

# 주제 분류 프로젝트

Data Centric

NLP\_15팀 (십오조가십오조)

# 김진재  
# 박규태  
# 윤선웅  
# 이정민  
# 임한택

# # Team Introduction



**김진재**



**박규태**



**윤선웅**



**이정민**



**임한택**

**Clustering**

**Data Augmentation**

**LLM Relabeling**

**EDA**

**Data Denoise**

**Data Augmentation**

**LLM Relabeling**

**Data Augmentation**

**Clustering**

**LLM Relabeling**

**Data Denoise**

**Data Augmentation**

**Clustering**

**Data Denoise**

**Back Translation**

**Data Augmentation**

# # Contents

## EDA

---

Exploratory Data Analysis

## Noised Data Processing

---

Augmentation w. LLM

Noun Removal

Back Translation

## Mislabeled Data Preprocessing

---

Augmentation w. LLM

Clustering

## Re-clustering

---

Re-clustering

## Review & Questions

---

Review & Questions

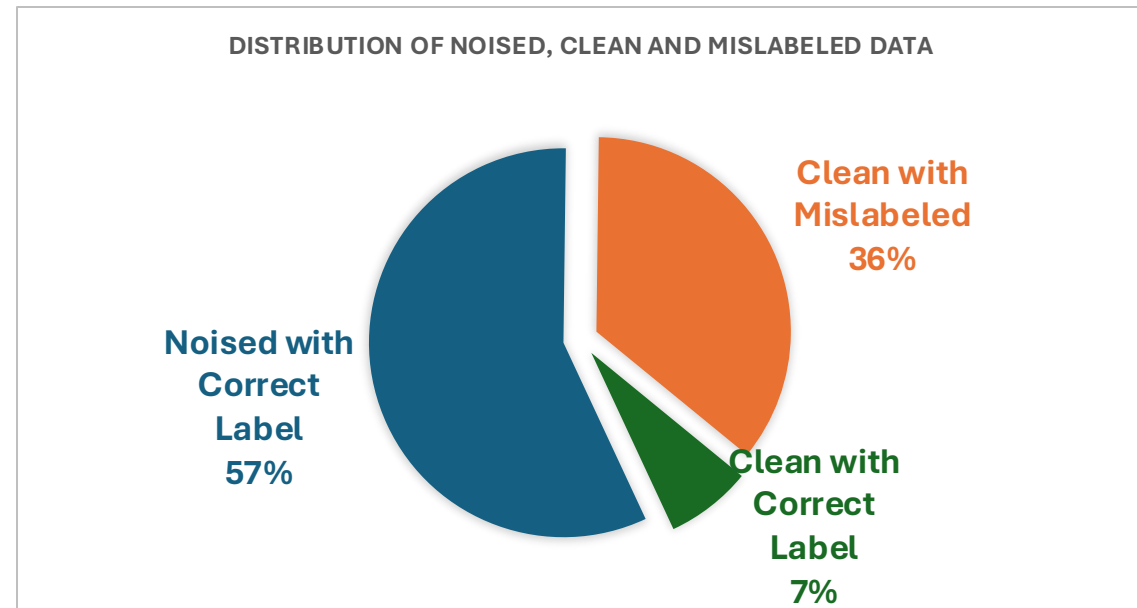
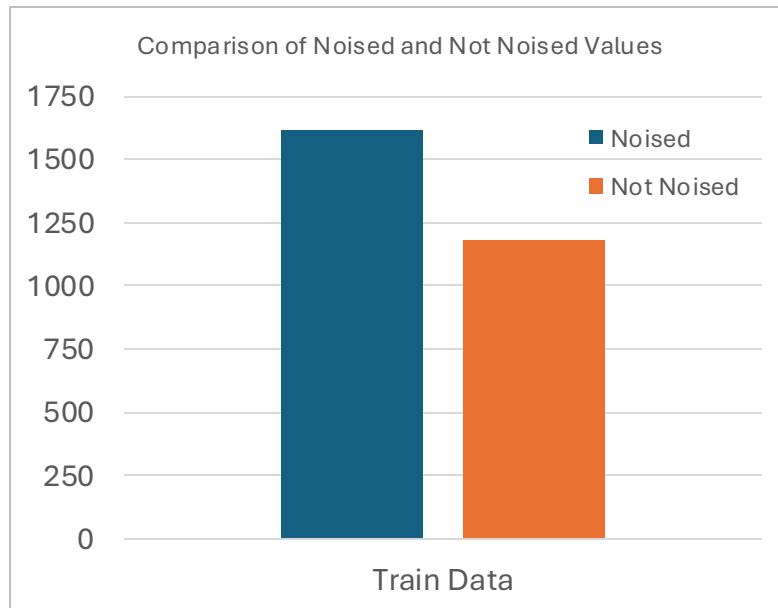
# EDA

---

## Exploratory Data Analysis

# Exploratory Data Analysis

- Train 데이터셋: 2800개
  - Noise 존재: 1600개 (57.14%)
  - Noise 없음: 1200개 (42.86%)
- 대회 제공 정보
  - 데이터 처리 중 Noised와 Mis-Labeled는 동시에 수행되지 않음



# Noised Data Processing

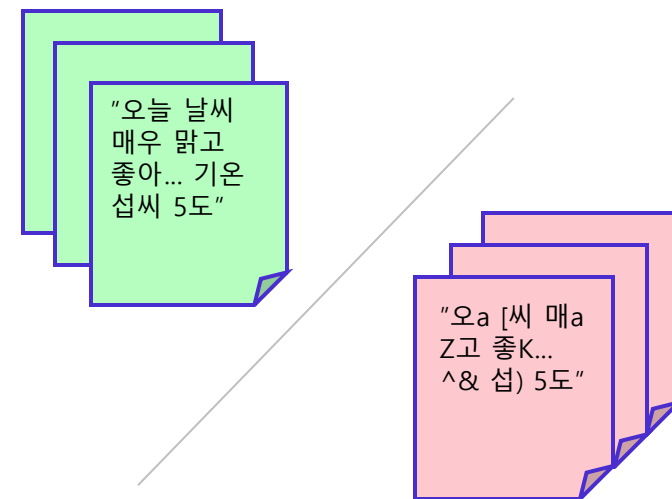
노이즈 데이터 처리

# Noised Data Preprocessing

- Noise 데이터 선별 작업 수행
  - Label 정보를 최대한 활용하기 위해 LLM을 활용한 Noised Data 선별 작업 수행
    - (LLM) **LGAI-EXAONE/EXAONE-3.0-7.8B-Instruct** 모델 활용: Temperature 0.1 + Prompt Engineering
  - 두 가지 버전의 Prompt를 활용하여 Noise 분류
    - Version 1. 2800개 Train Data 중 **1689개 Noise로 분류**
    - Version 2. 2800개 Train Data 중 **1619개 Noise로 분류**

당신에게 컨텍스트 하나가 주어질 때 노이즈가 낀 상태인지 파악해야 합니다. 노이즈의 종류는 특수문자와 영어 철자입니다. 단 '...', ',', '...' 과 한자는 노이즈가 아닙니다. ... 설명을 붙이지 말고 노이즈가 끼었던 거라면 'noised', 노이즈가 없는거라면 'nanoise'를 출력하세요.

당신에게 컨텍스트 하나가 주어질 때 왼쪽과 오른쪽의 텍스트를 비교해서 오른쪽에 비해서 왼쪽이 노이즈가 많이 낀 상태인지 파악해야 합니다. 설명을 붙이지 말고 노이즈가 끼었던 거라면 'noised', 노이즈가 없는거라면 'nanoise'를 출력하세요.



# Noised Data Preprocessing – LLM Selection

Rank	Model Name	Model Size	AVG	AVG_kr_eval	...
...	...	...	...	...	
#8	ko-gemma-2-9b-it	<10B	0.6048	0.5071	
...	...	...	...	...	

[Horangi 한국어 LLM 리더보드](#)

Rank	Model Name	Model Size	AVG	범용적언어성능	...
#1	Gpt-4o-2024-08-06	api	0.8192	0.8105	
#2	chatgpt-4o-latest	api	0.8176	0.8036	
...	...	...	...	...	
#23	EXAONE-3.0-7.8B-Instruct	<10B	0.661	0.6529	
...	...	...	...	...	
#31	ko-gemma-2-9b-it	<10B	0.6021	0.5731	
...	...	...	...	...	

[Horangi: W&B Korean LLM Leaderboard](#)

3

- 참고 지표: W&B 한국어 리더보드
  - LLM의 한국어 관련 능력을 종합적으로 평가하여 순위를 매긴 리더보드
- 선정 기준
  - 성능 (W&B 한국어 리더보드 기준)
  - V100\*1 GPU에서의 구동 가능성
  - 휴리스틱
    - 여러 모델을 직접 import하여 사용
    - 출력 내용과 속도를 비교하여 경험적 선택



# Noised Data Preprocessing – LLM Selection

Rank	Model Name	Model Size	AVG	AVG_kr_eval	...
...	...	...	...	...	...
#8	ko-gemma-2-9b-it	<10B	0.6048	0.5071	...
...	...	...	...	...	...

[Horangi 한국어 LLM 리더보드](#)

Rank	Model Name	Model Size	AVG	범용적언어성능	...
#1	Gpt-4o-2024-08-06	api	0.8192	0.8105	...
#2	chatgpt-4o-latest	api	0.8176	0.8036	...
...	...	...	...	...	...
#23	EXAONE-3.0-7.8B-Instruct	<10B	0.661	0.6529	...
...	...	...	...	...	...
#31	ko-gemma-2-9b-it	<10B	0.6021	0.5731	...
...	...	...	...	...	...

[Horangi: W&B Korean LLM Leaderboard](#)

3

- **LGAI-EXAONE/EXAONE-3.0-7.8B-Instruct**
  - KoMT 벤치마크에서 높은 점수 기록
- **CohereForAI/aya-expense-8b**
  - Arena-Hard 벤치마크에서 높은 Win-Rate 달성
- **rtzr/ko-gemma-2-9b-it**
  - Horangi Leaderboard 상에서 높은 AVG 점수 기록

➡ 증강에는 최종 위 세 모델을 활용

# Noised Data Preprocessing – Augmentation w. LLM

- Noised 데이터의 라벨링 정보를 활용
  - Noised 데이터의 복원 및 복원된 데이터 기반 증강 수행
- 앞서 언급한 Gemma, EXAONE, aya 세 가지 LLM을 활용해서 Noised 처리된 데이터 복원 및 증강 수행
- 복원된 데이터를 이용하여 증강한 결과, 복원을 여러 번 거치며 텍스트를 지속 복원시키는 효과 확인

"당신은 신문 기자입니다. 당신은 신문 기사 제목을 새롭게 만드는 임무를 부여받았습니다. ... 이 때 모든 설명을 붙이지 말고 새로 만든 기사 제목 한 개만 출력하세요."

...

"오a [씨 매a  
Z고 좋K...  
^& 섭) 5도"



"오 날씨  
맑고 좋은...  
기온은 섭씨  
5도"

	증강 전	증강 후
# of datasets	1680	<b>24600</b> <b>(+1364%)</b>
F1 Score	0.7531	<b>0.8456</b> <b>(+9.25%P)</b>

# Noised Data Preprocessing – Noun Removal

- 성능 향상을 위한 명사 제거
  - 증강을 위해 LLM이 생성한 데이터에는 유사한 단어를 포함하는 경우가 빈번하다는 점 발견
  - 증강을 통해 얻은 데이터에서 많이 겹치는 명사 Text를 제거
  - 핵심 명사 데이터의 제거 방법론 연구
    - Okt 라이브러리 활용

< 특정 카테고리에 편향되어 등장하는 명사 >

Noun	Count	Target Count	Target List
갤럭시	183	4	4(163), 5(17), 1(2), 3(1)
이란	149	4	6(139), 5(6), 1(3), 3(1)
농구	83	4	1(74), 0(6), 3(2), 6(1)
프랑스	83	4	6(66), 0(8), 1(7), 2(2)

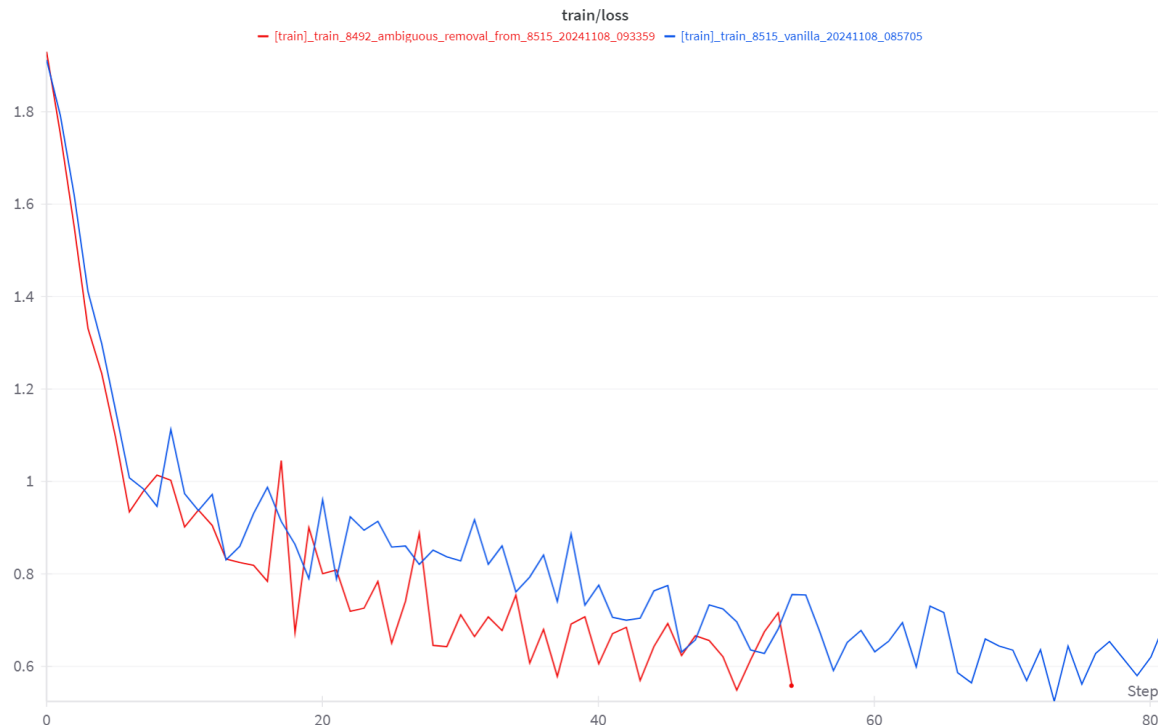
- Count: 데이터셋 내에서 등장한 횟수
- Target Count: 해당 명사가 등장한 라벨 종류의 개수
- Target List: {등장한 라벨}{라벨 내 등장 빈도}

< 여러 카테고리에서 유사하게 등장하는 명사 >

Noun	Count	Target Count	Target List
가치	19	6	5(9), 2(3), 3(2), 0(2), 6(2), 1(1)
대두	19	6	3(5), 2(4), 6(3), 1(3), 4(2), 0(2)
분위기	19	6	1(7), 2(5), 3(4), 0(1), 5(1), 4(1)
빅데이터	19	6	3(6), 4(4), 5(4), 1(2), 2(2), 6(1)

# Noised Data Preprocessing – Noun Removal

- 데이터 내에서 불확실한 명사 토큰이 하나 이상 존재하는 경우 해당 데이터를 제거
  - 데이터를 제거한 분량에 비해 성능 감소폭이 매우 적음
- 선별에 따른 **효과적인 제거 가능성** 확인



Public	Before	After
# of datasets	20,155	<b>7,668</b> <b>(-62%)</b>
F1 Score	0.8484	<b>0.8422</b> <b>(-0.6%P)</b>

Public	Before	After
# of datasets	15,588	<b>10,411</b> <b>(-34%)</b>
F1 Score	0.8515	<b>0.8492</b> <b>(-0.2%P)</b>

# Mislabeled Data Processing

오라벨 데이터 처리

# Mislabeled Data Preprocessing – Clustering

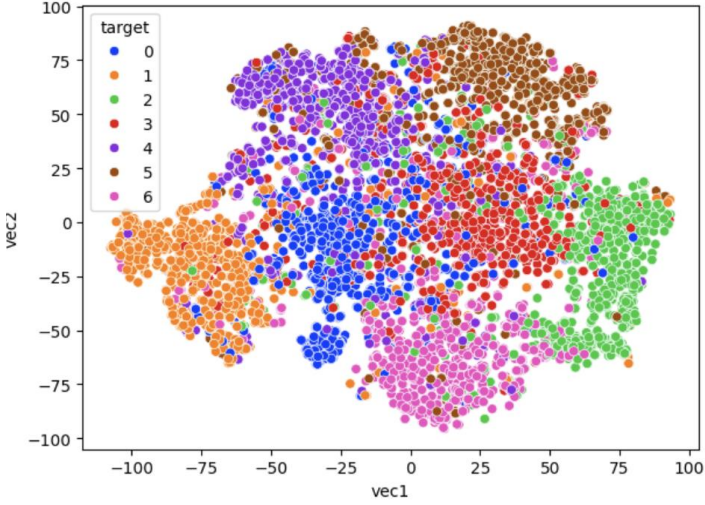
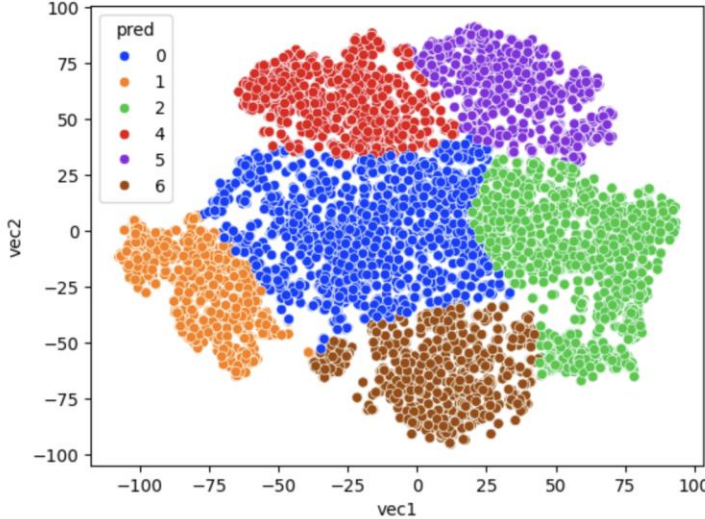
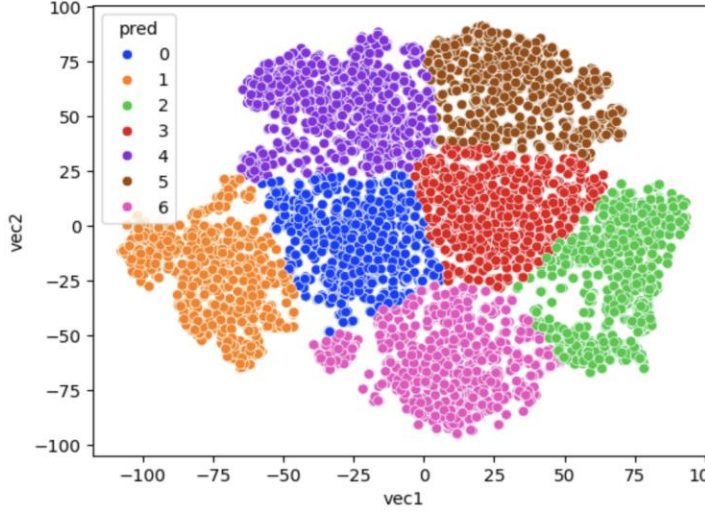
- Version 2. 2800개 Train Data 중 1619개 Noise로 분류
  - 나머지 데이터 1181개를 오라벨 데이터로 추정
- 각각 GMM(Gaussian Mixture Model), K-Means로 Re-labeling

## < GMM vs K-Means >

특징	GMM(Gaussian Mixture Model)	K-Means Clustering
모델링 방식	확률기반 모델, 각 클러스터를 가우시안 분포로 모델링	비확률적 모델, 클러스터 중심점(centroid)을 기반으로 작동
적합성	데이터의 분포가 가우시안일 때 효과적	데이터가 구형 분포일 때 효과적
결과 해석	각 클러스터에 대한 확률적 해석가능	각 클러스터에 대한 명확한 할당 가능

# Mislabeled Data Preprocessing – Clustering

- 임베딩 모델 : ssunbear/bert-base-finetuned-ynat (F1 : 0.8315)
- 당시 SOTA 데이터셋으로 klue/bert-base를 Fine-tuning

오라벨 데이터 분포	GMM (Gaussian Mixture Model)	K-Means Clustering
		
<p>원본 F1: 0.7806</p>	<p>적용 X 라벨 7개 -&gt; 라벨 6개</p>	<p>적용 O F1: 0.8143</p>

# Mislabeled Data Preprocessing – Clustering

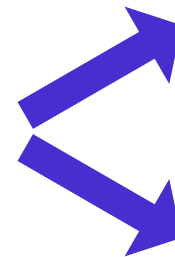
- Version 2. 2800개 Train Data 중 1619개 Noise로 분류
  - 나머지 데이터 1181개를 오라벨 데이터로 추정
- Mislabeled 데이터 1181개 교정 (F1: 0.8143) -> 3543개로 증강 (F1: 0.8283)
- 각각 **EXAONE-3.0-7.8B-Instruct**와 **aya-expanse-8b**로 Prompting 통해 증강

## [Instruction]

다음 뉴스 기사 제목과 동일한 카테고리에서 나올 기사 제목 1개를 생성 하시오.

답변은 오직 제목 1개 만을 포함해야 하며 불가능한 경우는 없다.

갤노트8 주말 27만대 개통...시장은 불법 보조금  
얼룩



## [LGAI-EXAONE/EXAONE-3.0-7.8B-Instruct]

갤럭시 S23 5G 출시 첫 주말, 15만대 이상  
판매...  
사전예약 초과 속 재고 부족 우려"

## [CohereForAI/aya-expanse-8b]

아이폰X 출시 후 갤럭시8 판매량과 불법  
보조금 경쟁 심화



# Backtranslation

---

역번역

# Back Translation

- 역번역 수행
  - 목적: LLM 증강에서의 단어 풍부성 부족을 해결하기 위해 토큰의 다양성 확보
  - 종류
    - 기본 역번역(Standard Back-Translation): 가장 기본적인 방법 (예시: 한국어 - 영어 - '한국어')
    - 순환 역번역(Circular Back-Translation): 여러 번의 반복 순환을 통해 품질 향상 (예시: 한국어 - 영어 - 한국어 - 영어 - '한국어')
    - 피봇 역번역(Pivot-Based Back-Translation): 중간에 다른 언어를 추가 (예시: 한국어 - 영어 - 일본어 - '한국어')
- API 사용량 제한(DeepL Free API)으로 인해, **기본 역번역** 방법 채택

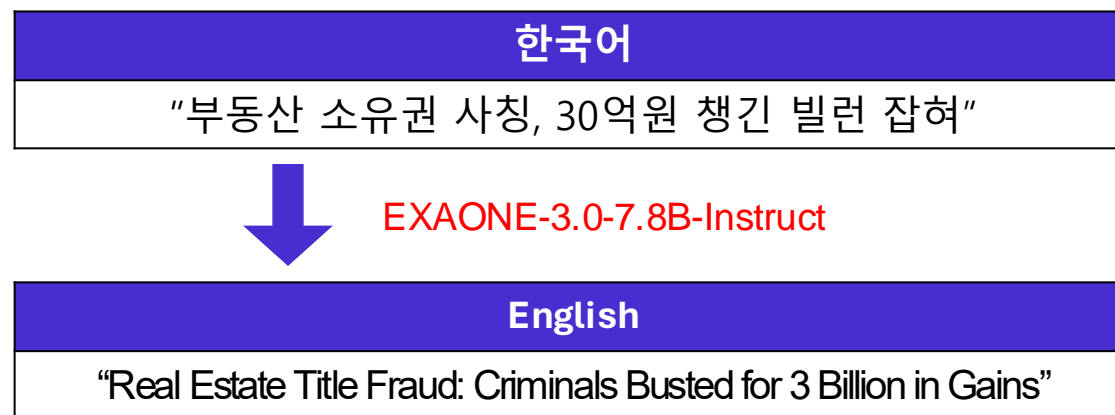
< 한국어 → 영어: LLM 모델을 이용한 번역 >

## [ System Message ]

당신은 뉴스 기자입니다. 입력받은 문장에서 노이즈를 식별하고 의미적으로 완전한 기사 타이틀을 생성한 것만 출력합니다.

## [ Prompt ]

어려운 영어 기사 타이틀로 바꿔주세요: {}



# Back Translation

- 역번역 수행
  - 목적: LLM 증강에서의 단어 풍부성 부족을 해결하기 위해 토큰의 다양성 확보
  - 종류
    - 기본 역번역(Standard Back-Translation): 가장 기본적인 방법 (예시: 한국어 - 영어 - ' 한국어 ')
    - 순환 역번역(Circular Back-Translation): 여러 번의 반복 순환을 통해 품질 향상 (예시: 한국어 - 영어 - 한국어 - 영어 - ' 한국어 ')
    - 피봇 역번역(Pivot-Based Back-Translation): 중간에 다른 언어를 추가 (예시: 한국어 - 영어 - 일본어 - ' 한국어 ')
- API 사용량 제한(DeepL Free API)으로 인해, **기본 역번역** 방법 채택

< 영어 → 한국어: DeepL Free API를 이용한 번역 >

```
...
import deepl
translator = deepl.translator("API_KEY")
...
results = translator.translate_text(batch,
                                   source_lang="EN",
                                   target_lang="KO")
```



# Back Translation

- 역번역 수행
  - 목적: LLM 증강에서의 단어 풍부성 부족을 해결하기 위해 토큰의 다양성 확보
  - 종류
    - 기본 역번역(Standard Back-Translation): 가장 기본적인 방법 (예시: 한국어 - 영어 - '한국어')
    - 순환 역번역(Circular Back-Translation): 여러 번의 반복 순환을 통해 품질 향상 (예시: 한국어 - 영어 - 한국어 - 영어 - '한국어')
    - 피봇 역번역(Pivot-Based Back-Translation): 중간에 다른 언어를 추가 (예시: 한국어 - 영어 - 일본어 - '한국어')
- API 사용량 제한(DeepL Free API)으로 인해, **기본 역번역** 방법 채택

## < 역번역 적용 결과 성능 분석 >

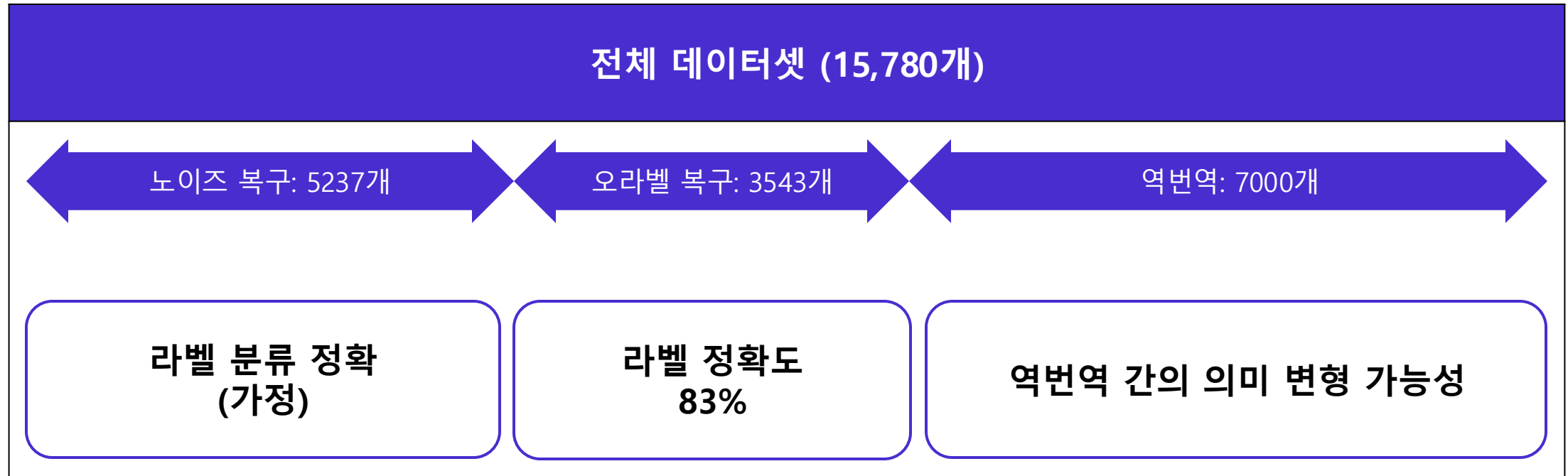
Public	Before	After
Accuracy Score	0.8607	<b>0.8611</b>
F1 Score	0.8564	<b>0.8571 (+0.07%P)</b>

# Re-clustering

---

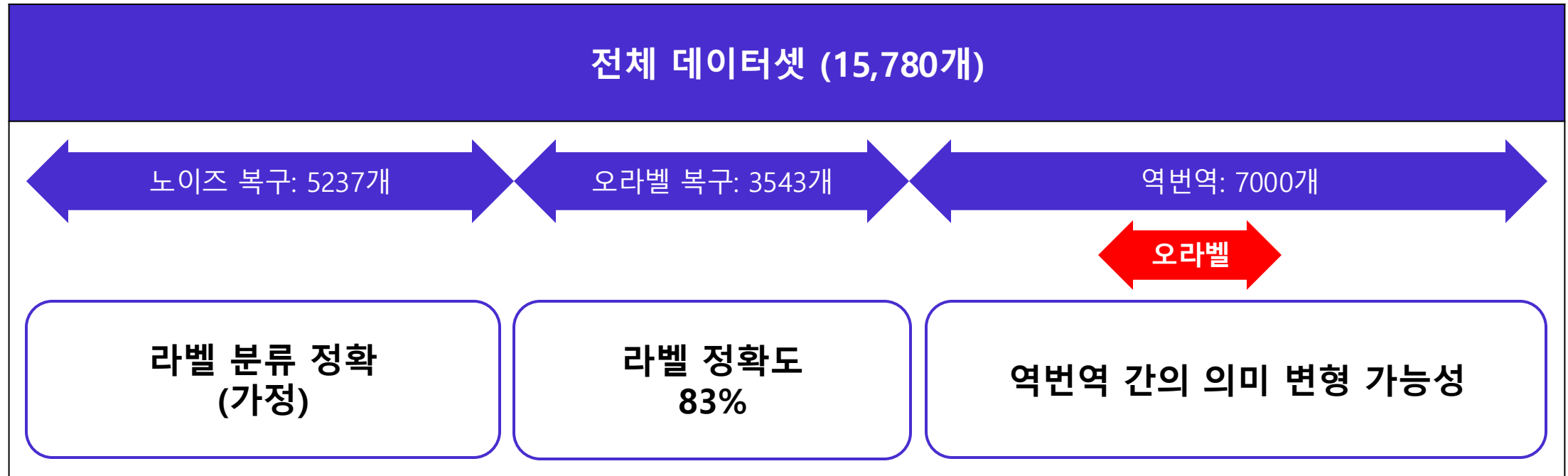
재군집화

# Re-clustering



- 클러스터링을 기반 모델 F1 0.8571점 달성
  - 다른 데이터셋들을 전체적으로 다시 클러스터링을 하여 라벨 재구성 필요성이 생김
- 클러스터링 모델이 F1 Score 1 가까이 기록하는 모델이 아니므로 일부 데이터셋만 재구성
  - 1000 ~ 2000개 단위로 여러 번 테스트

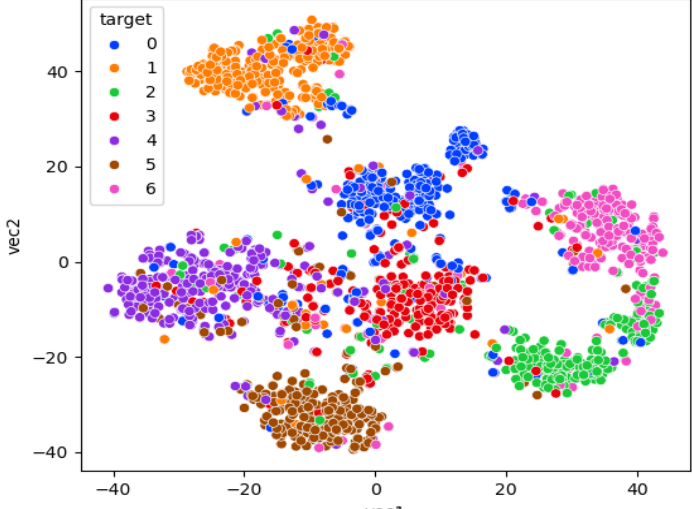
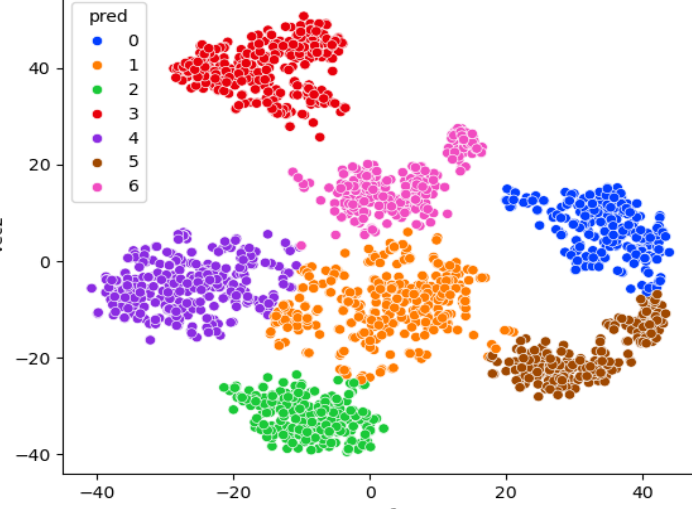
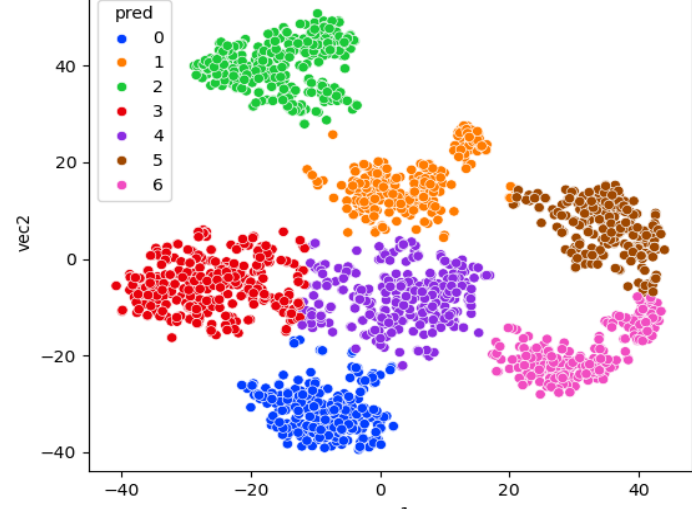
# Re-clustering



- 8780행 이후 : 역번역으로 클러스터링 모델과 다른 라벨 분포
  - 2000개의 데이터셋을 기준으로 300~448개의 Mislabeled 처리
- 가장 많은 Mislabeled을 배출한 10000 ~ 12000행 리클러스터링
  - 448개의 Mislabeled 보정

# Re-clustering

- 임베딩 모델 : ssunbear/bert-base-finetuned-ynat-v2 (F1 : 0.8571)
- 당시 SOTA 데이터셋으로 klue/bert-base를 Finetuning

오라벨 데이터 분포	GMM (Gaussian Mixture Model)	K-Means Clustering
		
역번역 데이터셋의 오라벨 분포 확인	435개 변형, 적용 X	448개 변형, 적용 O F1: 0.8584




# Re-clustering

< Re-clustering 적용 결과 성능 분석 (Public Dataset) >

Public	Before	After
Accuracy Score	0.8616	<b>0.8613</b>
F1 Score	0.8575	<b>0.8573 (-0.02%P)</b>

< Re-clustering 적용 결과 성능 분석 (Private Dataset) >

Private	Before	After
Accuracy Score	0.8623	<b>0.8623</b>
F1 Score	0.8583	<b>0.8584 (+0.01%P)</b>
	<b>SOTA #2</b>	<b>SOTA #1</b>

The background of the slide is a light blue gradient with several translucent, overlapping circles in various shades of purple, pink, and blue. The circles have a soft, glowing effect and are scattered across the frame.

감사합니다  
Q & A