## **1월 3주차 회의록 (21.01.16)**

4조 김태희, 안세훈, 정영훈, 고준호

**겨울방학 PLAN**

**1월 1-2주차**

- API 사용 가능여부 확인  
- 신뢰도 분석 모델 조사  
- KCELECTRA 모델 조사  
- UI develop 계획 수립

**1월 3-4주차**

- 도메인 type, range 설정 => 도메인 변경- 크롤링 자동화, 최적화, db 구축  
- 파인튜닝 스터디

**2월 1-2주차**

- 파인튜닝 테스트  
- 데이터 태깅 시작  
- 최종 UI 프로토 타입 구현

**2월 3-4주차**

- 서버 배포  
- 신뢰도 분석 모델 추가  
- 파인 튜닝 완료

**1월 3-4주차 진행사항**

**[API 활용 가능 여부 조사]**

**네이버 API**

- 네이버 플레이스 API는 제휴 업체만 활용가능  
- 자체 웹사이트와 회사명 등 제휴 제안서를 작성해야 되기 때문에 대학생이 이용하기는 어려움  
- Open API는 단순 검색 쿼리나 쇼핑 추이 등 리뷰 데이터는 활용불가

텍스트이(가) 표시된 사진

자동 생성된 설명

**구글 API**- 구글 맵에서 제공하는 open API는 존재  
- 하지만 리뷰양이 네이버에 비해 현저히 적음 (1/10) 수준  
- 리뷰 도메인으로는 적합하지 않음

**결론**

- 현재 구현되어 있는 구조 그대로 자체 크롤링 툴을 이용  
- 주기적으로 데이터 최신화 및 크롤링 최적화 필요  
- 전체 도메인이 아닌 범위를 정하여 정적인 DB 구축 예정   
(ex. 2021년도 까지의 성남시 전체 카페 리뷰)

**[모델링 스터디]**

기존 사용 모델: KoBERT => 신규모델 : KcELECTRA

**KCELECTRA**

공개된 한국어 Transformer 계열 모델들은 대부분 한국어 위키, 뉴스 기사, 책 등 잘 정제된 데이터를 기반으로 학습한 모델입니다. 한편, 실제로 NSMC와 같은 User-Generated Noisy text domain 데이터셋은 정제되지 않았고 구어체 특징에 신조어가 많으며, 오탈자 등 공식적인 글쓰기에서 나타나지 않는 표현들이 빈번하게 등장합니다. KcELECTRA는 위와 같은 특성의 데이터셋에 적용하기 위해, 네이버 뉴스에서 댓글과 대댓글을 수집해, 토크나이저와 ELECTRA모델을 처음부터 학습한 Pretrained ELECTRA 모델입니다. 기존 KcBERT 대비 데이터셋 증가 및 vocab 확장을 통해 상당한 수준으로 성능이 향상되었습니다. KcELECTRA는 Huggingface의 Transformers 라이브러리를 통해 간편히 불러와 사용할 수 있습니다.

**Performance**

테이블이(가) 표시된 사진

자동 생성된 설명

Reference: KcELECTRA <https://github.com/Beomi/KcELECTRA>

Experiment: <https://bab2min.tistory.com/657>

**Finetuning**  
파인튜닝 Code

<https://github.com/Beomi/KcBERT-finetune>

네이버 쇼핑 리뷰 및 스팀 게임 리뷰 파인튜닝

<https://github.com/jaehyeongAN/KoELECTRA-finetuned-sentiment-analysis>

**모델링 계획**

1. KcELECTRA -> 네이버 쇼핑리뷰 데이터로 Fine Tuning -> 태깅 데이터로 Finetuning  
=> 높은 정확도를 가진 범용적 긍부정 감성분석 모델 완성

2. BERT 기반의 pororo 주제 분류 모델 -> 사전 훈련 모델 그대로 사용

3. 신뢰도 분석 모델  
- 현재 프로젝트는 네이버 플레이스의 짧은 리뷰(1-3줄) 이므로 가짜리뷰 판별이 쉽지 않음  
- 가짜 리뷰 시스템은 대부분 블로그 크롤링을 통해 광고성 포스팅을 판별  
- 따라서 학습을 통한 모델 구축 보다는 자체 알고리즘을 통해 너무 짧거나 의미 없는 리뷰를 필터링 해주는 방식 고려 예정