

답변 요약서

6조 T-SAN(티끌모아 태산)

1. 기술적인 부분이 초기 계획서보다 많이 발전 했습니다. 너무 많은 기능을 구현하기 보다 핵심 기능의 구현을 위해 노력하면 좋을 듯 합니다. (이경용 교수님)

계획서상의 내용에서 추가하여 1차 중간 발표에서 공공 데이터로 반환하는 기능을 추가했습니다. 하지만 정책적인 부가 기능을 추가하다보니, 기존 플랫폼의 비전인 효율적인 비용으로 데이터셋을 생성한다는 부분에서 많이 벗어나는 것 같다고 판단하였습니다. 그래서 이 내용을 삭제하고 교수님의 피드백에 따라 핵심 기능 구현에 집중하는 것으로 결정했습니다.

또한, 검수 과정에서 2차 레이블링을 요청하는 것을 삭제하였습니다. 2차 레이블링이 필요하다고 판단하는 것과 이를 위해 검수 작업을 중단시키고 있는 것에 한계가 있고 비효율적이라고 판단하였기에, 최대한 검수 과정에서 레이블링을 확정하는 것으로 수정하였습니다.

2. 잘못된 레이블링에 대한 검수를 할 때, 판별하는 구체적인 척도(기준)가 있는가? (이경용 교수님)

검수에 사용하는 척도는 유저의 신뢰도, 해당 프로젝트에서의 유저의 정답 개수가 될 것입니다. 검수 작업 과정에서 큰 분기점은 임시 레이블과 모델의 예측 레이블의 일치 여부입니다. 임시 레이블은 유저의 레이블들 중 유저의 신뢰도를 이용해 신뢰도가 높다고 판단한 레이블을 지정한 것입니다. 그래서 두 레이블이 일치 한다면 그 레이블이 옳바르다고 판단할 것입니다. 하지만 불일치 한다면 임시 레이블이 옳바른지, 모델의 예측 레이블이 옳바른지 판단 하는 과정이 필요합니다. 이 때는 임시 레이블 지정 이전의 유저의 레이블들을 확인합니다. 모델의 예측 레이블과 일치하는 유저의 레이블이 존재 한다면 이전에 옳바르다고 판단한 레이블들을 바탕으로 그 유저의 정답 개수를 구해서 검수 작업에 사용하게 됩니다. 현재 이 과정을 통해 잘못 지정된 임시 레이블이 모델의 예측 레이블로 조정되는 것을 확인하였습니다. 또한, 유저의 신뢰도를 바탕으로 임시 레이블의 신뢰도를 계산하여 신뢰도 상위 수 천개 데이터를 필터링하여 모델의 학습에 사용함으로써, 모델의 성능을 높이는 요소로 사용하고 있습니다.