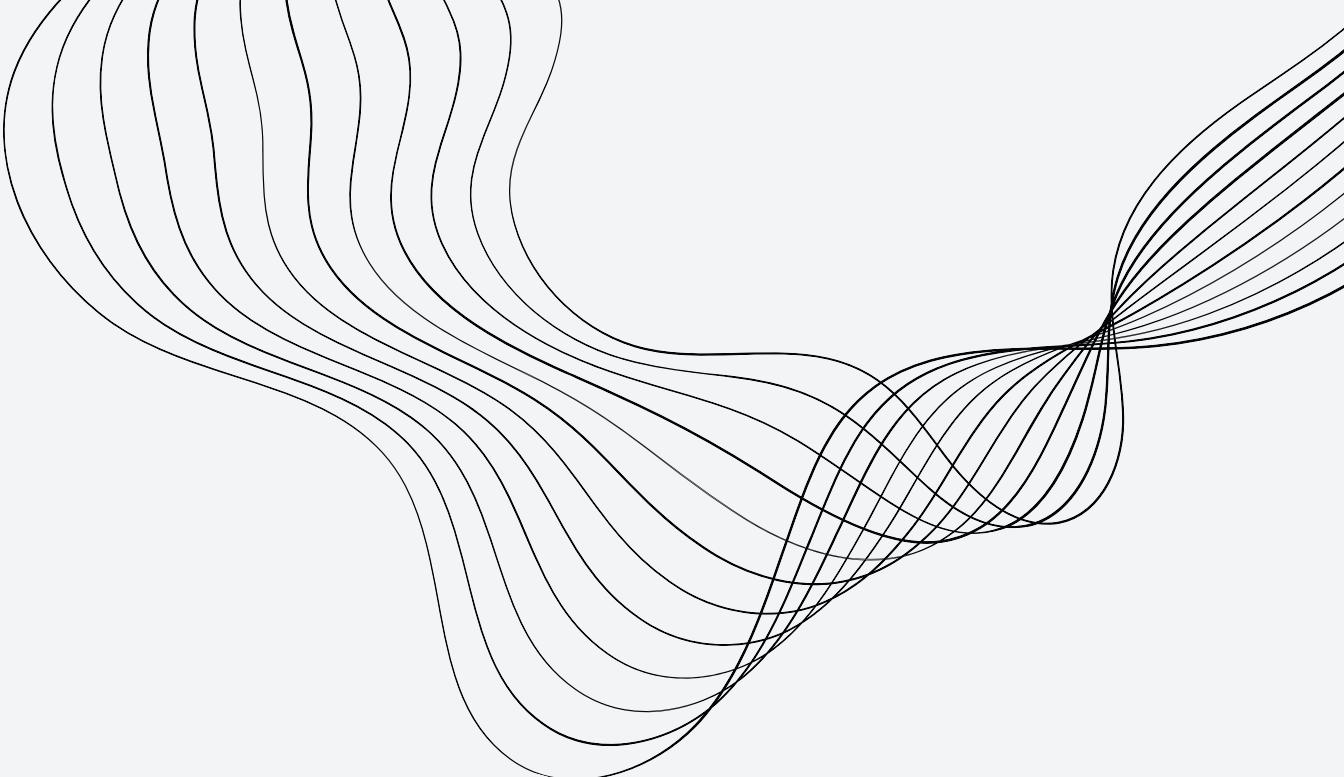
# MODEL

## KPF-BERT

- 한국어 데이터에 강한 모델
- 몇 십년 간 축적된 뉴스데이터로 학습한 모델
- 정제된 한국어 표현에 높은 성능

## OUR PROJECT

- 뉴스 기사 본문을 학습데이터로 사용
- 뉴스 기사를 진보와 보수로 분류

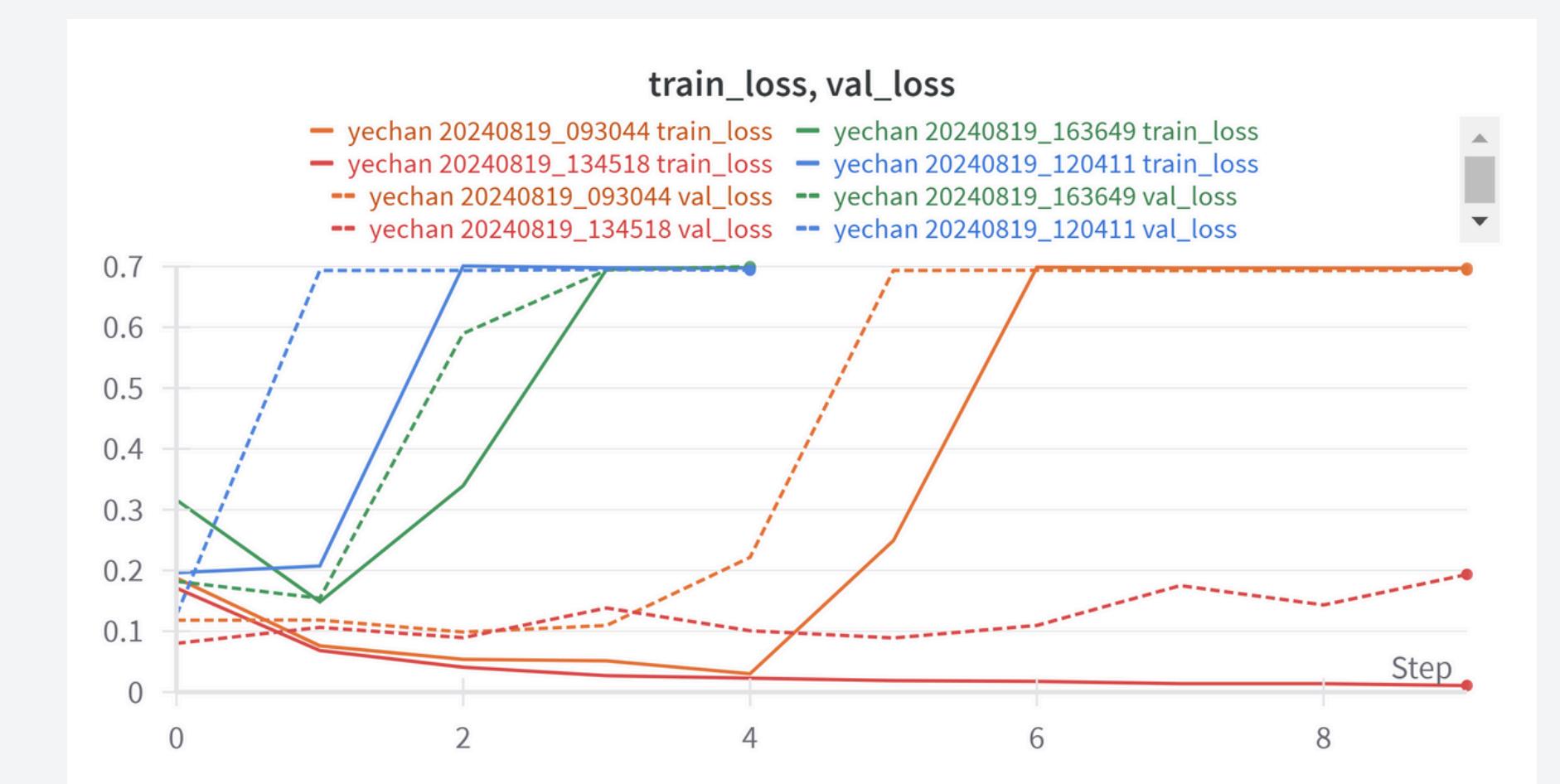
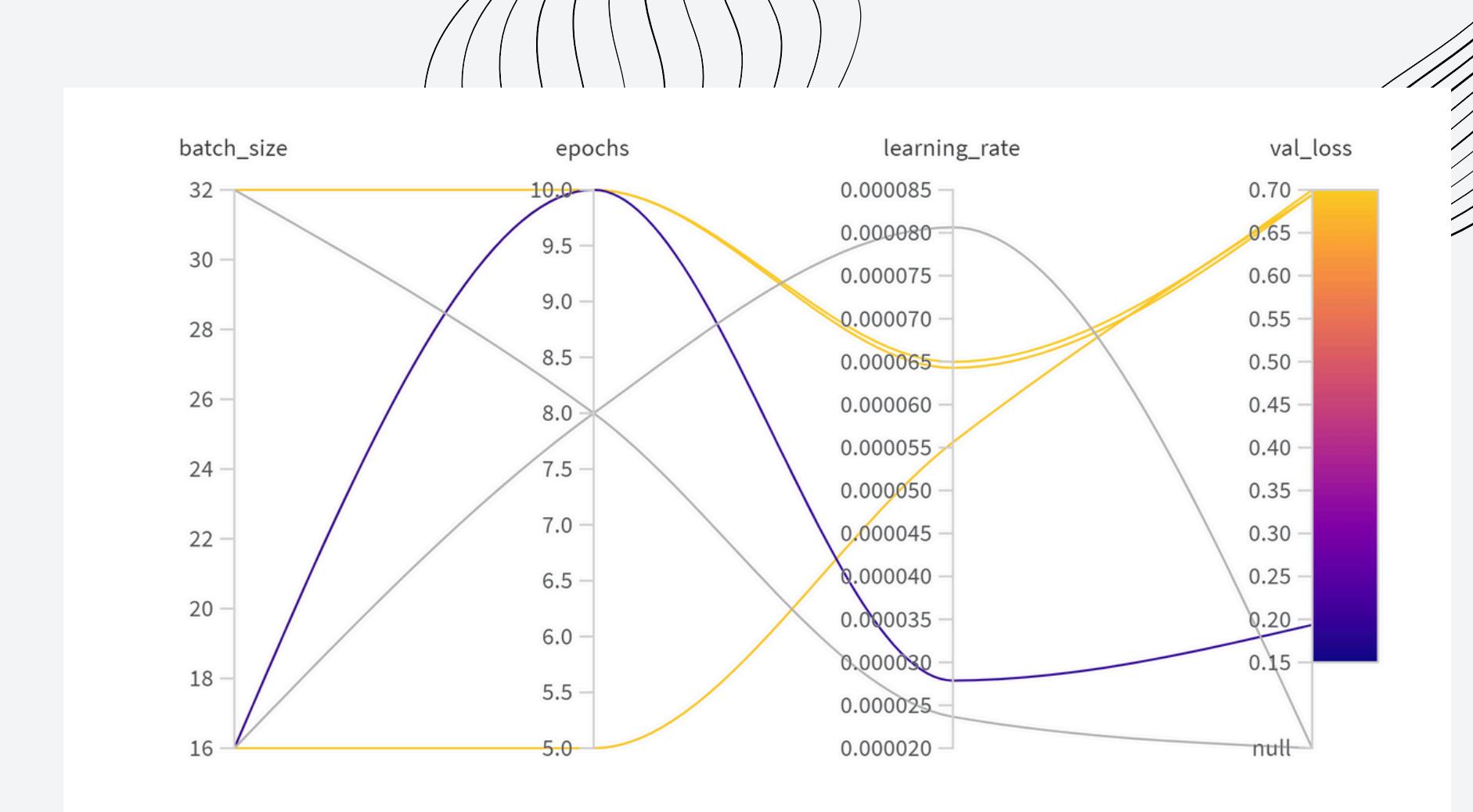


WHY NOT?

# MODEL

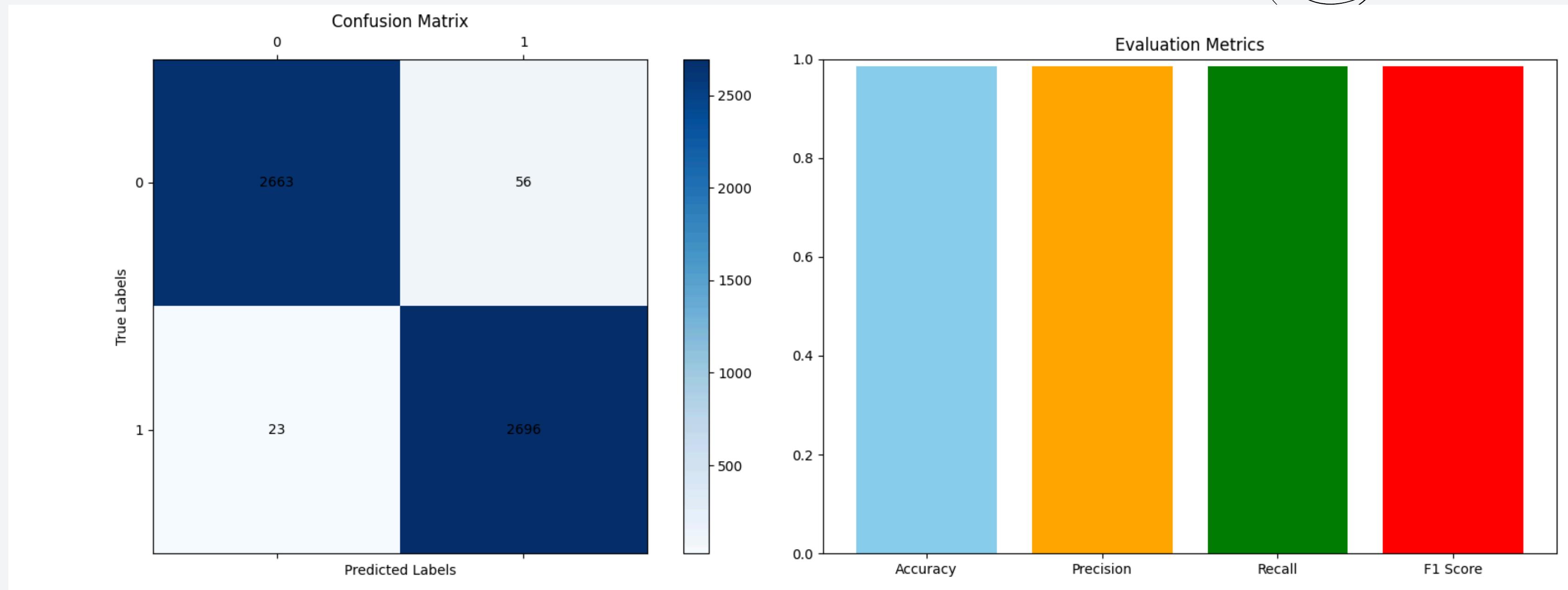
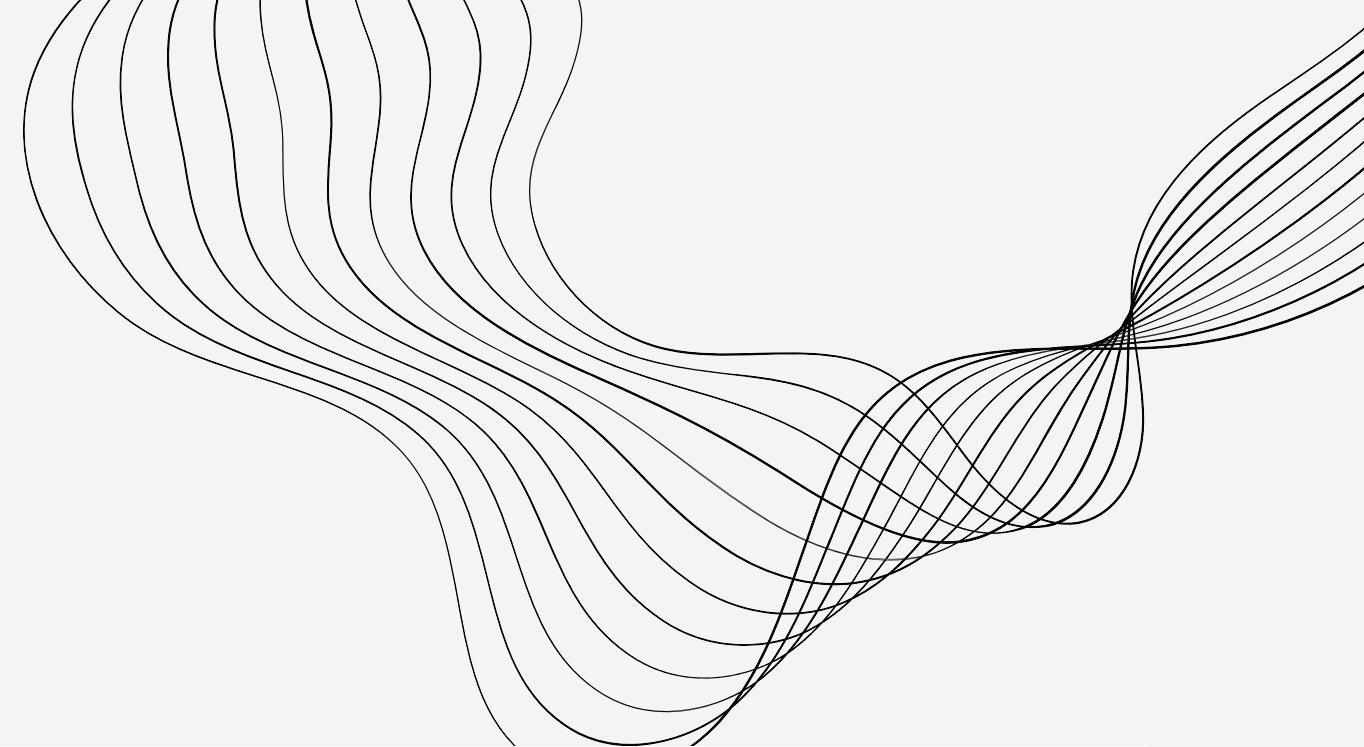
## KPF-BERT Finetuning

- Wandb logger 사용
  - Wandb Sweep를 통해 파라미터 조합 생성 (Bayes)
  - val\_loss가 반등하기 직전 Epoch를 Checkpoint로 저장
  - 모델 성능 평가 후 학습 반복 → 최고 성능 모델 저장
- 
- Batch\_Size: 16
  - Epochs: 10
  - Learning\_rate: 0.00002787
  - Optimizer: AdamW
  - Framework: PyTorch
  - EarlyStopping: Val\_loss



# MODEL

## Model Score



# DATA BASE & BACK-END



## MySQL



- 전세계적으로 널리 쓰이는 관계형 데이터베이스
- 이번 프로젝트에서는 정형 데이터만 사용할 예정이며 특히, 뉴스 데이터라는 긴 데이터를 불러오는데 MySQL의 빠른 읽기 성능은 큰 장점
- 오픈소스로서 대규모 데이터 처리와 높은 확장성을 제공



## FastAPI



- SQLAlchemy과 같은 다양한 ORM 라이브러리를 통한 데이터베이스 확장과 관리가 용이
- 복잡한 쿼리나 트랜잭션을 처리하기 용이
- 테스트 작성이 쉽고, 데이터베이스와 통합된 테스트 환경을 쉽게 구성할 수 있음

# DATA BASE & BACK-END

- 빅카인즈 csv에 크롤링 된 본문을 inner join하여 초기 db를 생성.
- like query 바탕으로 불러오는 초기 api 생성

STEP 1

- Fine-tuning된 모델을 통해 logit값 라밸링
- 유저값들을 기록하기 위한 유저 테이블 생성

STEP 2

## DataBase Structure

### <DB 'ybigta'>

Tables_in_ybigta
article
user_info

### <TABLE 'article'>

Field	Type	Null
article_id	varchar(255)	NO
title	text	YES
keyword	text	YES
content	text	YES
date	int	YES
image	text	YES
inference	json	YES

### <TABLE 'user\_info'>

Field	Type	Null
id	int	NO
user_id	varchar(255)	YES
average_logits	text	YES

# DATA BASE & BACK-END

- Fine-tuning된 모델을 통해 logit값 라밸링
- 유저값들을 기록하기 위한 유저 테이블 생성

STEP 2

- 유저가 선택한 기사를 바탕으로 성향을 판단 위한 모델 불러오는 api 추가
- 유저의 성향과 반대되는 기사를 가져오는 api 추가

STEP 3

- cron을 통한 DB article table 업데이트 자동화

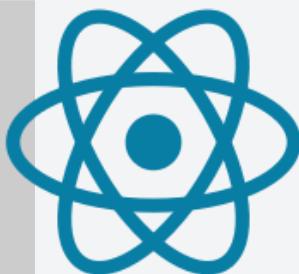
STEP 4

매일 새로운 뉴스레터를 제공해야 하므로,  
매일 새벽 3시에 자동으로  
5개 언론사에서 하루 동안  
의 기사를 크롤링 해 DB  
에 추가함.

# FRONT-END

## React

JavaScript의 프론트엔드 라이브러리로 컴포넌트 기반 구조를 통해 복잡한 UI를 효율적으로 만들 수 있다.



Frontend

HomePage

ArticleContent

SearchBar

MyPage

NewsletterPage

KeywordButtons

ArticleContents

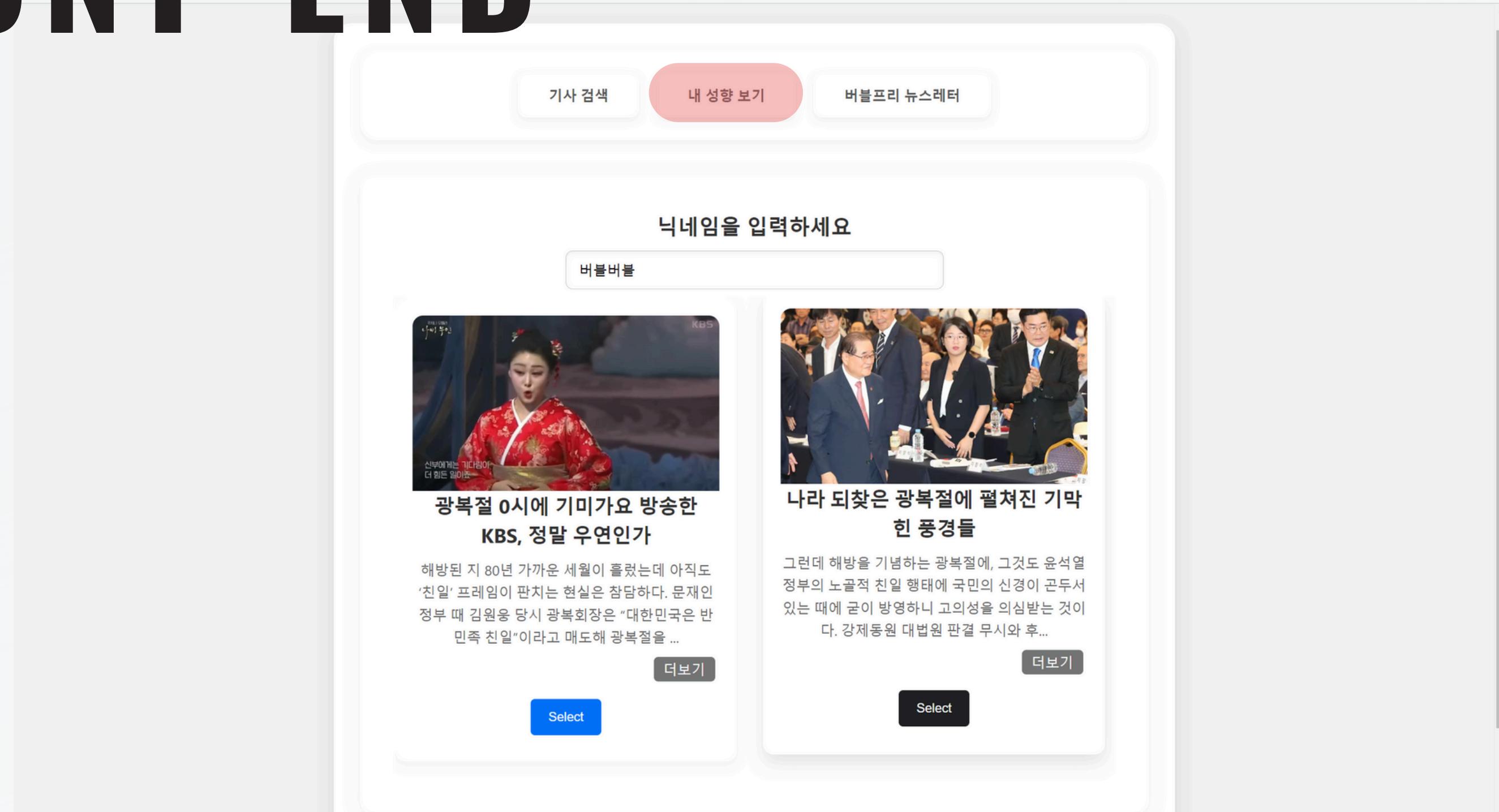
# FRONT-END

The screenshot shows a user interface for a newsletter. At the top, there are three buttons: '기사 검색' (highlighted in red), '내 성향 보기', and '버블프리 뉴스레터'. Below this is a title: '필터버블 해소를 위한 뉴스레터'. Underneath the title is a search bar with the placeholder 'Enter search query' and a dropdown menu set to 'Title'. A blue 'Search' button is to the right of the search bar. Below the search area, the text '기사 목록' is displayed, followed by the message '0개의 기사가 검색되었습니다.'

1. 기사 검색(HomePage)

: 검색을 통해 DB에 저장된 기사에 접근할 수 있음.

# FRONT-END



## 2. 언론 성향 테스트 (MyPage)

: 같은 주제를 다루는 두 기사 중, 마음에 드는 기사를 고르면 user 성향을 계산해 user DB에 저장.

# FRONT-END

기사 검색

내 성향 보기

버블프리 뉴스레터

## BubblePOP NewsLetter

국회 무혐의 영혼

[사설]檢 “김 여사 명품백 무혐의” 유사 사례도 ‘혈한 잣대’ 적용될까

김건희 여사의 '명품백 수수 의혹' 사건을 수사 중인 서울중앙지검이 '혐의 없음' 결론을 내렸다고 한다. 이창수 서울중앙지검장이 오늘 이원석 검찰총장에게 수사 결과를 보고하고, 이 총장이 수사심의 위원회를 소집하지 않는다면 김 여사는 조만간 불기소 처분된다. 이 사건은 재미교포 최재영 씨가 김 여사에게 300만 원 상당의 디올백을 선물하는 장면이 담긴 동영상을 한 인터넷 매체가 지난해 11월

이전 2024-08-22

### 3. 버블-프리 뉴스레터(NewsletterPage)

- : 닉네임을 입력하면, 해당하는 날짜의 기사 중 user의 성향과 가장 먼 기사를 최대 3개 추천함.
- : 선별된 기사의 키워드를 제공하고, 키워드 버튼을 클릭해 기사를 확인할 수 있음.

# OVERALL



사용자들은 평소에 접할 수 없었던 다양한 시각의 뉴스를 통해 정보의 폭을 넓히고, 뉴스 소비에 대한 새로운 접근 방식을 경험

‘뉴스 유목민’이라는 새로운 유저 페르소나의 확장

USER



단일한 시각에 치우치지 않고,  
다양한 목소리를 반영하는  
건강한 저널리즘을 실현

언론사의 신뢰도를 높이고,  
보다 넓은 독자층을 확보

MEDIA COMPANY

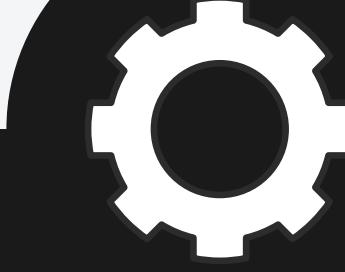


사회 전체의 정보 소비 문화에  
긍정적인 영향

더 나은 사회적 대화를  
촉진하는데 기여

SOCIETY

# TRIAL & ERROR



훈련 시 Val-loss가 무분별하게 팀  
Learning-rate와 batch-size  
변경해보았지만 이 두 파라미터보다  
Optimizer가 더 많은 상관관계가  
있음을 알게 됨

SGD -> Adam -> AdamW

OPTIMIZER



docker

팀원들과 협업할 때, 개발환경은  
세팅했으나 도커를 사용하지 않았음

추후, 배포하게 된다면 도커를 사용  
해서 다른 OS나 다른 컴퓨터에서도  
잘 작동할 수 있도록 해야 할 것임

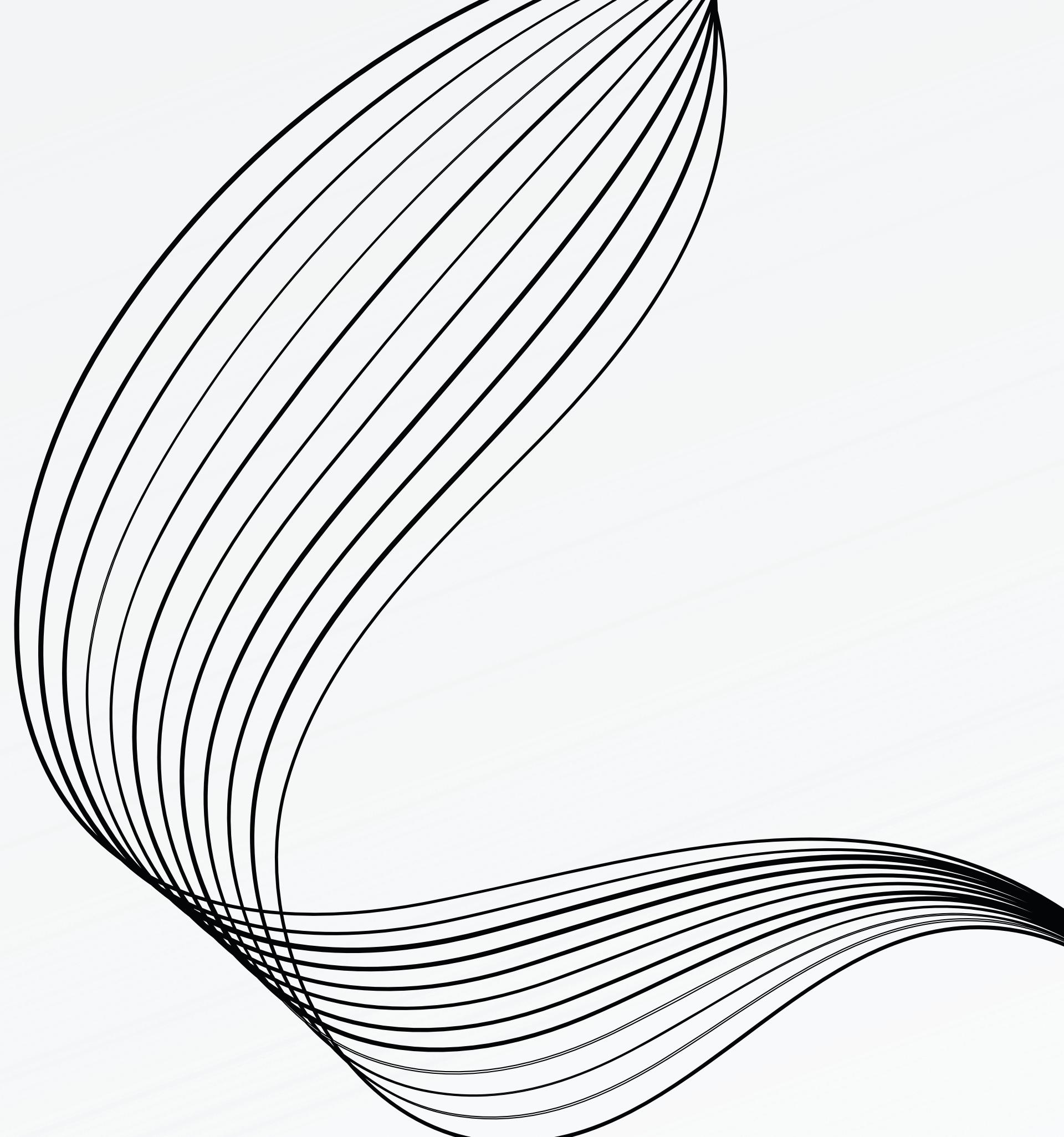
DOCKER



정보 손실을 최대한 피하기 위해  
특성 벡터로 Softmax 함수를  
적용하지 않은 Logits 값을 사용하려  
했으나, 기사의 길이나 종류가 다양해  
정규화해서 0-1의 범위를 가지는  
Softmax 예측확률 값을  
특성 벡터로 사용함.

SOFTMAX

# 프로젝트 시연



**THANK'S FOR  
WATCHING**

