

NLP 기초 프로젝트

News-HeadLine Topic Classification With a Data-Centric Approach

TEAM : 아기더키들(NLP-09조)



MEMBER : 한동훈_T7440

곽희준_T7308

김정은_T7325

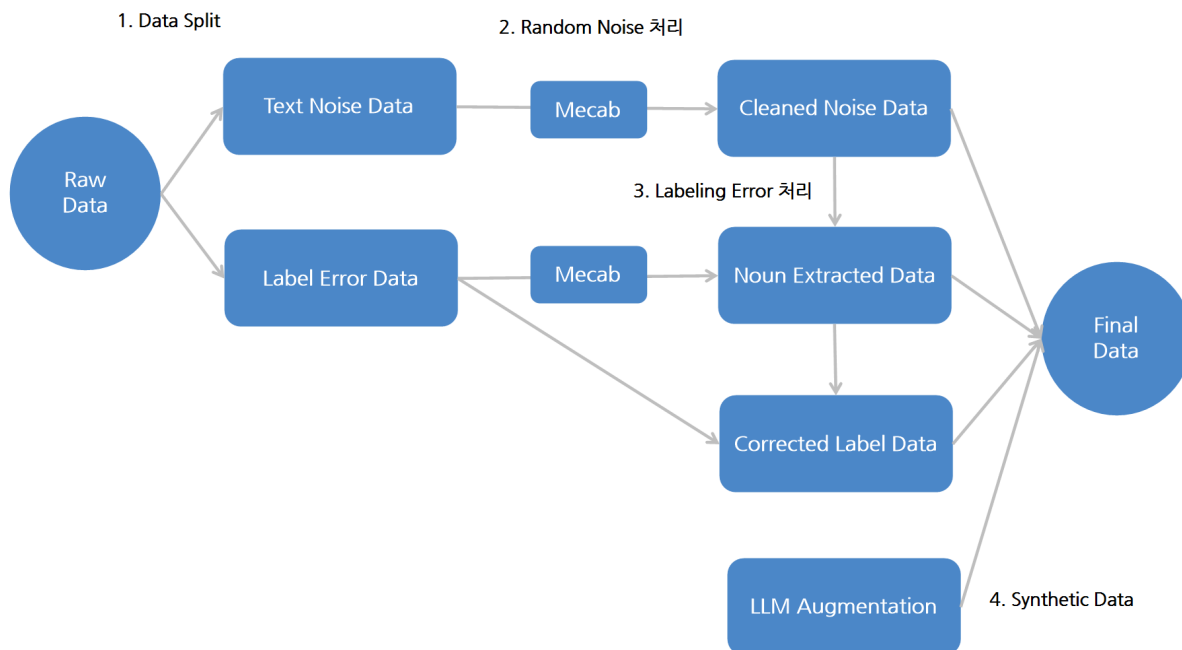
박동혁_T7335

[1. 대회 개요]

본 대회는 뉴스의 헤드라인을 통해 그 뉴스가 어떤 topic을 갖는지를 분류해내는 task로, 베이스라인 모델을 수정하지 않고 데이터 수정만으로 성능을 향상시키는 **Data-Centric** 접근 방법을 요구함. 공개된 생성 모델이나 **Data Augmentation, Filtering, Sampling** 등 다양한 데이터 조작 기법을 통해 성능을 개선할 수 있으며, 유료 API 사용 및 외부 데이터셋 활용은 금지함.

총 2800개의 train 데이터 중 1000개는 라벨을 임의로 바꾸었으며(**Labeling Error**), 1600개는 text에 임의의 char 중 20~80%를 다른 아스키코드로 대체하였으며(**Random Noise**), 나머지 200개의 데이터만 정상적으로 존재함.

[2. 프로젝트 수행 내용]



Preprocessing

1. Data Split

- Train 데이터의 text에서 noise가 없는 데이터와 noise가 있는 데이터로 각각 나누어 csv 파일로 저장
- Raw 데이터 확인 결과, 한글과 띄어쓰기를 영어, 숫자, 특수문자 등으로 대체하여 Noise를 추가함. 해당 [char \(아스키코드 33~126\)가 문장 내 20% 기준으로 존재](#)하는지 유무로 데이터 분리를 진행
- Text Noise 데이터 : 1608개 / Label Error 데이터 : 1192개

2. Random Noise 처리

- Text에 noise가 있는 데이터를 형태소 분석기 [Mecab을 활용하여 명사만 추출](#)하여 저장. 명사가 추출되지 않는 데이터는 제거 (Text Noise 데이터 : 1608개 -> 1606개)
- 위의 데이터를 기반으로 labeling에 오류가 있는 데이터를 relabeling하기 위해 text에 noise가 없는 데이터도 명사만 추출 (Label Error 데이터 : 1192개)

3. Labeling Error 처리

- 전처리한 text noise 데이터를 [klue/base-bert 모델에 학습](#)시키고 명사만 추출한 text noise가 없는 데이터를 relabeling을 진행
- Relabeling한 데이터의 ID를 기준으로 명사 추출 전 clean한 데이터의 labeling을 덮어씌움

Augmentation

1. Synthetic Data

- LLM을 활용하여 다양한 주제의 뉴스 헤드라인을 생성하여 Augmentation을 진행
-> 사용한 모델 : [allganize/Llama-3-Alpha-Ko-8B-Instruct](#)
- Noise 처리 : 생성한 text에도 noise가 존재함을 확인하여 Mecab을 활용하여 명사만 추출 (증강 데이터 : 7338개)


- **Label** 추가 : 앞서 relabeling한 방식과 동일한 모델로 생성한 text에 맞는 labeling을 진행

Baseline Model

1. Train Data 학습

- 앞서 진행한 [Random Noise 제거한 데이터 + Labeling Error 처리한 데이터 + Mecab을 적용한 Clean 데이터 + LLaMA를 활용해 증강한 데이터](#)를 concat하여 Baseline Model에 학습 (최종 데이터 : 11330개)

[3. 수행 결과]

내등수 9	NLP_09조		0.8420	0.8405	29	23h
----------	---------	---	--------	--------	----	-----