DACON 수상작리뷰
[쇼핑몰 리뷰 평점 분류 AI 해커톤 - DACON]

[쇼핑몰 리뷰 평점 분류 AI 해커톤]

◈ 주제 및 문제점

👉 쇼핑몰 상품 리뷰 텍스트 분석을 통한 평점 예측

- 상품 리뷰 텍스트와 평점 사이의 관계
- 다중분류 (1점, 2점, 4점, 5점)
- 평가기준: accuracy
- 네이버 쇼핑 데이터

◈ 데이터

- train.csv 샘플 ID, 쇼핑몰 리뷰 텍스트, **상품 평점 (target)**
- test.csv : 샘플 ID, 쇼핑몰 리뷰 텍스트
- sample_submission.csv : 샘플 ID, **상품 평점 (target)**

◈ 코드 리뷰

(1) 모델 파라미터 세팅, 가중치 저장 폴더 설정

🔷 사용한 모델: 사전 학습된 'kykim/electra-kor-base', 'kykim/funnel-kor-base'

✅ Hugging Face Hub에 있는 한국어 사전 학습 언어 모델

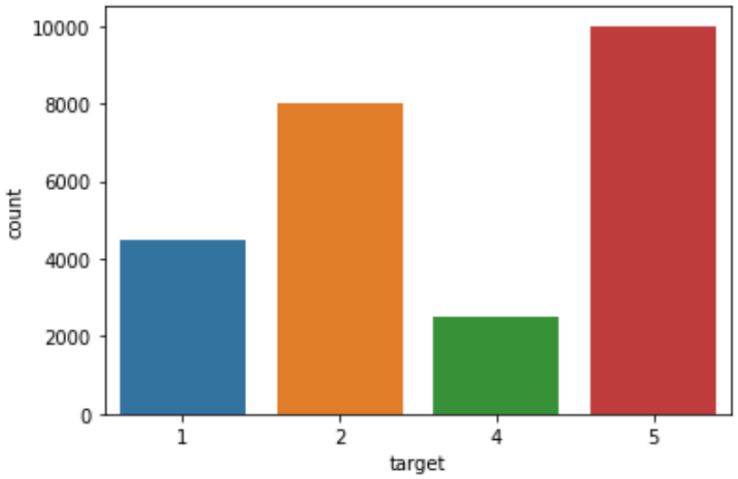
① kykim/electra-kor-base: electra 기반, Discriminator-Generator 구조, 용도 (한국어 문장분류, 감정분석, 질의응답 등)

② kykim/funnel-kor-base: funnel transformer 기반, 토큰 시퀀스를 압축하면서 처리, 긴 한국어 문서 처리에 유용

🔷 가중치 저장 폴더

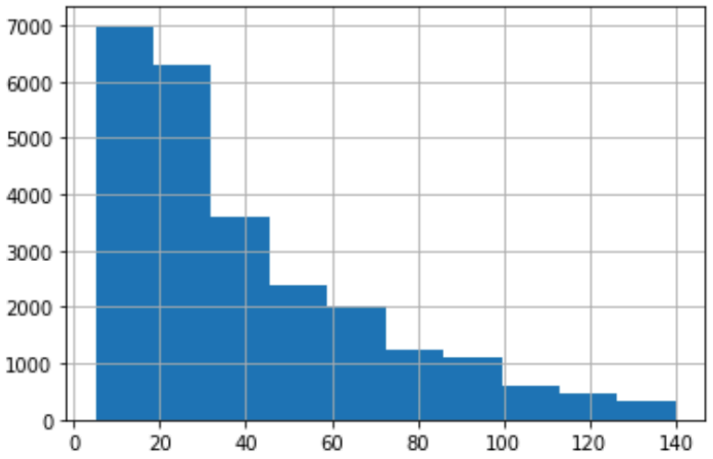
(2) Data Load & EDA

🔷 타겟값 분포



- 타겟값 매핑 {1:0, 2:1, 4:2, 5:3}

🔷 리뷰 텍스트 길이 분포



(3) 텍스트 전처리

🔷 텍스트 정제

- 이모지 처리 (!pip install emoji==0.6.0) : 모든 이모지를 포함하는 긴 문자열을 생성, 이모지도 허용 문자로 포함시키기 위한 처리
- 한글·영어·숫자·특정 특수문자·이모지 빼고 다 제거
- URL 제거
- 특수문자, 불필요한 외국어 문자 → 공백
- 문자열 양 끝 공백 제거
- 같은 문자 반복 → 2번으로 축약

🔷 띄어쓰기 검사

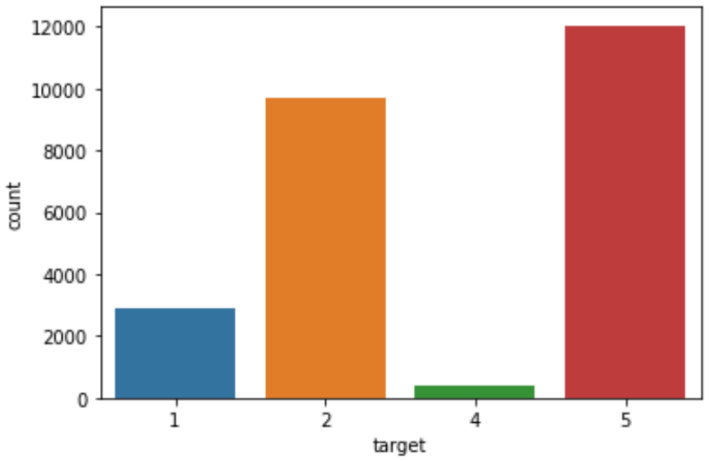
- pykospacing 패키지 사용

✅ PyKoSpacing 패키지

- 문장에서 잘못된 띄어쓰기를 찾아 올바른 위치에 공백 삽입

(4) 모델 학습 및 예측

- 🔷 Before: Accuracy 계산 함수 생성, 데이터셋, 조기중단 설정
- 🔷 StratifiedKFold 이용
  - fold 마다 평가 루프 종료시, 최소 loss와 그때의 accuracy 출력
  - fold 마다 최소 loss가 작은 모델들의 평균 정확도, 평균 손실 출력
- 🔷 예측: 앙상블 활용
  - 최종 예측된 타겟값 분포



◈ 다른 수상작

- 2위: 타겟값의 데이터 불균형 → 데이터 증강 방법 사용 / 마찬가지로 HuggingFace Transformer 활용, KoElectra&RoBERTa 모델 학습
- 3위: Back Translation 처리 (영어, 일본어), HuggingFace 사용

◈ **배울점**