**팀원 구성 및 역할**

트위터 크롤링, 감정분석 데이터셋 생성, 감정분석 모델 학습, 발표자료 및 발표

**담당역할에 따른 수행내용 및 성과**

트위터 공식 API 및 python libarary tweepy를 사용하여 트위터에서 내용을 바탕으로 트윗을 스크래핑하는 기능을 구현함.

텍스트이(가) 표시된 사진

자동 생성된 설명이러한 스크래핑 기능을 활용하여 암호화폐 상위 10종에 대한 언급을 포함하는 트윗을 수집한 뒤, 해당 내용들을 광고 및 불필요한 데이터인지, 아니면 실제 유저의 데이터인지 직접 labeling을 수행하여 데이터셋을 생성함.

총 1500여개의 데이터셋을 생성했고, 라벨링은 특정 사이트, 유튜브를 홍보하는 단순광고, 여러 트위터 봇의 트윗, 이벤트에 참여하는 일반유저 글은 YES로, 그외 일반 사용자 트윗은 NO로 라벨링을 진행하였음.

텍스트이(가) 표시된 사진

자동 생성된 설명

생성한 데이터셋을 바탕으로 pre-trained 된 distilbert 모델에 fine-tuning을 진행함.

학습 모델로 distilbert를 사용한 이유는, distilbert가 기존 bert 대비 모델의 크기가 40% 감소했고 연산속도는 60% 증가했으나 성능은 97% 수준을 유지하는, 퍼포먼스 위주의 모델이기 때문에 서버 성능을 위해 선택했음.

학습은 tensorflow와 keras, huggingface의 transformer와 trainer를 사용하여 dataset labeling, tokenizing, fine-tuning, evaluation 과정을 거쳐 학습을 진행했음.

학습 결과는 KERAS의 EVALUATE 기능을 사용하여 평가했음.

학습시 학습 데이터셋의 20%를 분리하여 테스트용 데이터셋으로 분리했고, 나머지 80%로 학습을 진행한 뒤 EVALUATE 기능을 사용하여 테스트용 데이터셋에 대한 결과 비교를 진행함

텍스트이(가) 표시된 사진

자동 생성된 설명

EVALUATION 결과는 accuracy 0.8366으로 약 84%의 정확도를 보였음.

**문제해결**

데이터셋 형성 과정에서 어디까지가 광고 및 불필요한 데이터고, 어디까지가 유효한 데이터인지 판단이 모호함. 이벤트 참여 글은 실제 유저 글이지만 감정분석을 통한 여론분석에 사용 가능한 데이터인가? 트위터 봇의 암호화폐 관련 뉴스는 유효한 데이터인가?

* 단순광고를 포함해 트위터 봇 전부, 실제 유저의 이벤트 참여 글 모두 불필요 데이터로 판단

학습 결과 정확도가 떨어지는 문제

* underfitting 문제라고 판단, EPOCH 횟수를 늘려 정확도를 끌어올림. overfitting이 발생하지 않도록 적절한 EPOCH=10으로 조정