

# 1. 개요

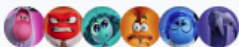
## A. Data-Centric 주제 분류 프로젝트

텍스트 주제 분류는 다양한 형태의 문서를 효율적으로 분류하고 정리할 수 있는 중요한 작업으로, 특히 뉴스 기사와 같은 대량의 텍스트 데이터를 체계적으로 관리하는 데 필수적이다. 본 프로젝트에서는 한국어 뉴스 기사 제목을 7개 카테고리 로 분류하는 작업을 수행했다.

본 프로젝트의 특징은 모델 개선에 집중하는 model-centric 접근 방식이 아닌, 데이터 품질 향상에 초점을 맞춘 data-centric 접근 방식을 채택했다는 점이다. 주어진 데이터에는 ASCII 코드 노이즈와 잘못된 레이블이 섞여 있으며, 데이터 전처리와 증강만으로 시스템 성능을 향상시키는 방식이다.

## B. 결과

- 평가 기준: Macro F1-score (모든 레이블에 동일한 중요도 부여)
- 성능 평가: 0.8321 (베이스라인 0.5980 대비 +0.2341 향상)

Rank	Team Name	Team Member	accuracy	f1
My Rank 13	NLP_06조		0.8348	0.8321

[그림1] 리더보드 최종 결과

# 2. 데이터 전처리

## A. 텍스트 노이즈 정제

텍스트 데이터의 노이즈 정도를 정량적으로 측정하기 위해 두 가지 지표를 활용했다. 첫째, ASCII 코드 문자의 비율로, 한글이 아닌 특수문자나 영문자 비중이 높을수록 텍스트 오염 가능성이 높다고 판단했다. 둘째, 형태소 분석을 통해 한국어 토큰 비율을 측정하여 의미 있는 한국어 단어나 문장 구조의 보존 정도를 확인했다.

이렇게 식별된 오염 텍스트는 LLaMA 모델을 활용한 few-shot 학습과 프롬프트 엔지니어링을 통해 정제했다. 노이즈가 경미한 데이터는 별도 정제 과정을 거치지 않았으며, 심각한 손상으로 복구가 불가능한 텍스트의 경우 "복구 불가"를 출력하도록 했다. 노이즈 정도에 따른 구체적인 텍스트 정제 예시는 [표1]과 같다.

노이즈 수준	원본 텍스트	정제된 텍스트
약함	美성인 6명 중 1명꼴 배우자·연인 빛 띠안은 적 있다	-
중간	m 김정) 자주통일 새,?r열1나가야1보	'김정은' 자주통일 새 시대 열린다...나가야 할 보안의 시대
	pl美대선앞두고 R2fr단 발] \$비해 감시 강화	미 대선 앞두고 R2F단 발... 비해 감시 강화
심함	x콩-P면금[T 나%\...g트=물J1h나고 지>철 %기고	"복구 불가"

[표1] 노이즈 수준에 따른 원본 및 정제 텍스트 예시

## B. 레이블 노이즈 정제

cleanlab 라이브러리를 활용해 원본 학습 데이터의 레이블 품질 점수를 측정한 결과, 0.47로 나타나 전체 데이터의 절반 이상에서 레이블 신뢰도가 낮음을 확인했다. 이를 해결하기 위해 두 가지 자동화된 재레이블링 방식을 적용하였다. 첫째, Llama, SBERT, K-means 군집화를 결합한 방법으로, Llama로 생성한 자연어 레이블이나 원본 텍스트를 SBERT로 임베딩한 후 군집화하여 대표 레이블을 재부여했다. 둘째, cleanlab 기반의 self-training 방식으로, 일부 복구된 데이터로 학습한 baseline 모델을 통해 레이블 오류를 검증하고, cleanlab이 예측한 새로운 레이블을 부여했다.

### 3. 데이터 증강

데이터셋의 크기와 다양성을 확장하기 위해 데이터 증강을 수행했다. Validation 결과를 바탕으로 데이터가 부족하거나 F1 점수가 낮은 레이블을 대상으로 텍스트 노이즈가 중간 수준인 데이터를 선별적으로 증강했다. 주요 기법으로 MLM과 역번역을 사용했다.

#### A. MLM (Masked Language Model)

MLM 기법은 텍스트의 문맥적 자연스러움을 향상시키기 위해 적용되었다. ASCII 코드나 [UNK] 토큰을 [MASK] 토큰으로 대체한 후 MLM 모델이 예측하도록 했다. 반복 등장하는 토큰은 후처리를 통해 중복을 제거하여 텍스트의 자연스러움을 개선했다.

#### B. 역번역 (Back Translation)

역번역 기법은 데이터의 다각적 표현을 얻어 문장의 다양성을 높이기 위해 사용되었다. Google Translator, DeepL API, Meta AI의 NLLB 등 다양한 번역 도구와 모델을 활용했다. 이를 통해 텍스트의 표현이 풍부해지고 모델 성능이 개선되는 효과를 얻었다.

### 4. 합성 데이터 생성

#### A. LLaMA 프롬프트 엔지니어링

모델이 각 레이블의 특성을 더 잘 학습할 수 있도록 LLaMA 모델에 프롬프트 엔지니어링을 적용하여 뉴스 기사 제목을 생성했다. 우선 레이블에 대해 임의로 선택된 데이터를 few-shot 예시로 제공하여, 모델이 레이블별로 다양한 표현을 학습하도록 유도했다. 이에 더해 validation 결과 분석을 통해 모델이 특정 레이블을 잘 탐지하지 못하는 유형을 파악했다. 주요 오분류 문제(예: 생활문화 또는 과학 뉴스를 사회로 오분류, IT 기업의 경제 뉴스를 IT과학으로 오분류)에 대해 프롬프트 엔지니어링과 few-shot 예시를 적용하여 각 유형별로 합성 데이터를 추가 생성했다. 이를 통해 전반적인 분류 성능을 개선할 수 있었다.

처리 방식	데이터
MLM	기존: pWd 2018 바^셀"나는 갤럭시Sc 맞! 준비]한창...5천여명 몰릴 W 생성: 갤럭시S6 .0삼성 바셀나는 갤럭시S 맞팔 준비에한창...천여명 몰릴듯
BT	기존: 듀얼심 아이폰 하반기 출시설 솔솔...알뜰폰 기대감 역번역: 듀얼심 아이폰 하반기 기대감 출시
PE(label = 3)	생성 1: 공동주택 수요자 1명당 1만 원의 장기 임대 보조금 지급 생성 2: 코로나19 이후 재건을 위한 정책 개선 필요성 논의

[표2] 증강 및 합성 데이터 생성 예시

### 5. 최종 결과

본 프로젝트에서는 체계적인 데이터 전처리 및 증강 과정을 통해 최종 데이터셋을 구성했다. 각 팀원이 시도한 다양한 방법 중 가장 효과적이었던 기법들을 통합하여 최종 데이터셋을 구성했다. 이 과정에서 중복된 텍스트 데이터를 제거하여 데이터 품질을 향상시켰고, 동일 텍스트에 상이한 레이블이 할당된 경우에는 Cleanlab 라이브러리를 활용하여 데이터 레이블을 재할당함으로써 데이터 일관성을 확보했다. 이를 통해 최종적으로 F1 점수 0.8321을 달성했으며, 데이터 정제 및 증강이 실질적인 성능 향상에 기여했음을 확인했다. 각 팀원의 개별 시도와 성능 개선 과정, 그리고 최종 결과에 대한 자세한 내용은 부록에서 확인할 수 있다.

## Appendix. 실험 결과

아래 표에서 각 팀원의 개별 시도 및 성능 개선 결과를 확인할 수 있다.

리더보드의 베이스라인 F1 점수는 0.5980이었으며, 이 점수를 기준으로 각 실험의 성능을 비교했다.

구분	팀원	데이터셋 버전	F1-Score (Test)
개별 시도	오수현	text filtering(LLaMA)	0.6677
		label filtering(CleanLab)	0.7380
		합성 데이터(LLaMA)	0.7553
		증강(Synonym Replacement)	<b>0.7627</b>
	이상의	ASCII Noised	0.3922(f1 valid)
		ASCII Noised + MLM	0.7449
		ASCII Noised + MLM + BackTranslation	0.8215
		ASCII Noised + MLM + BackTranslation + Relabeling	<b>0.8307</b>
	이정휘	데이터 전처리 - 레이블(KMeans), 텍스트(Gemma) 정제	0.6561
		데이터 증강 - 역번역(NLLB)	0.6574
		데이터 증강 - 합성데이터(Gemma)	0.6660
		데이터 전처리 - 레이블(Cleanlab) 정제	<b>0.6775</b>
	정민지	텍스트 정제 (LLaMA)	0.5021
		레이블 정제 (LLaMA, SBERT, KMeans)	0.6876
		합성 데이터 - 모든 레이블 별 100개 생성 (LLaMA)	0.7183
		합성 데이터 - 품질 낮은 3개 레이블 별 200개 생성 (LLaMA)	<b>0.7458</b>
데이터 종합	서태영	데이터 전처리(ASCII) + relabeling(llama)	0.1732
		데이터 병합 - 중복 제거	0.8276
		데이터 병합 - 중복 relabeling	<b>0.8289</b>