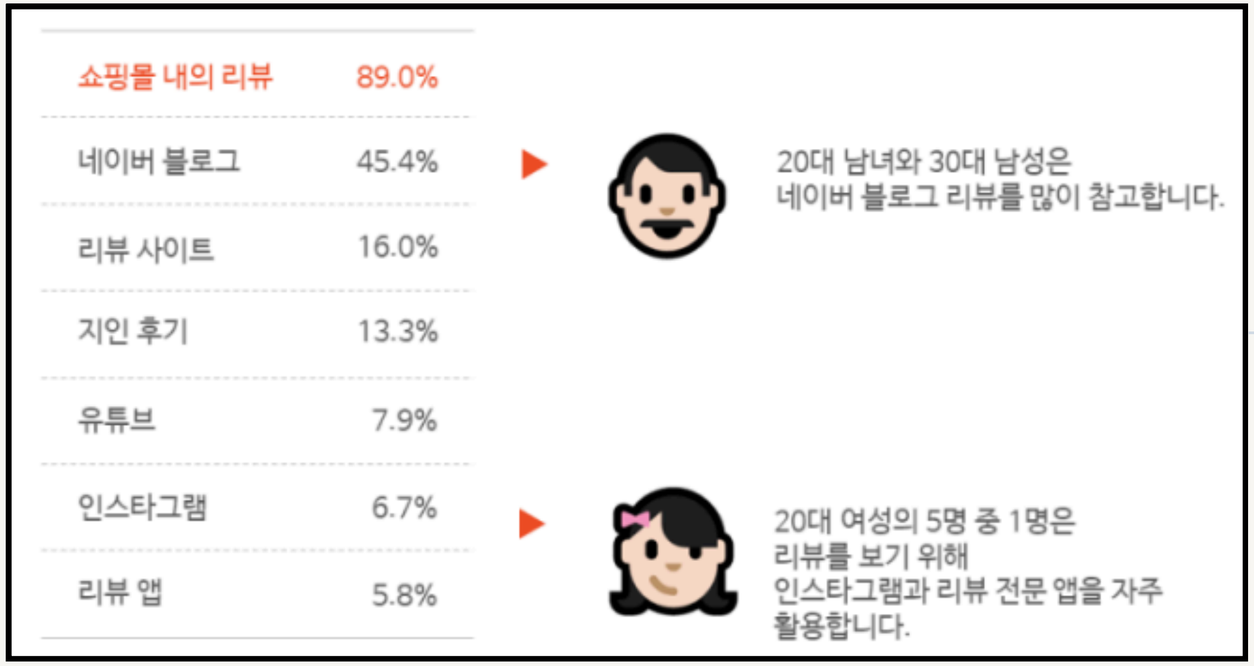
# KoBERT 기반 광고성 리뷰 분석 확장 프로그램

20191546
이상현

# 목차

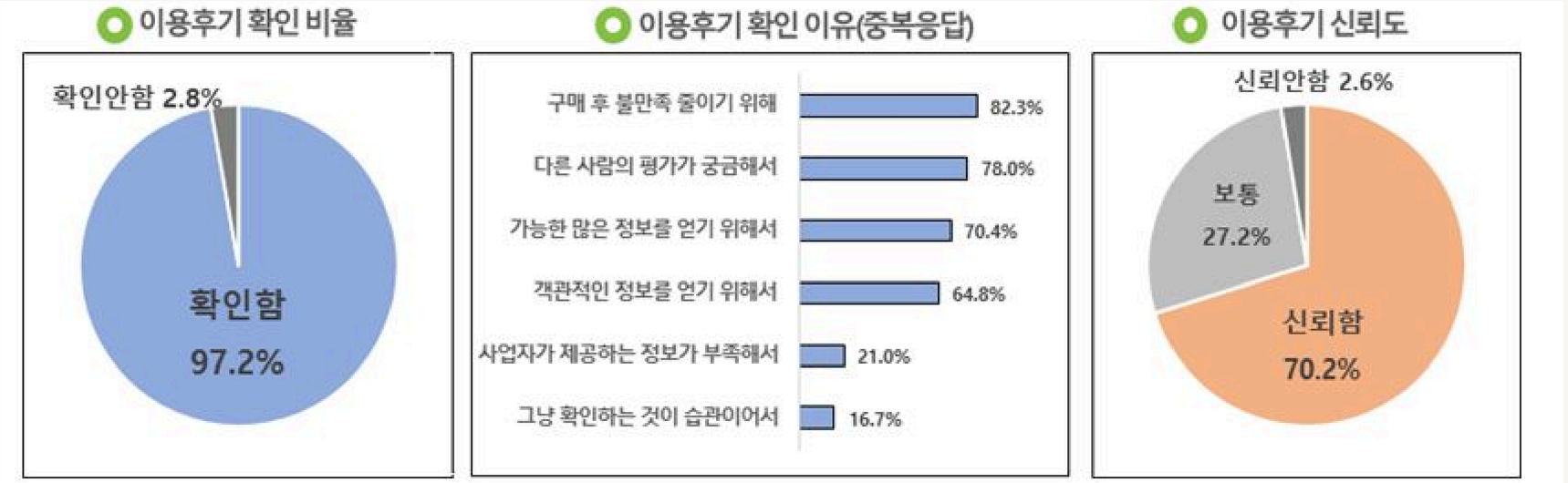
# 01 배경 및 필요성




## 국내 소비자들, 리뷰를 통한 구매 대다수

출처 : 마케터가 알아야 할 모바일 쇼핑과 리뷰에 관한
모든 것 |Trend Meetup
https://blog.opensurvey.co.kr/article/trendmtp_03/

# 01 배경 및 필요성




## 소비자 이용후기 확인률 97.2%, 신뢰도 70.2%

출처 : 출처 : 소비자경제(http://www.dailycnc.com)

## 02 기술 스택

### 크롬 익스텐션



### 서버

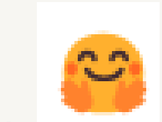


### 크롤링



### 딥 러닝

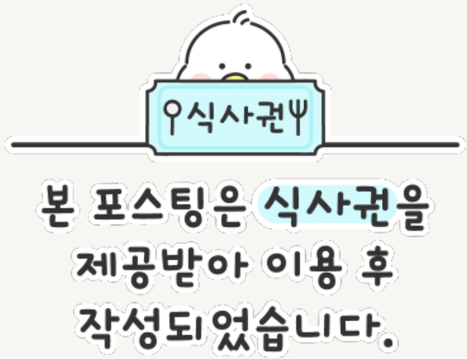
KoNLPy

transpormers

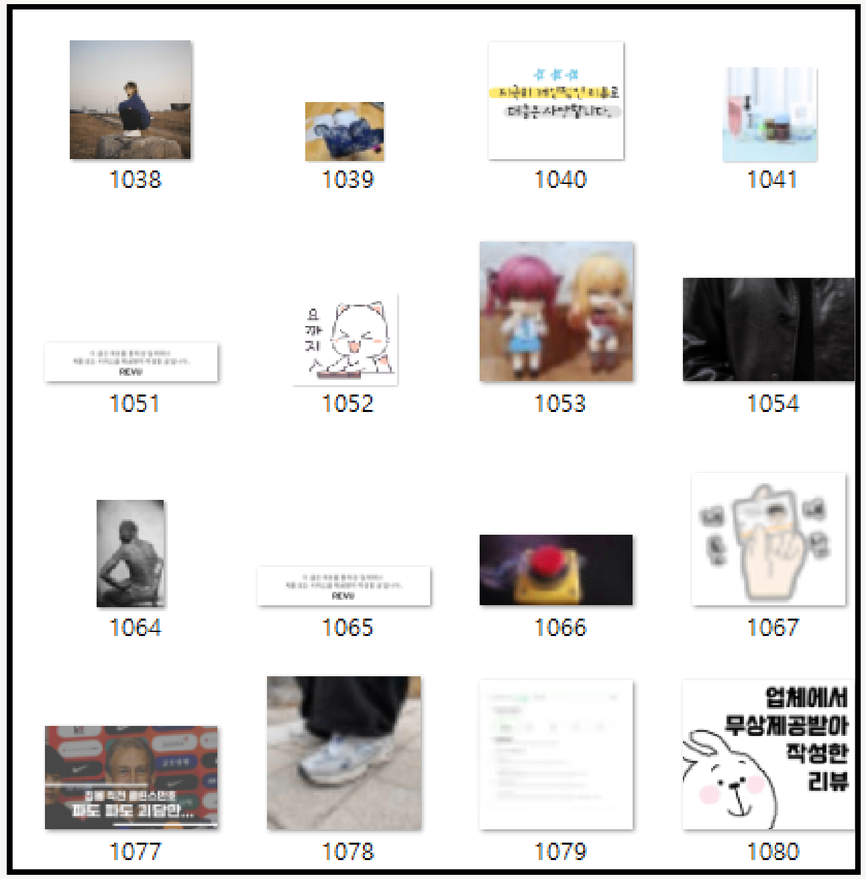
### 이미지 텍스트 추출

# OCR 텍스트 추출




- 네이버 블로그에서 광고리뷰는 이미지로 명시하는 경우가 많음
- 네이버 클로바 OCR 서비스를 통해 이미지에서 텍스트 추출

# 데이터 수집




| cnt | writer | date | url | title | content | ocr_data | comments | empathy_c | writer_revie | blog_is_promotion |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 행복하게 | 2024. 4. 2 | https://blo | 초보 운전 - | 초보운전4 | 포스팅은원 | 13 | 22 | 540 | 1 |
| 2 | 무한설비 | 2024. 5. 3. | https://blo | 화곡역 맛집 | 화곡역맛집 | THANKYo | 11 | 40 | 653 | 1 |
| 3 | 마음가는대 | 2024. 4. 1 | https://blo | 삼성 인덕션 | 삼성인덕션 | 므과 | 3 | 11 | 203 | 1 |
| 4 | 비파이어스 | 2024. 5. 3. | https://blo | 휴대용 미니 | 휴대용미니 | 선풍기추천,실사용리뷰 | | 4 | 95 | 0 |
| 5 | 오렌지 | 2024. 4. 2 | https://blo | 액상형 전자 | 액상형전자 | 포스팅은원 | 2 | 10 | 558 | 1 |
| 6 | RENARENA | 2024. 4. 9. | https://blo | 화곡동윤곽 | 화곡동윤곽 | 관리강미경 | 4 | 17 | 101 | 1 |
| 7 | 촤밍전 | 2024. 5. 4. | https://blo | 다낭 곱창 - | 다낭곱창쌀 | VILLADEN | 2 | 7 | 72 | 0 |
| 8 | 쿠우욱 | 2024. 4. 3 | https://blo | 무소음선풍 | 무소음선풍 | 이 글은 레 | 14 | 98 | 95 | 1 |
| 9 | SHIR | 2024. 5. 5. | https://blo | 기흥 이케아 | 기흥이케아 | 레스토랑아기랑이용리 | | 4 | 359 | 0 |
| 10 | 우라이더 | 2024. 4. 2 | https://blo | [강추 영상] | [강추영상] | 선거에대한 | 37 | 118 | 3 | 0 |
| 11 | konbella | 2024. 4. 2 | https://blo | 인스타 팔로 | 인스타팔로워 | 늘리기se | 없음 | 3 | 69 | 1 |
| 12 | 잡다 | 2024. 4. 2 | https://blo | 코웨이 얼음 | 코웨이얼음 | ₩본 포스팅 | 없음 | 1 | 140 | 1 |
| 13 | 너구리 | 2024. 5. 3. | https://blo | 대구 동성로 | 대구동성로 | 걸리버막창 | 16 | 54 | 36 | 0 |

- **Selenium, Beautiful Soup를 활용하여 네이버 블로그 글을 6473개 크롤링**
- **수집 목록 : 공감 수, 이웃 수, 본문, 본문 마지막 이미지등**

# 데이터 라벨링

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6464 | 소녀 | 2024. 2. 5. | https://blo | [리뷰] 데이 | [리뷰]데이터분석방법론:개요[리뷰]데이터분석방법론:개요20 | | 1 | 5 | 0 |
| 6465 | 럭철 | 2023. 12. | https://blo | [아츠제로] | [아츠제로]원피스쌍용도모모노스케피규어리뷰안 | 5 | 7 | 471 | 0 |
| 6466 | 한솔아카데 | 2024. 2. 2 | https://blo | 전기공사기 | 전기공사기사필기책5주완성,한솔아카데미출간리 | 없음 | 3 | 112 | 1 |
| 6467 | 더키럽 | 2024. 1. 2 | https://blo | 24신상 샤 | 24신상샤넬탑핸들시즌백가 | YEAH | 4 | 24 | 268 | 1 |
| 6468 | staybetter | 2023. 12. | https://blo | 플레이모빌 | 플레이모빌9060아쿠아리움 | J | 7 | 11 | 17 | 1 |
| 6469 | 카츠 | 2024. 2. 1 | https://blo | 마운틴 하 | 마운틴하우스(MountainHouse)간편식리뷰(1)요즘 | | 10 | 21 | 93 | 0 |
| 6470 | 미리내맨 | 2023. 12. | https://blo | [호텔객실 | [호텔객실리뷰]켄싱턴호텔여 | 7=( ( | 26 | 21 | 81 | 1 |
| 6471 | leedana | 2023. 12. | https://blo | [선유도역 | [선유도역카페/선유도카페:: | 마이어\| Oㅅ | 7 | 13 | 33 | 0 |
| 6472 | 그린 | 2024. 1. 3 | https://blo | [도서리뷰] | [도서리뷰]오직한사람의차지 | 김금희소설오직한사 | 7 | 14 | 9 | 0 |

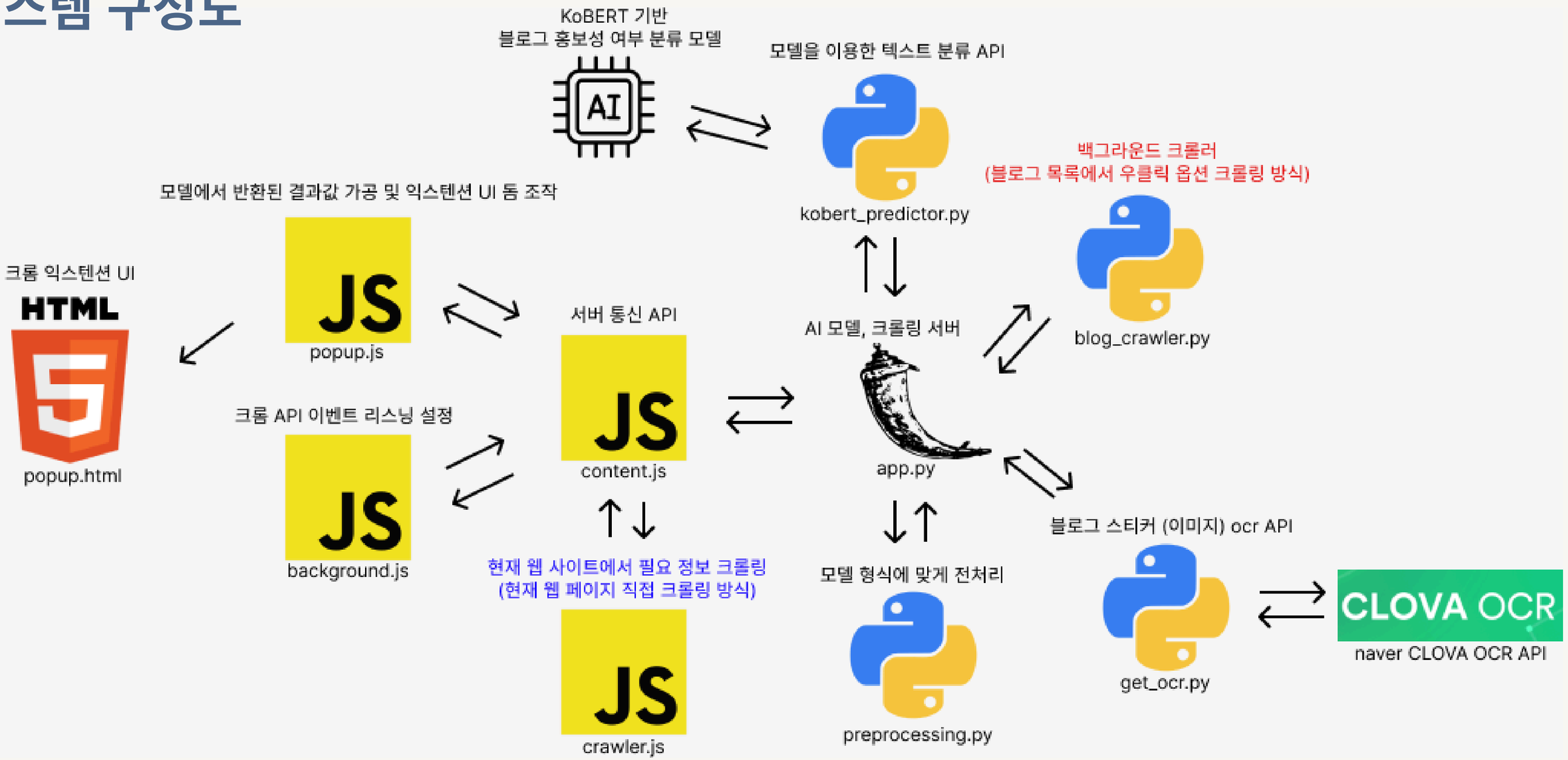
- 본문 데이터 또는 OCR 추출 텍스트에 측정 키워드 ( 지원받고, 제공받고, 서포터즈, 파트너스 등)의 가 있을 경우 1으로 라벨링
- 광고법 제정 이후의 글들만 크롤링 해왔기에 나머지는 0으로 라벨링

# 데이터 전처리

| combined_text |
| --- |
| 초보 운전 4일 10시간 자동차운전연수리뷰초보운전 4일 10시간 자동차운전연수리뷰안녕하세요 . 초보 운전 4일 10시간 자동차운전연수리뷰를해볼까하 |
| 화곡역 맛집 조밥 맛있다 이현우 스시 리뷰 ( 주차 , 메뉴 ) 인스타 계정 을 만들다 아래 카드 누르다 들어오다 블로그 에서 소개 되다 다양하다 맛집 리뷰 |
| 삼성 인덕션 3 구 추천사 용법 및 리뷰 삼성 3 구 인덕션 프리 스탠 딩 사 용법 및 리뷰 안녕하다 ! 주방 인테리어 구상 하 면서 로 망중 하 나 이다 인덕션 |
| 휴대 용 미니 선풍기 추천 , 실 사용 리뷰 !https://smartstore.naver.com/skillfulbrother/products/10240351343 휴대 용 미니 선풍기 추천 , 실 사용 리뷰 |
| 액 상형 전자담배 순위 1 위 하 카시 그 니 처 리뷰 액 상 전자담배 순위 # 액 상 전자담배 순위 # 액 상형 전자담배 순 위액 상형 전자담배 순위 1 위 하 |
| 화곡동 윤곽 관리 강미경 피부 & 전 신 관리 리뷰 숨다 화곡동 윤곽 관리 전문점 강미경 피부 & 전 신 관리 다녀오다 .( 간판 강미경 뷰티 연구소 라고 또 |
| 다낭 곱창 쌀국수 맛집 포틴 ( 퍼틴 ) 13 솔직 리뷰 다낭 살이 중 인 촤밍 전입 니 다 ! 오늘 다낭 곱창 쌀국수 맛집 포틴 ( 퍼틴 ) 13 솔직 리뷰 전해 드리 |
| 무 소음 선풍기 추천 아이오 랩 BLDC 선풍기 아이브 리즈 리뷰 무 소음 선풍기 추천 아이오 랩 BLDC 선풍기 아이브 리즈 리뷰 글 . 사진 ⓒ 쿠우 욱 지난 |
| 기흥 이케아 레스토랑 아기 랑 이용 리뷰 기흥 프리미엄 아울렛 에서 미니 기차 회전목마 타고 저녁 을 먹다 집 출발 하 지만 집 에가 서다 뭐 먹다 ? 집 |
| [ 강추 영상 ] 선거 대한 7년 만의리뷰 , 미국 정치 한국 정치 7년 만의제대로된리뷰를봤다 . 이런 리뷰 있다 정치가 존재 의미 가지다 . 내 가정 치의 영역 |
| 인스타 팔로워 늘리다 self 설정 리뷰 인스타 팔로워 늘리다 self 설정 리뷰 저 여러 SNS 경험 후 에도 인스타그램 을 가장 선호 . 현재 주로 일상 을 담다 |
| 코웨이 얼음 정수기 렌탈 추천 받다 리뷰 얼음 정수기 렌탈 # 얼음 정수기 렌탈 # 얼음 정수기 렌탈 추천 코웨이 얼음 정수기 렌탈 추천 받다 리뷰 요즘 |
| 대 구 동성로 걸리버 막창 여섯 번 먹다 보고 첫 리뷰 대구 정말 막창 대동단결 하 도시 이다 . 원래 유명하다 삼겹살 집 에도 거의 막창 있다 , 막창집 에 |

- 불용어 제거 ( '의', '가', '이', '은' 등 )
- NaN 값 및 빈 문자열 처리, 줄바꿈 문자 제거, 형태소 분석 및 어간 추출, 토큰 결합
- KoBERT 입력 형식으로 변환
- 데이터 분할 및 데이터 로더 생성

# 시스템 구성도


KoBERT 기반
블로그 홍보성 여부 분류 모델
AI
모델을 이용한 텍스트 분류 API
kobert_predictor.py
백그라운드 크롤러
(블로그 목록에서 우클릭 옵션 크롤링 방식)
blog_crawler.py
모델에서 반환된 결과값 가공 및 익스텐션 UI 동 조작
크롬 익스텐션 UI
HTML
popup.html
JS
popup.js
서버 통신 API
AI 모델, 크롤링 서버
JS
content.js
app.py
크롬 API 이벤트 리스닝 설정
JS
background.js
현재 웹 사이트에서 필요 정보 크롤링
(현재 웹 페이지 직접 크롤링 방식)
JS
crawler.js
모델 형식에 맞게 전처리
preprocessing.py
블로그 스티커 (이미지) ocr API
get_ocr.py
CLOVA OCR
naver CLOVA OCR API

# 딥러닝 모델 : 트랜스포머

문장 토큰화 및 임베딩

포지셔널 인코딩 (Positional Encoding)

멀티헤드 셀프 어텐션 (Multi-head Self-attention)

Residual (Add) & Normalization

피드포워드 뉴럴 네트워크 (FFNN)

Dense 레이어

드롭아웃 (Dropout)

# 딥러닝 모델 : 트랜스포머

문장 토큰화 및 임베딩

result

포지셔널 인코딩 (Positional Encoding)

멀티헤 ntion)

```
Average training loss: 0.13
Validation Loss: 0.14
Validation Accuracy: 0.96
Precision: 0.97
Recall: 0.94
F1 Score: 0.96
```

피드포워드 뉴럴 네트워크 (FFNN)

↓

Dense 레이어

↓

드롭아웃 (Dropout)

# 기대효과

- 정보 신뢰성 향상: 사용자에게 신뢰할 수 있는 정보를 제공하여 만족도와 플랫폼 신뢰도 증가.
- 마케팅 전략 개선: 기업이 시장 동향을 파악하고 마케팅 전략을 효과적으로 수립.
- 사용자 만족도 증가: 리뷰 글을 읽기 전에 광고 여부를 미리 파악하여 신뢰할 수 있는 정보를 제공.

# 향후 계획

- 전처리 및 모델 개선: 텍스트 패턴, 감정 분석 등을 추가로 고려하는 정교한 전처리 과정을 도입할 계획
- 추가적인 데이터 수집 및 라벨링: 다양한 도메인의 리뷰 데이터를 추가로 수집하고, 보다 철저한 라벨링 과정을 통해 데이터셋의 품질을 향상시킬 계획
- 속도 개선: 모델의 속도를 향상시키는 최적화를 실시하여 사용자 경험을 향상시킬 계획

# 감사합니다

20191546 이상현

2024.06.12