DACON 수상작리뷰
[Private 1st] Pre-Processing + BERTweet-large + K-fold - DACON

[2023 용인시 SW/AI 해커톤]

◈ 주제 및 문제점

👉 트위터 영어 텍스트에 포함된 감성(pos,neg,neu) 분류

- 평가기준: Macro F1 score

◈ 데이터

- train.csv : id, text(트위터 텍스트), sentiment(target)
- test.csv : id, text(트위터 텍스트)
- sample_submission.csv : id, sentiment(target) (neu-0, pos-1, neg-2)

◈ 코드 리뷰

(1) 사전 학습 모델

🔷 BERT, RoBERTa, BERTweet 시도 ⇛ BERTweet 선택

(2) 테스트 정규화 (전처리)

🔷 유저이름 → @user 로 변환
🔷 사이트 주소 → http 변환
🔷 영어문자 이외 텍스트 → 제거

(3) 모델 학습

🔷 **새로운 함수를 생성**

- bertweet-large 모델 사용
- K-FOLD 사용

◈ 배울점

- 사전 학습 모델 여러 개 시도, 최대 토큰 수 여러개 시도, 데이터 증강 시도를 하여 제일 좋은 성능을 낸 것을 보며 다양한 시도를 해보는 것이 중요하다고 생각함.
- 먼저 알고리즘 함수를 생성한 다음 데이터에 적용해보는 과정이 인상깊음.