

# 당신도 중고 거래왕이 될 수 있습니다!

#Classification #Text-Extraction #Text-Generation

NLP-16  
(NLPRIME)



**boostcamp** aitech

# 목차

1. 팀 소개
2. 프로젝트 개요
3. 시연
4. 카테고리 분류모델
5. 해시태그 모델
6. 향후 개선 방안

# 1. 팀 소개

# 1. 팀 소개 – NLPRIME

---



@김아경\_T2259

#추출모델설계  
#텍스트전처리



@김현욱\_T2069

#이미지전처리  
#분류모델검증



@김황대\_2071

#생성모델설계  
#프로토타입설계



@박상류\_T2083

#생성모델설계  
#텍스트전처리



@정재현\_T2205

#데이터수집  
#ES설계 및 구현



@최윤성\_T2230

#PM  
#분류모델설계

## 2. 프로젝트 개요

## 2. 프로젝트 개요

---

- **프로젝트 의도**

- 중고 판매글을 소비자에게 조금 더 노출 시킬 수 없을까?
- 카테고리 설정을 편하게 할 수 없을까?
- 사용자가 해시태그를 고민하는 시간을 줄일 수 없을까?

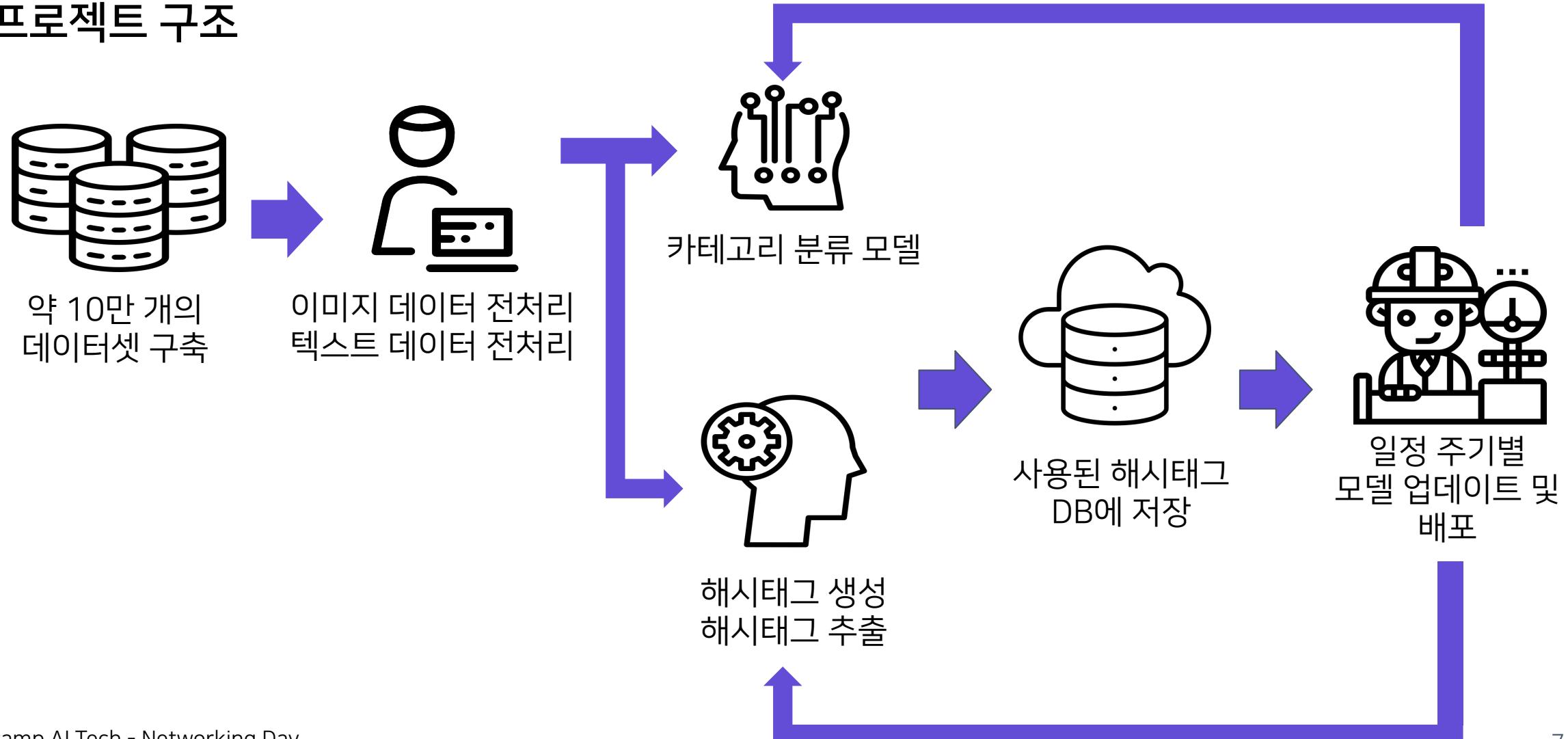
- **프로젝트 목적**

- 생성 모델, 추출 모델로 해시태그를 만들어 주자!
- 사용자가 이미지와 제목을 입력하면 자동으로 카테고리를 만들어 주자!



## 2. 프로젝트 개요

- 프로젝트 구조



### 3. 시연



## 4. 카테고리 분류모델

# 4. 카테고리 분류모델

- Single Modality 분류 모델의 문제점

- 현재 중고거래 플랫폼에서는 상품 제목을 통한 카테고리 분류 기능 제공
- 하지만, Text 만으로 분류하는 경우 다음과 같은 문제 발생

브랜드 or 모델명만 작성한 경우



한성 gk898b

카테고리 선택 >

노트북/데스크탑 > 노트북/넷북 > 기타 제조사

노트북/데스크탑 > 모니터

제목에 상품을 명시하지 않은 경우



6만원에 팝니다

카테고리 선택 >

모바일/태블릿 > 스마트폰 > 애플

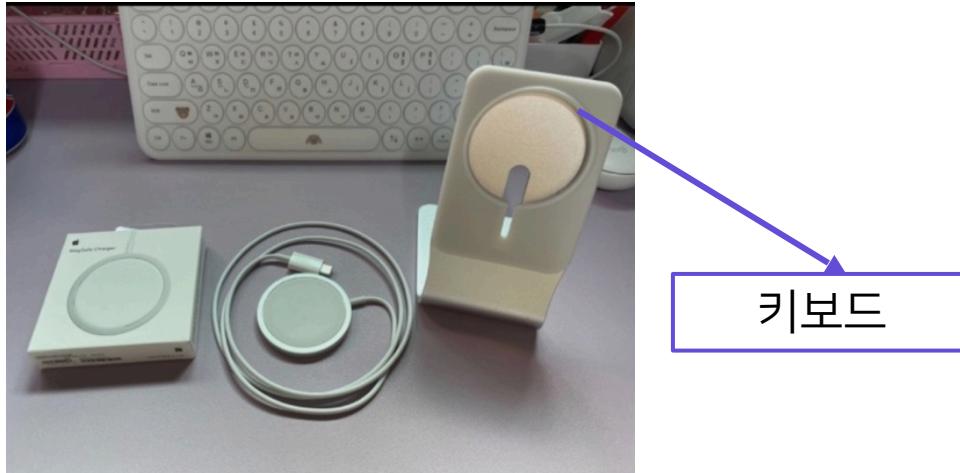
티켓/쿠폰 > 상품권/쿠폰 > 외식/주유

## 4. 카테고리 분류모델

- Single Modality 분류 모델의 문제점 (cont.)

- 현재 중고거래 플랫폼에서는 상품 제목을 통한 카테고리 분류 기능 제공
- Image 만으로 분류하는 경우도 문제 발생 가능

판매하지 않는 다른 상품이 함께 등록될 경우



애플 맥세이프 무선 충전기 + 거치대

미개봉 박스 이미지를 등록하는 경우

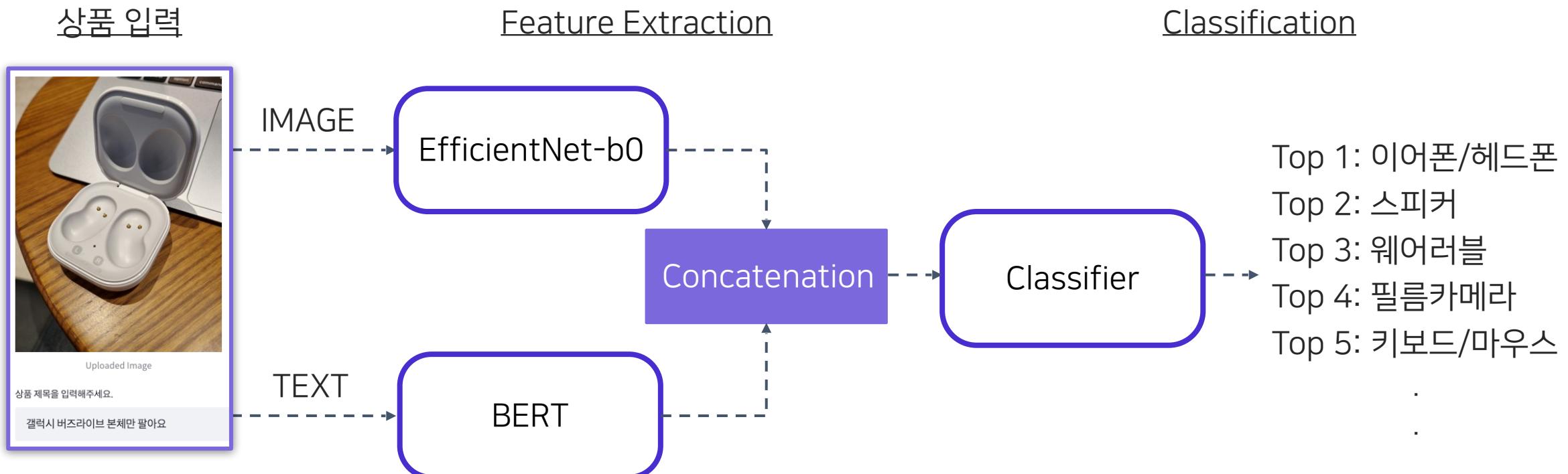


스타벅스 머그컵 (미개봉)

# 4. 카테고리 분류모델

## • 1차 모델링

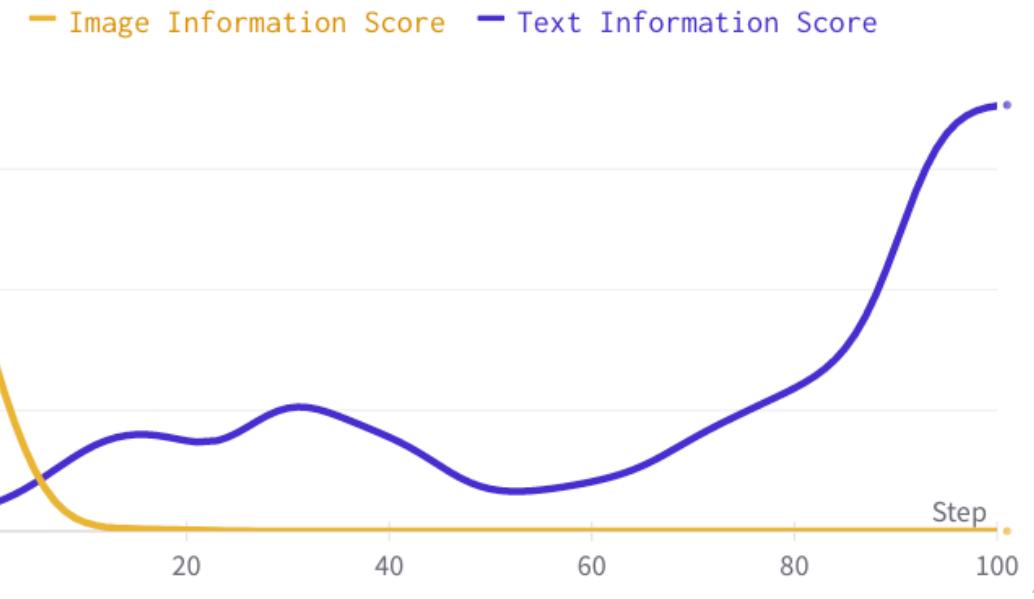
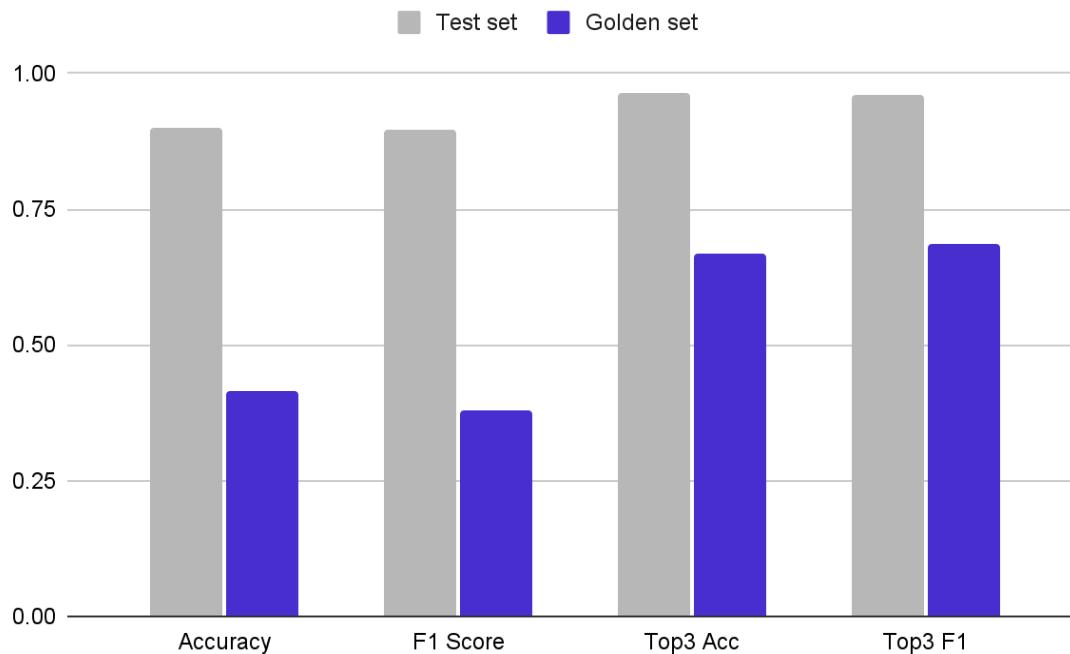
- Image backbone 모델은 EfficientNet-b0, Text backbone 모델은 BERT
- 각 modality data로부터 feature를 추출한 뒤 concatenated feature 를 최종 분류에 사용



## 4. 카테고리 분류모델

- 1차 모델링 결과

- Golden dataset에 대한 낮은 분류 성능
- Biased modality의 Information Score가 압도적으로 높음

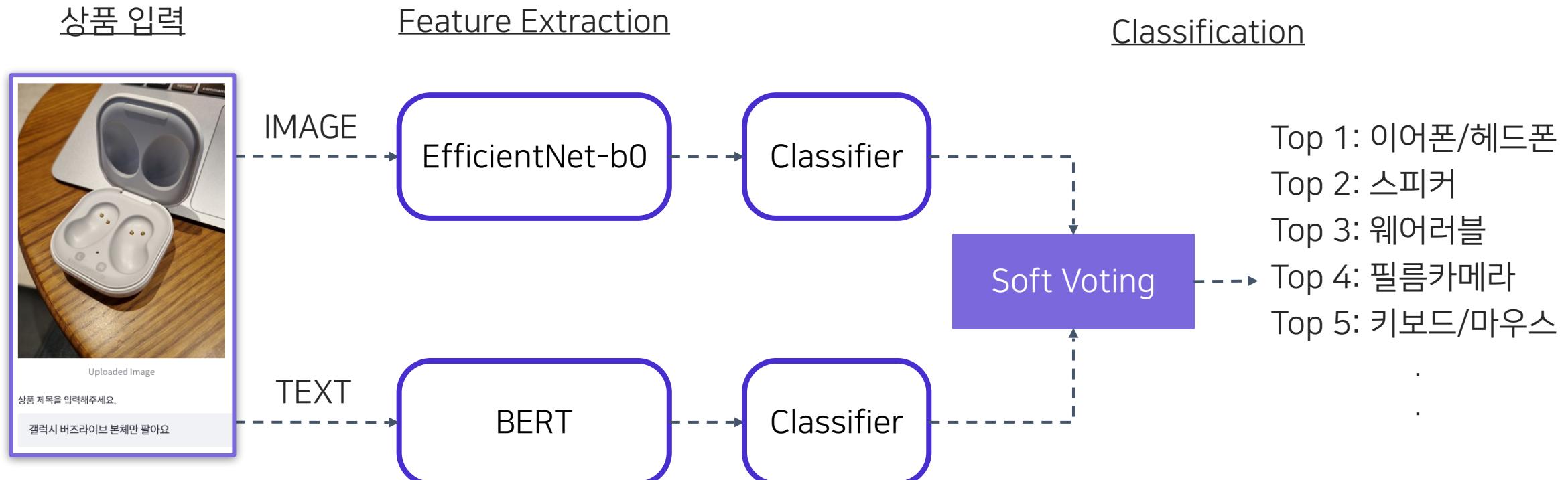


Itai Gat, et al. "Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies." NeurIPS. 2020.

# 4. 카테고리 분류모델

## • 2차 모델링

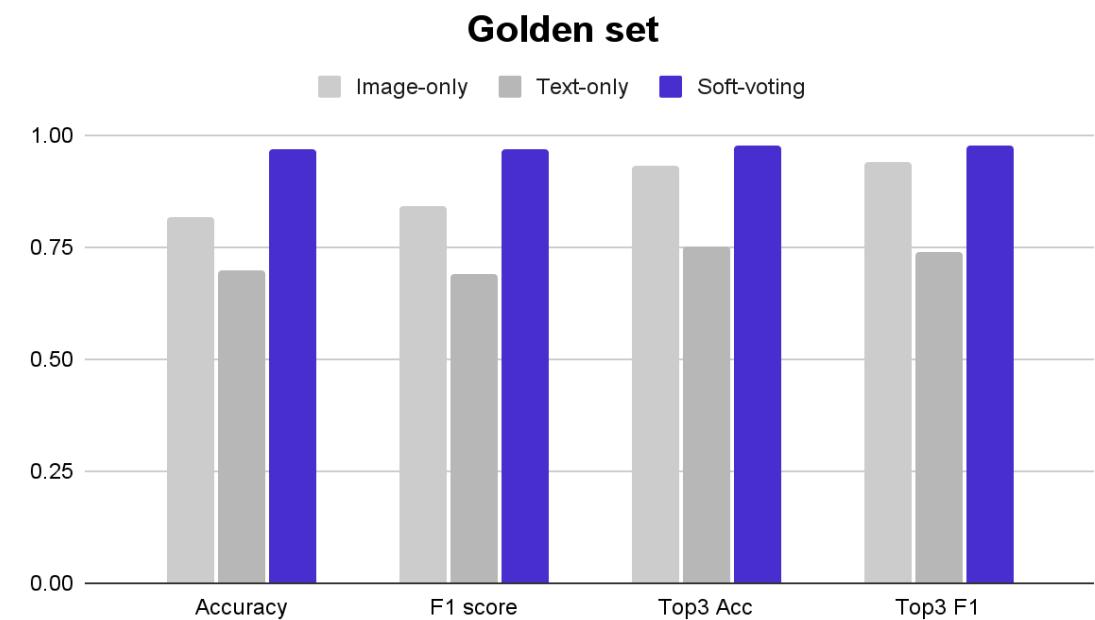
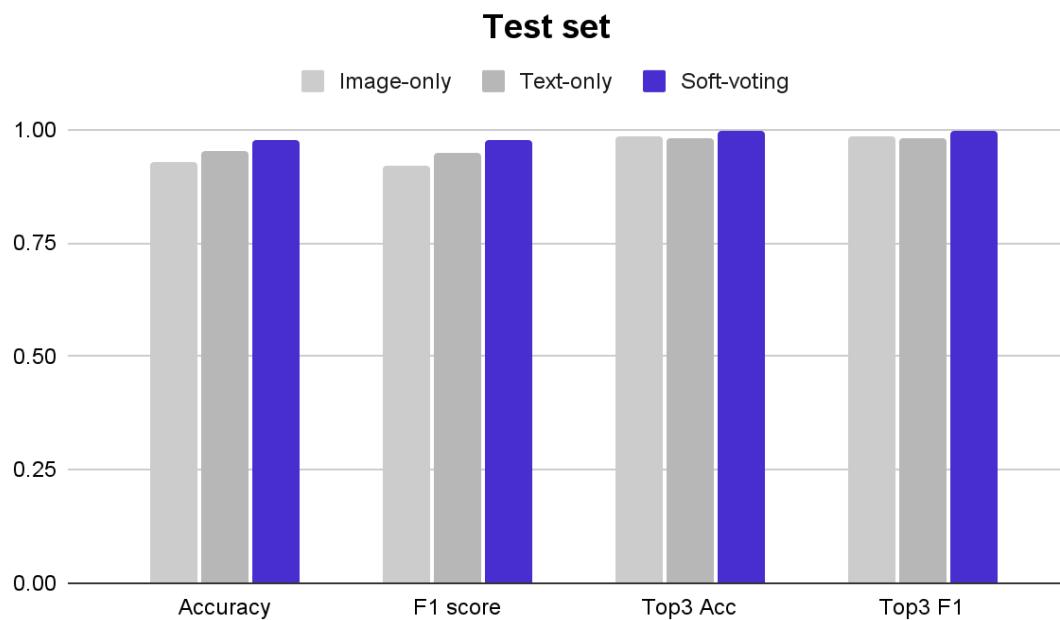
- Single Modality만 사용하는 분류모델을 분리하여 학습
- 각 분류모델을 통해 나온 예측확률을 바탕으로 최종 단계에서 Soft Voting



## 4. 카테고리 분류모델

- 2차 모델링 결과

- Single Modality 모델의 Test set 분류 성능은 좋았지만 Golden set 성능은 떨어짐
- 하지만, Voting model은 Test set, Golden set 모두 높은 분류성능



## 5. 해시태그 모델

# 5. HashTag

- HashTag 모델

#갤럭시북 #NT751QCJ #15.6인치 #i510210U

판매 중고제품의 상표, 모델명, 스펙 등의 상세한 정보를 추출

→ 추출 모델



삼성 갤럭시북 플렉스알파 (NT751QCJ-K03/C)

#삼성노트북 #중고노트북 #삼성전자 #galaxybook

제목, 본문에 나타나지는 않지만 실제 사용자들이 판매율을  
높이기 위해 등록한 해시태그를 생성

→ 생성 모델

상품 최초구매일 : 2020년 Q3

...

인텔 코어 i5-10210U (4코어, 8스레드)  
15.6인치 Full HD 터치패널 (IPS, 600 nits)

...

택배거래 안 받습니다...

# 5.1 HashTag 추출모델

---

- HashTag 추출 방법

- 제약 사항

- 2주간의 짧은 프로젝트 기간
    - DL Model을 사용하기 위해서는 10만개의 데이터 라벨링 필요

- 해결 방법

- TF-IDF 기반 핵심 단어 추출
    - 추출의 속도를 높이기 위한 ElasticSearch
    - 적절한 해시태그를 추출하기 위한 Vocab과 Query 고도화



## 5.1 HashTag 추출모델

- HashTag vocab 생성

삼성 갤럭시 Z플립3 크림 직거래 선호

안녕하세요.

두달 전에 샀는데 사용하기 힘들어서 팔아요.

용량은 256GB. 풀박스입니다.

모든 기능 정상이고 상태 좋아요.

70만원 직거래, 번개 페이 선호합니다.

택배는 72만원. 감사합니다.

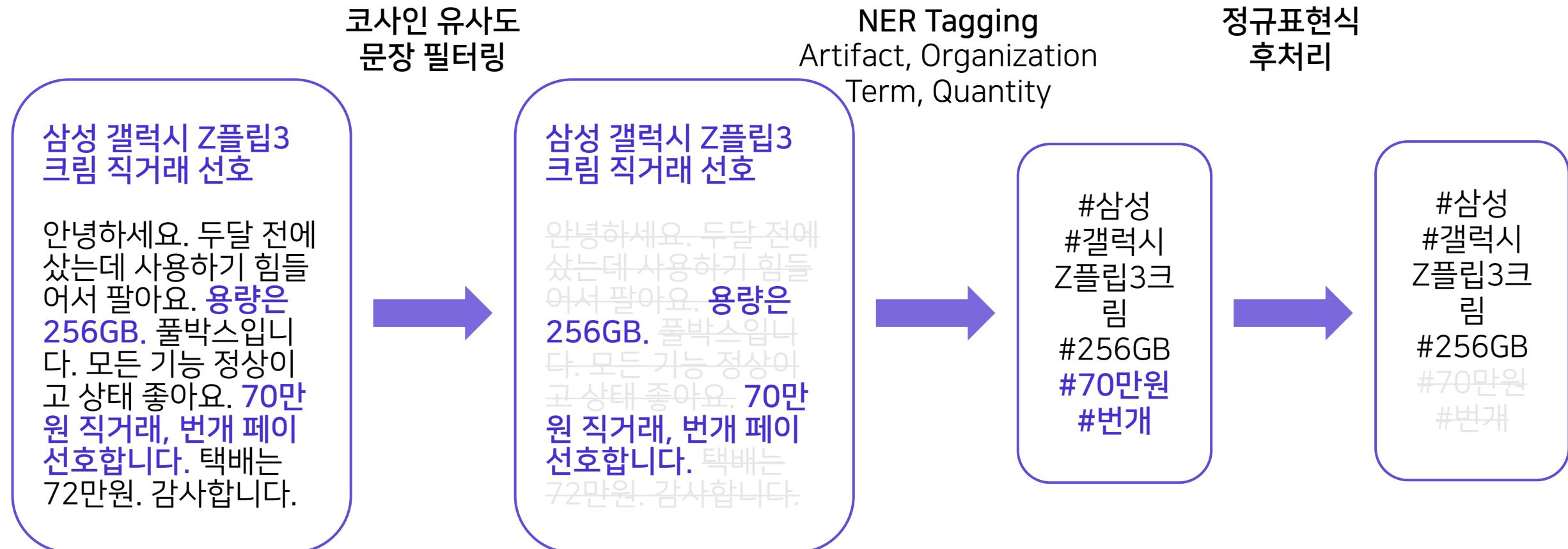
KoNLPy kkma  
명사 추출



#삼성 #갤럭시 #플립 #크림 #직거래  
#선호 #달 #전 #사용 #용량 #풀 #박스  
#모든 #기능 #정상 #상태 #번개  
#페이 #선호 #택배

# 5.1 HashTag 추출모델

- HashTag vocab 생성 개선방안



## 5.1 HashTag 추출모델

- Elastic Search

g102라일락 미개봉급팝니다  
로지텍g102입니다.  
새상품과 종일합니다  
미개봉싸게가져가세요.  
장비 다 정리중이라 싸게 처분합니다.  
신용잇는사람에게구입하세요 !  
 택배거래 선호합니다.

nori\_tokenizer

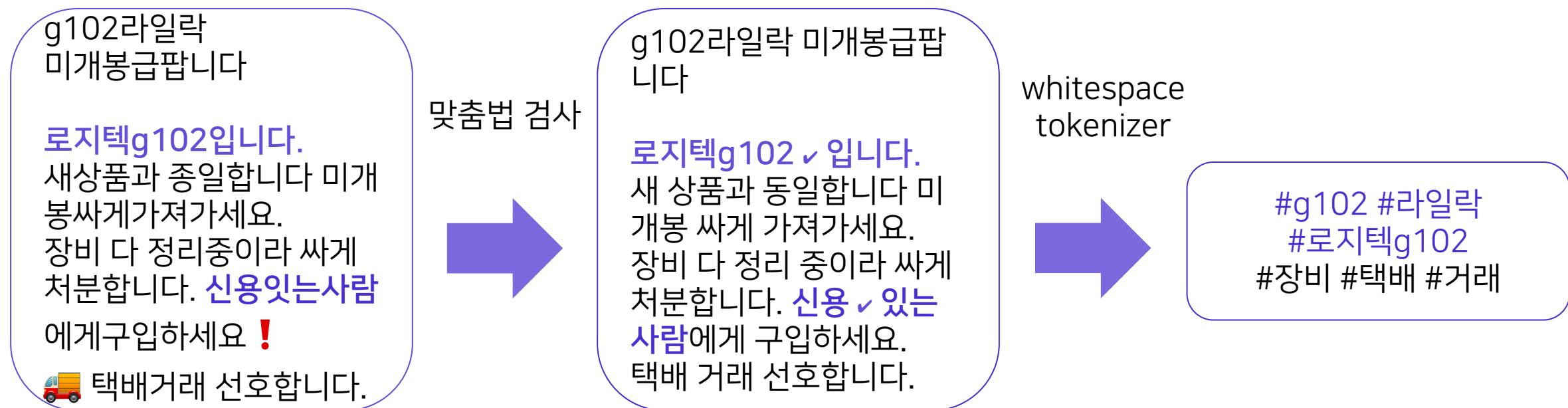


추출되지않음!!

# 5.1 HashTag 추출모델

- Elastic Search 개선방안

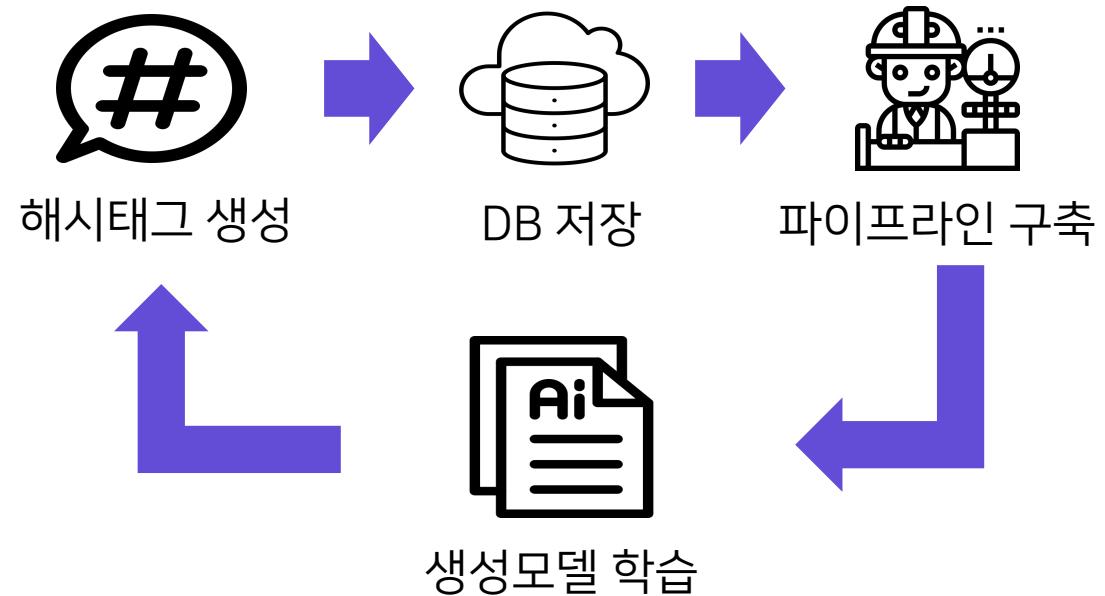
- 사용자가 입력한 쿼리에 대해서 맞춤법 검사 전처리
- ES의 검색 토크나이저로 white space 사용



## 5.2 HashTag 생성모델

- GPT-2 (base : skt/kogpt2-base-v2)

- 실제 사용하는 해시태그를 생성하기 위해 생성 모델을 사용
- 번개 장터 실제 DATA 10만개를 이용한 Fine-Tuning 및 진행
- 약 100명의 설문조사를 통해 정성적 평가지표를 반영함.



<s> 갤럭시 s10 팝니다 <sep> 손상없이 깨끗한 ... 할인해드립니다. <sep> 갤럭시s10, 깨끗한핸드폰 </s>

제목

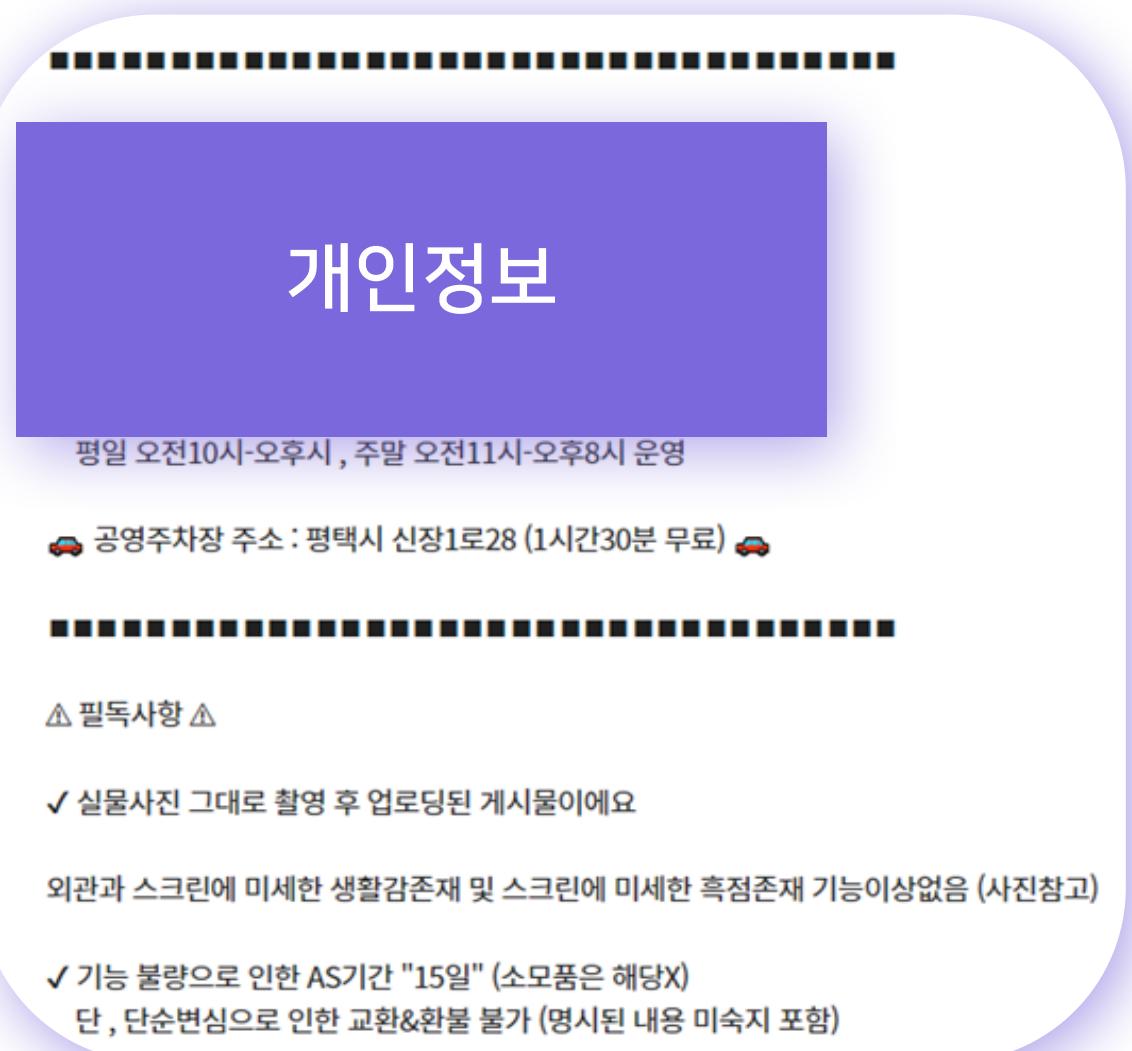
상품설명

해시태그

## 5.2 HashTag 생성모델

- 제한된 시간, 많은 데이터

- 수집된 데이터에는 사용자 개인정보, 광고성 문장 등 전처리가 필요한 데이터
- 정규표현식으로 전처리가 불가능한 데이터 형식  
ex) 공1공 -공0공공-영00공
- 제한된 시간의 많은 데이터를 전처리 하기위한 방법이 필요함



## 5.2 HashTag 생성모델

- Cosine 유사도와 NER Tagging

- 제목과 문장단위로 나눈 본문으로 Cosine 유사도를 구함  $\Rightarrow$  0.3 Threshold
- 정규표현식으로 제거되지 않았던 개인 정보, 광고성 문장 등이 제거됨
- NER Tagging으로 지역명이 포함된 HashTag를 삭제

S급 아이폰12미니 64GB 민트

64GB 아이폰 12미니입니다.  
아이폰 액정깨끗합니다.  
~~구입후 7일이내 가능 이상시 교환,환불가능 합니다.~~  
~~(단순변심,고객과실 제외).~~  
전화주세요 공10-일2삼사-5육

#부산핸드폰 #부산 #아이폰 #김해핸드폰



S급 아이폰12미니 64GB 민트

64GB 아이폰 12미니입니다.  
아이폰 액정깨끗합니다.  
~~구입후 7일이내 가능 이상시 교환,환불가능 합니다.~~  
~~(단순변심,고객과실 제외).~~  
~~전화주세요 공10-일2삼사-5육~~

#핸드폰 #아이폰

## 5.2 HashTag 생성모델

- 생성모델 결과



삼성 갤럭시북 플렉스알파 (NT751QCJ-K03/C)

상품 최초구매일 : 2020년 Q3

...

인텔 코어 i5-10210U (4코어, 8스레드)

15.6인치 Full HD 터치패널 (IPS, 600 nits)

...

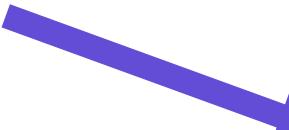
택배거래 안 받습니다...



전처리 전 데이터로 학습한 모델

#갤럭시북 #대전노트북 #당일

#파름신오신날



전처리 후 데이터로 학습한 모델

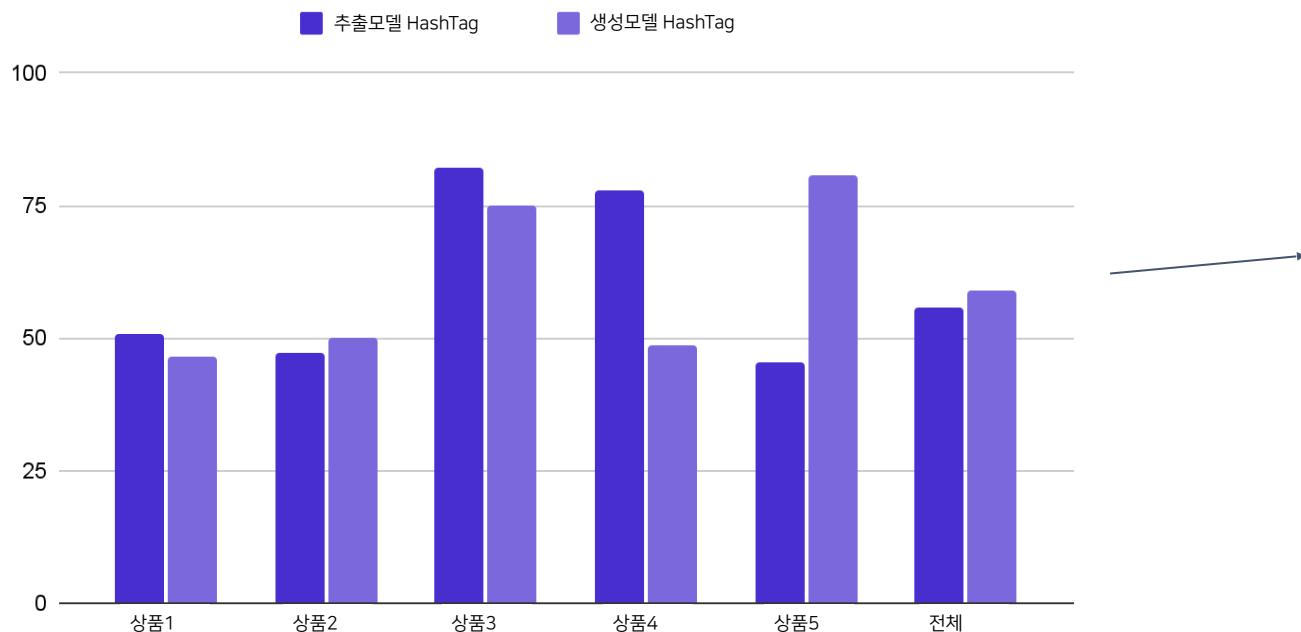
#갤럭시북플렉스알파 #갤럭시북플렉스

#삼성노트북플러스 #삼성노트북

#노트북

## 5.3 HashTag 평가

- 설문조사 개요
  - 평가인원: 50명
  - 평가내용: 모델로 생성된 HashTag 적합도 및 향후 사용의향 평가
- 설문조사 결과 – HashTag 적합도

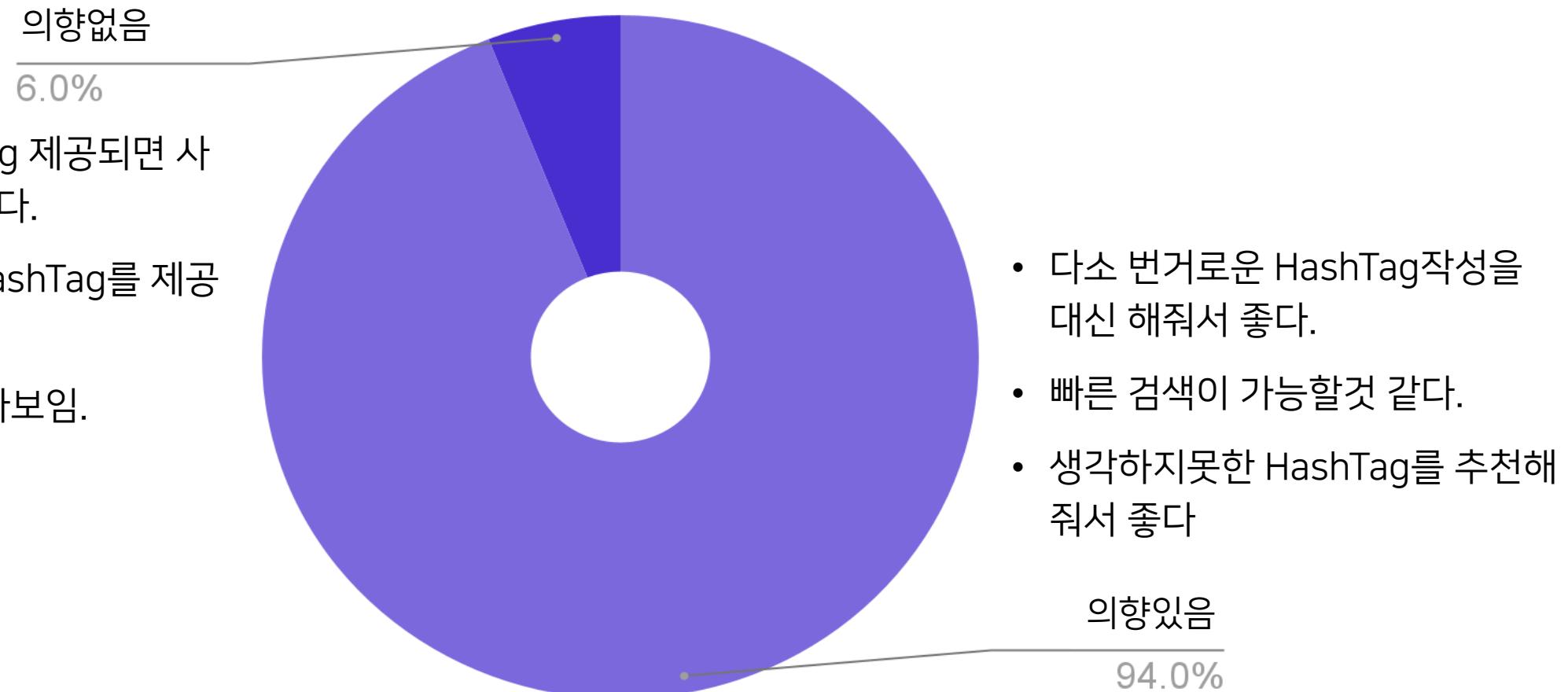


### 부적합한 이유:

- 상품과 연관없는 HashTag 추천
- 상품설명에는 있지만 굳이 상품검색할때는 쓰지 않을것 같음
- HashTag가 구체적이지 않음
- 미개봉급 → 미개봉 같이 다소 혼란이 올 수 있는 단어가 들어감
- 다소 HashTag가 길게 생성됨

## 5.3 HashTag 평가 (cont.)

- 설문조사 결과 - 향후 사용의향 평가



## 6. 향후 개선 방안

# 6. 향후 개선 방안

---

- HashTag 추출모델
  - 새로운 데이터를 통한 지속적인 업데이트 및 Vocab 고도화
  - 해시태그로 사용될 단어로 분리해 낼 수 있는 NER Model 개발
- HashTag 생성모델
  - 사용자가 직접입력한 데이터를 수집, 모델에 지속적인 업데이트
- 공통
  - 가전/디지털 이외의 다양한 카테고리로 범위 확대
  - 중고거래 플랫폼에 특화된 HashTag 생성

# Q&A