INHA AI

2025 SW 중심대학 디지털 경진대회

AI 부문

TEAM	SINSA

2025.08.12.

미디어 리터러시

미디어를 비판적으로 분석하고 이해하며,
미디어를 책임감 있게 활용하고, 의미 있는 정보와
문화를 생산하고 공유할 수 있는 능력

Introduce

미디어 리터러시 1.0

이 뉴스가 진짜인가?

미디어 리터러시 2.0

이 글의 작성자는 사람? AI?

미디어 리터러시 1.0

이제는 근원적인 질문

생성형 AI의 등장은 정보 소비 방식을 바꿈

콘텐츠 내용 뿐 아닌, 출처부터 분별해야함

이 글의 작성자는 사람? AI?

미디어 리터러시 1.0

단계1. 전형적인 가짜뉴스

[단독] 인천 송도 해상서 길이 15m
거대 고래 사체 발견! “신종 심해어 소행 추정” 충격

검증 과정

출처 확인 → 교차 검증 → 전문가 검증 → 이미지 검증

미디어 리터러시 1.0

단계2. 사실과 허위를 교묘하게 섞은 정보

거짓

정부, 청년 주거안정 대책 발표... 강남 3구는 제외되어
"결국 부자 감싸기" 비판 여론 확산

검증 과정

핵심 사실 확인 → 핵심 주장 확인 → 의도 파악

미디어 리터러시 2.0

단계3. 생성형 AI를 활용한 그럴듯한 허위 정보

교묘한 거짓

[분석] 2025년 하반기, 대한민국 부동산 시장의 '퍼펙트 스톰'이 온다

검증 과정

저자 기관 확인 → AI 판별 도구 → AI 특징 분석

미디어 리터러시 2.0

단계3. 생성형 AI를 활용한 그럴듯한 허위 정보

교묘한 거짓

[분석] 2025년 하반기, 대한민국 주요 산업의 트렌드 분석 보고서

무분별한 수용의 위험,
사회적으로는 건강한 여론 형성 저해

생성형 텍스트 판별 챌린지를 통해
미디어 리터러시 역량 강화에 기여 의의

검증 과정

저자 기관 확인 → AI 판별 도구 → AI 특징 분석

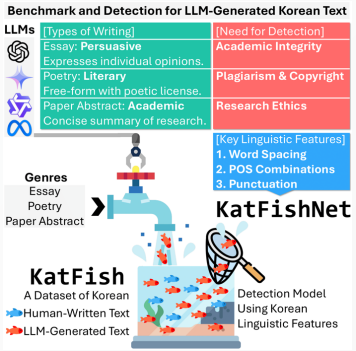
Data Analysis

데이터 분석

접근1) KatFishNet 아이디어 기반 언어 특성 분석 시도

KatFishNet: Detecting LLM-Generated Korean Text through Linguistic Feature Analysis

Shinwoo Park Shubin Kim Do-Kyung Kim Yo-Sub Han*
Yonsei University, Seoul, Republic of Korea
pshkhh@yonsei.ac.kr, shubs919@yonsei.ac.kr, kdky95@yonsei.ac.kr, emmous@yonsei.ac.kr



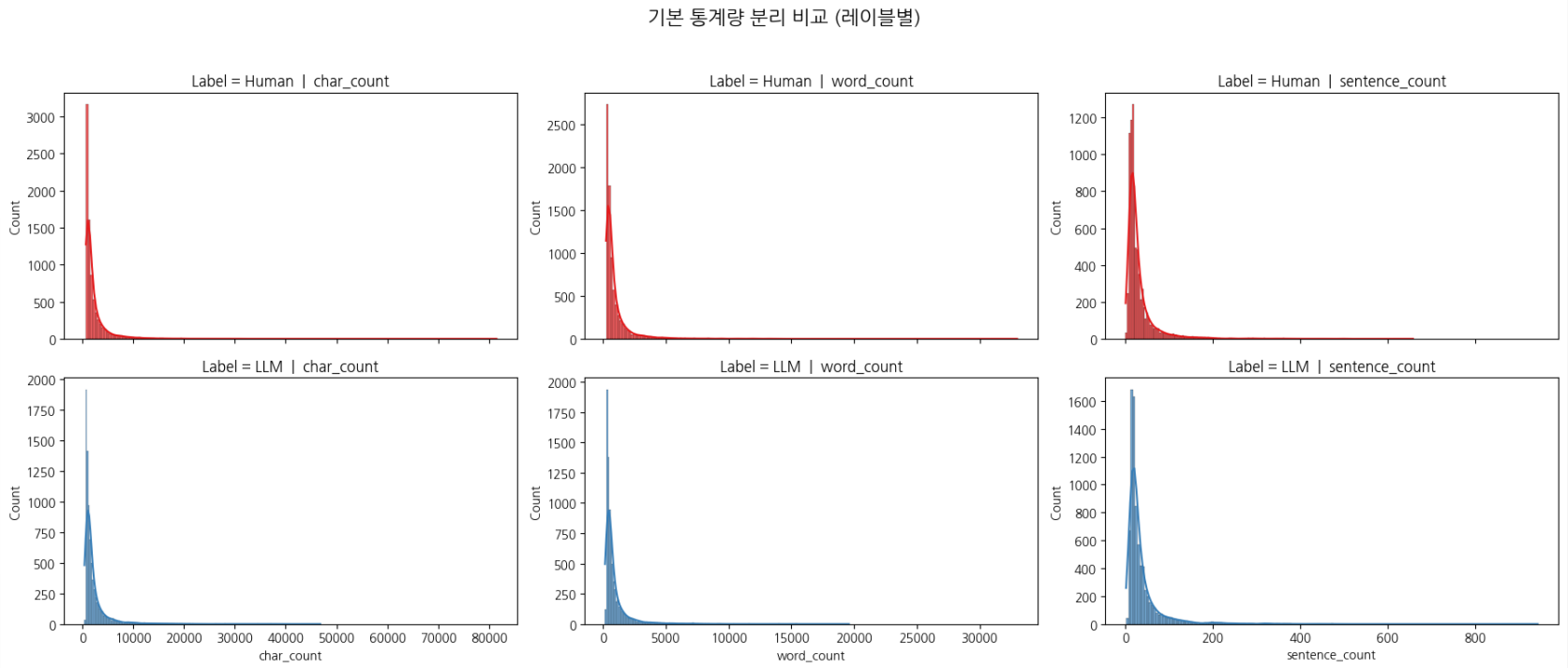
KatFishNet: Detecting LLM-Generated Korean Text through Linguistic Feature Analysis

- 쉼표 사용 패턴, 품사 분포, 단어 개수 등 단순한 언어 특성만을 이용하여 AI/human 분류

- 위 아이디어를 참고하여 주어진 train.csv 데이터에 대해 언어 특성 분석 시도

데이터 분석

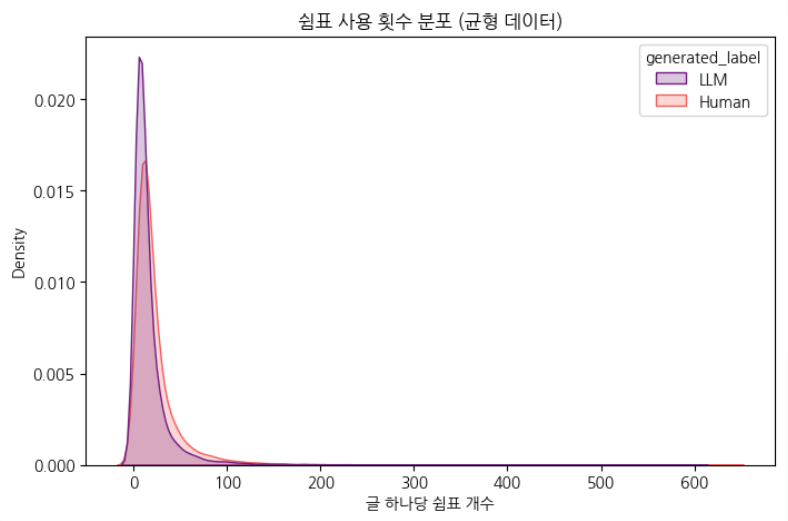
접근1) KatFishNet 아이디어 기반 언어 특성 분석 시도

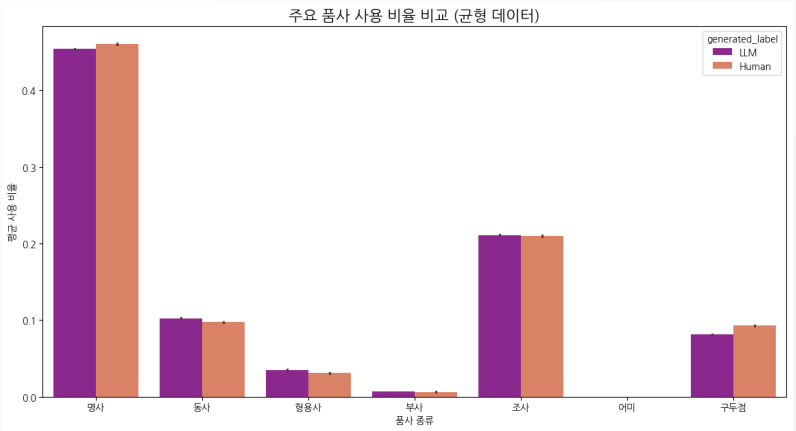


- Human vs LLM: 글자수, 단어수, 문장 수 비교

데이터 분석

접근1) KatFishNet 아이디어 기반 언어 특성 분석 시도

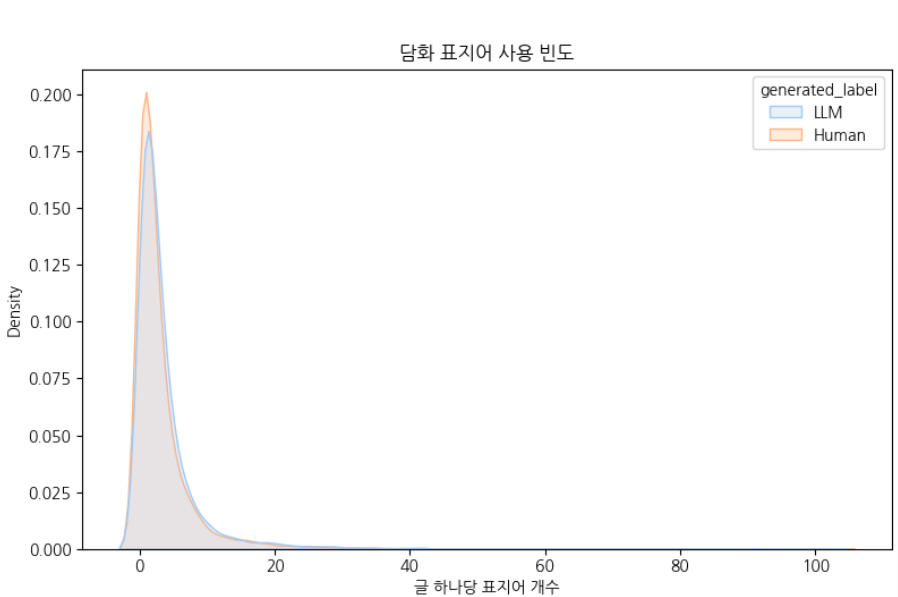




- Human vs LLM: 쉼표 사용 횟수 비교, 품사 사용 비율 비교

데이터 분석

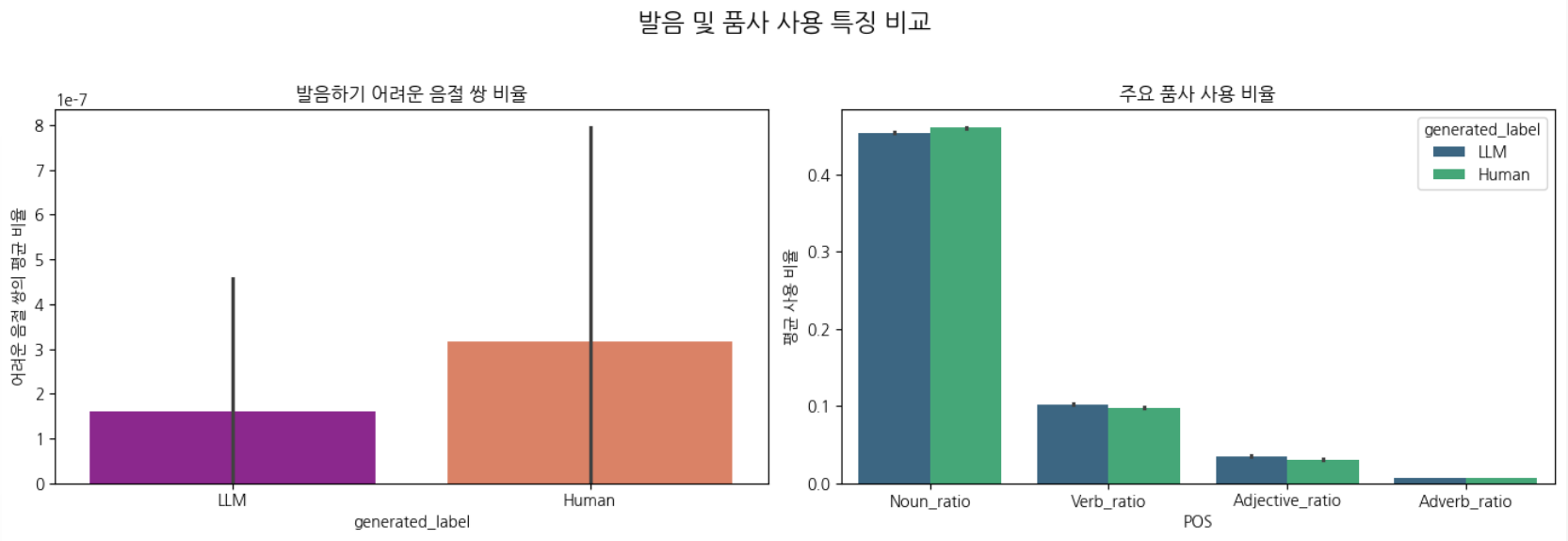
접근1) KatFishNet 아이디어 기반 언어 특성 분석 시도



- Human vs LLM: 담화 표지어 사용 빈도

데이터 분석

접근1) KatFishNet 아이디어 기반 언어 특성 분석 시도



- Human vs LLM: 발음 어려운 음절 쌍 비율 비교, 주요 품사 사용 비율 비교

데이터 분석

접근1) KatFishNet 아이디어 기반 언어 특성 분석 시도

- 데이터 분석 결과, Human vs LLM 언어적 특성 분포의 유의미한 차이를 얻기 어려움
- 현재 train.csv에서 generated 1인 데이터에는 AI가 일부 문장이나 문단만 작성된 데이터도 포함됨

- 한계 : 일반적인 분류 학습시 높은 정확도를 기대하기 어려움

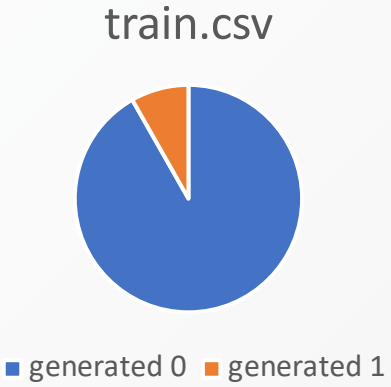
데이터 분석

접근2) 데이터 증강

[데이터의 문제점]

- 기존 train.csv의 generated 1 데이터는 온전히 AI가 생성한 것이 아님

- 또한 generated 0: generated 1 이 91.7:8.3 정도로 매우 불균형함.



- train 데이터는 한 샘플 당 여러 문단으로 구성되지만, 평가 데이터는 한 문단으로만 구성됨

데이터 분석

접근2) 데이터 증강

[generated 0을 이용한 데이터 증강]

- generated 0 샘플들만 추출하여 문단 단위로 나누어 새로운 샘플들로 구성(약 112만개의 샘플)

- 새로운 샘플 중 약 40%를 LLM 모델을 통하여 재생성

- LLM 모델로는 rtzr/ko-gemma-2-9b-it, SEOKDONG/llama3.1_korean_v1.1_sft_by_aidx 사용

데이터 분석

접근2) 데이터 증강

[generated 0을 이용한 데이터 증강]

- 클래스 비율을 약 2:3 (generated 0 : generated 1) 수준으로 맞춘 균형 잡힌 데이터셋 확보

  gemma_human.csv, llama_human.csv

- 총 샘플 수 97,172 → 1,125,652로 약 11.6배 증가

- 클래스 불균형을 해소하고 샘플 수를 대폭 확대한 만큼, 딥러닝 모델의 성능 향상을 기대할 수 있음

Augmentation

Train Overall Process

- **순서도**

문단 단위 분리 및 저장

데이터 증강 (LLaMA, Gemma)

본문 전체 학습 (슬라이딩 윈도우)

수도 라벨링 기반 학습

증강 데이터 학습

앙상블

총 6단계의 파이프라인

Gemma Augmentation

- 입력 데이터
  - 문단의 집합을 문단으로 분리한 본문 중 레이블이 0인 것 (title, paragraph 포함)
- LLM 설정
  - 모델 : rtzr/ko-gemma-2-9b-it
  - 프롬프트 5종 (수필, 설명글, 뉴스, 요약, 칼럼)
  - temperature=0.7 / top_p = 0.95
- 샘플링 전략
  - 전체 문단 중 2~3개 마다 1개 랜덤 선택 (스타일별 균등 분할)

LLaMA Augmentation (1/2)

- 입력 데이터
  - 앞서 자른 label이 0(사람이 작성한)인 문단 데이터 (title, paragraph 포함)

- LLM 설정
  - SEOKDONG/llama3.1-Korean_v1.1_sft_by_aidx
  - 모델 유형 : decoder-only / left-padding
  - temperature=0.7 / top_p=0.95
  - 프롬프트 5종 (수필, 설명글, 뉴스, 요약, 칼럼)

LLaMA Augmentation (2/2)

- 샘플링 전략
  - 총 NUM_SAMPLES개 랜덤 샘플링
  - 스타일별 균등 분할 → 5등분 후 각각 스타일 변환

Prompt (1/2)

프롬프트 예시

수필

[지시]
당신은 **감수성이 풍부한 작가**입니다. 주어진 글을 바탕으로 '**수필**' 형식으로 문체를 바꿔주세요.
원문의 **핵심 정보는 유지**해야 합니다. 오직 변환된 수필 본문만 생성하고, 그 어떤 부연 설명이나
코멘트도 절대 추가하지 마세요. 줄 바꿈 없이 생성해주세요.

설명글

[지시]
당신은 **뛰어난 설명가**입니다. 주어진 글의 내용을 독자들이 **알아듣기 쉽게 설명**해주세요.
오직 변환된 설명 글 본문만 생성하고, 그 어떤 부연 설명이나 코멘트도 절대 추가하지 마세요.
줄 바꿈 없이 생성해주세요.

뉴스

[지시]
당신은 **냉철한 기자**입니다. 주어진 글에서 **감정적인 표현은 모두 배제**하고,
객관적인 사실과 정보만을 사용하여 **육하원칙에 따라 간결하고 명료한 '뉴스 기사'** 형식으로 문체를 바꿔주세요.
오직 변환된 뉴스 기사 본문만 생성하고, 그 어떤 부연 설명이나 코멘트도 절대 추가하지 마세요.
줄 바꿈 없이 생성해주세요.

Prompt (2/2)

프롬프트 예시

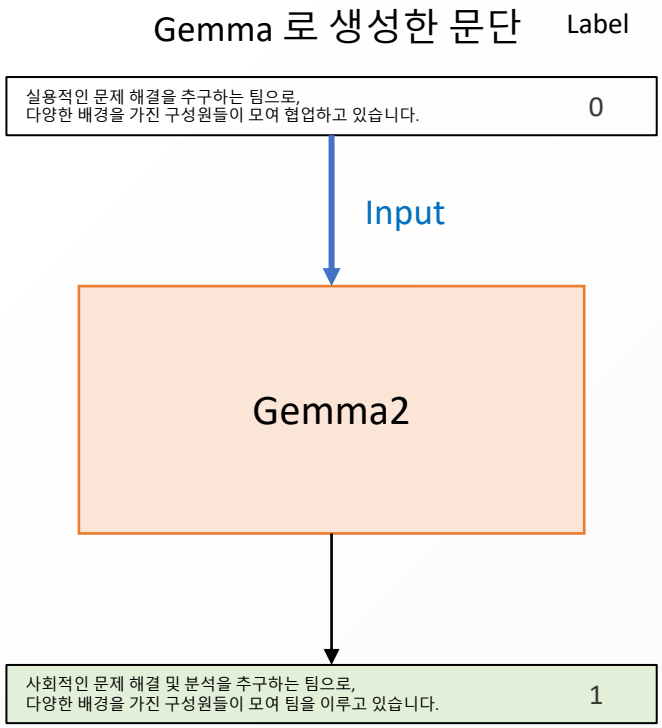
요약

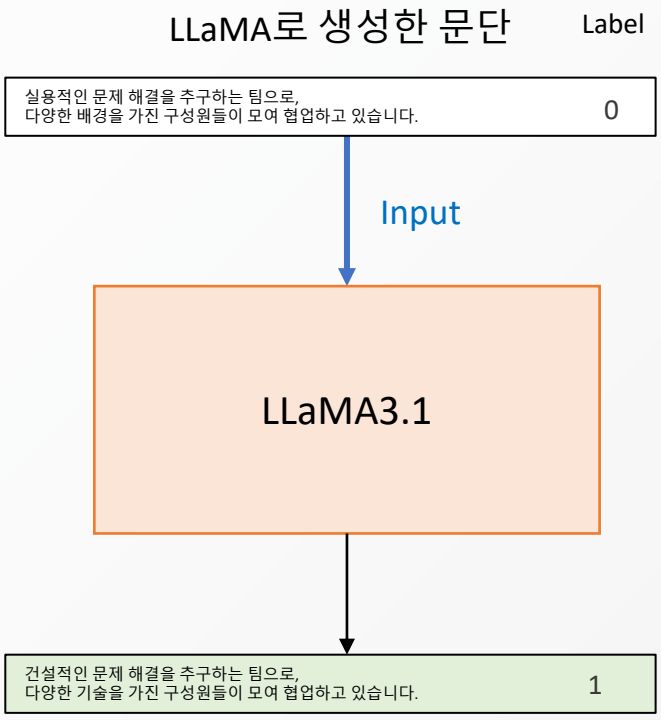
[지시]
당신은 뛰어난 **요약 전문가**입니다. 주어진 글의 핵심 내용을 간결하게 **요약** 해주세요.
단, **원래 글의 문체와 톤은 최대한 유지하면서 요약**해야 합니다. 스타일을 바꾸지 마세요.
오직 요약된 본문만 생성하고, 그 어떤 부연 설명이나 코멘트도 절대 추가하지 마세요.
줄 바꿈 없이 생성해주세요.

칼럼

[지시]
당신은 **논리적인 칼럼니스트**입니다. 주어진 글의 내용을 바탕으로, '**에세이**' 형식으로 문체를 바꿔주세요.
오직 변환된 에세이 본문만 생성하고, 그 어떤 부연 설명이나 코멘트도 절대 추가하지 마세요.
줄 바꿈 없이 생성해주세요.

Augumentation : Gemma, LLaMA





Train

Train Overall Process

- **순서도**

문단 단위 분리 및 저장

데이터 증강 (LLaMA, Gemma)

본문 전체 학습 (슬라이딩 윈도우)

수도 라벨링 기반 학습

증강 데이터 학습

앙상블

총 6단계의 파이프라인

Train-Full-Text(개요) (1/3)

- **사용데이터 및 전처리**
  - train.csv
    - ✓ **DeBERTa는 최대 토큰이 512**이기에 Full Text를 받지 못함
    - ✓ 이를 개선하기 위해 **Sliding Window**를 활용하여 Window로 나누어 **전체 Text의 내용을 학습**에 사용

- **사용 모델**
  - Backbone
    - ✓ team-lucid/**deberta-v3-base**-Korean
  - CLS 임베딩 기반 Downstream Classifier용 Custom head 구성

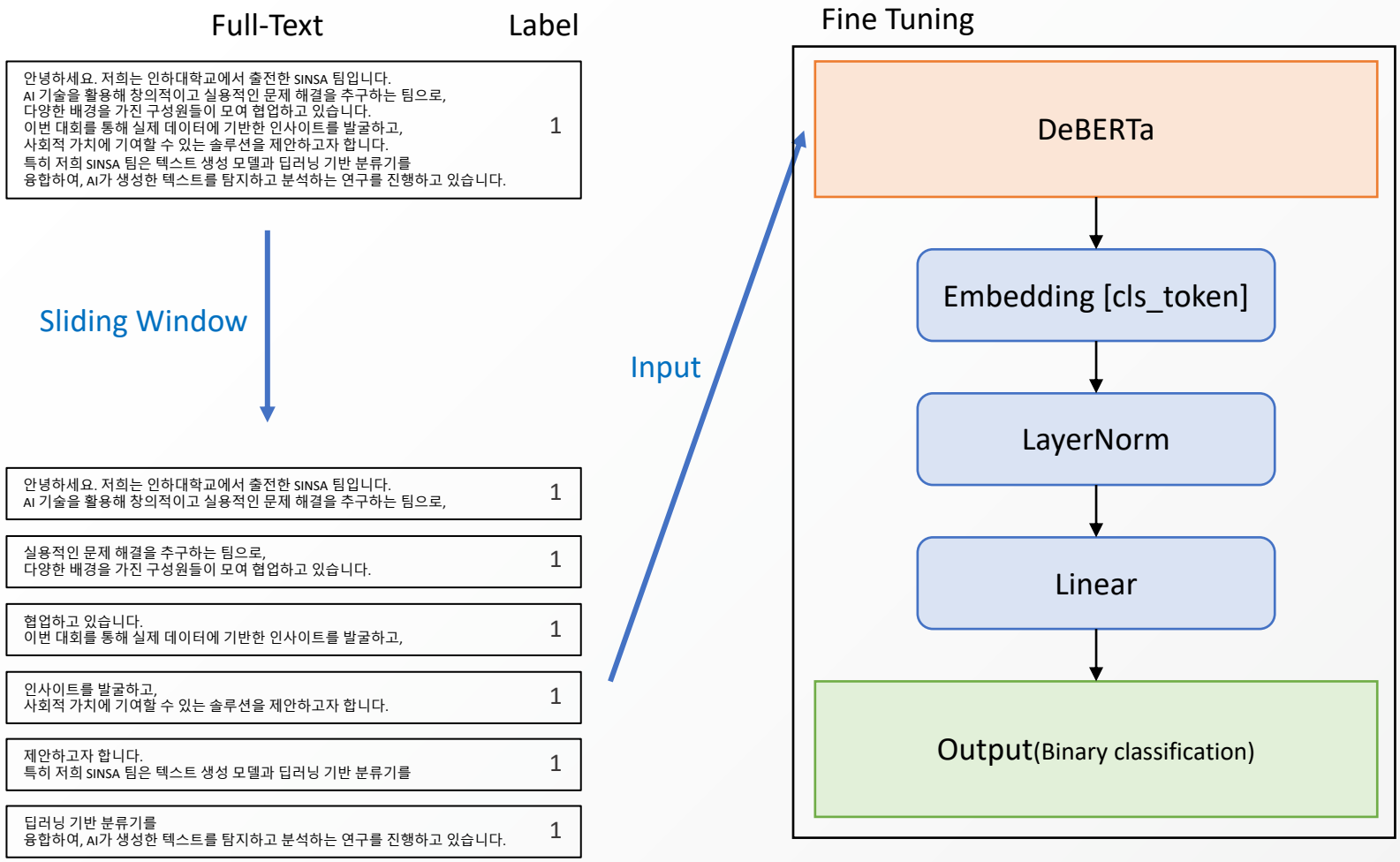
Train-Full-Text(개요) (2/3)

- **학습 전략**
  - Batch_size = 16 / Loss = BCEWithLogitLoss / Optimizer = Adam / lr = 2e-5 / Epochs = 3

    window_size = 400 / overlap = 100

  - Validation 없이 **모두 Train Set 사용**

Train-Full-Text(아키텍처) (3/3)



Train-pseudo-labeling-1(개요) (1/3)

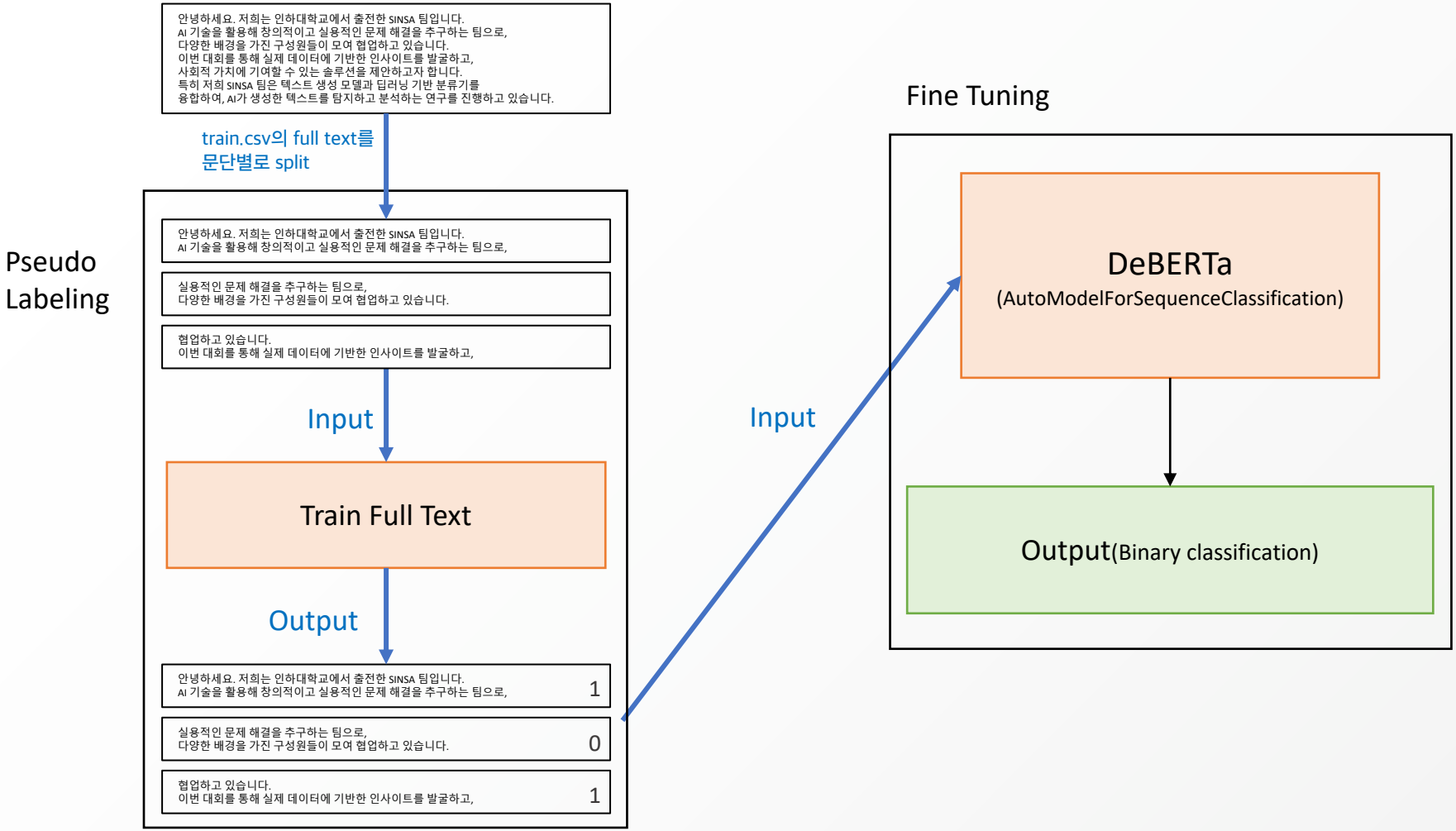
- **사용데이터 및 전처리**
  - train_pseudo_label.csv
    - ✓ Train-Full-Text(Sliding Window사용)에서 DeBERTa를 학습한 모델로 train.csv **문단단위 Pseudo Labeling**
    - ✓ Majority class가 Minority class의 6배가 되도록 **Undersampling**

- **사용 모델**
  - Backbone : team-lucid/**deberta-v3-base**-Korean
  - AutoModelForSequenceClassification으로 **Classifier 자동 구성**

Train-pseudo-labeling-1(개요) (2/3)

- **학습 전략**
  - Batch_size = 4 / lr_scheduler_type = “cosine” / lr = 1e-5 / Epochs = 3 / weight_decay = 1e-2
  - Validation Set 20% / Train Set 80%

Train-pseudo-labeling-1 (아키텍처)

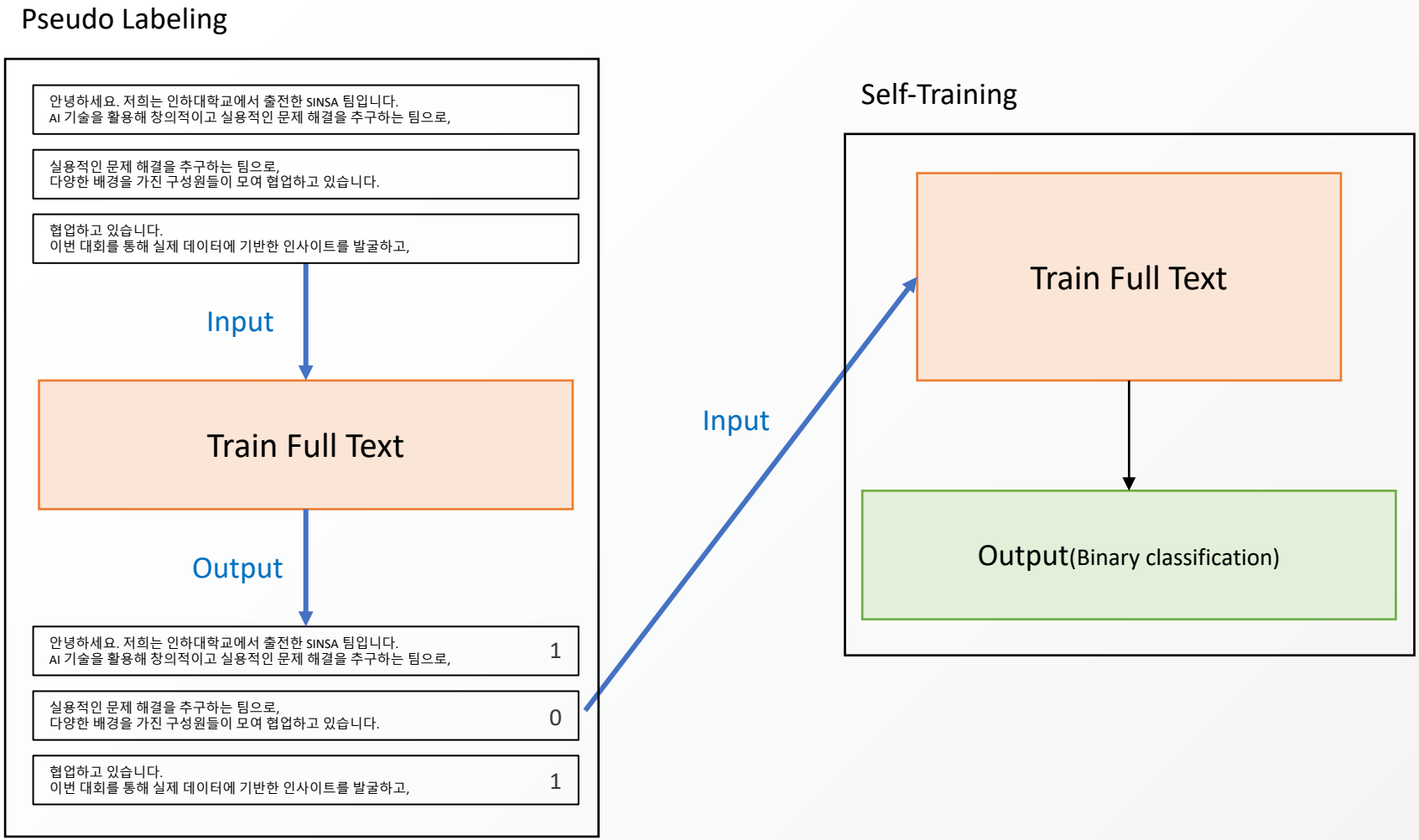


Train-pseudo-labeling-2(개요) (1/2)

- **사용데이터 및 전처리**
  - train_pseudo_label.csv
    - ✓ Train-Full-Text(Sliding Window사용)에서 DeBERTa를 학습한 모델로

      train.csv 문단단위 Pseudo Labeling
- **사용 모델**
  - 이전에 학습한 **Train-Full-Text 모델에 Pseudo Labeling한 데이터로 Self-Training**
- **학습 전략**
  - Batch_size = 4 / lr_scheduler_type = “cosine” / lr = 1e-5 / Epochs = 3 / weight_decay = 1e-2
  - **Validation Set 20% / Train Set 80%**

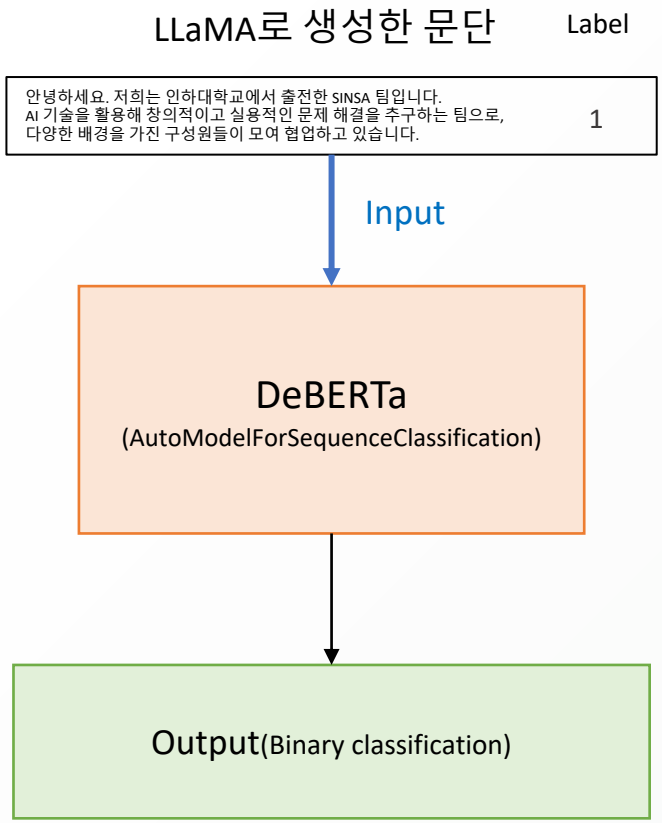
Train-pseudo-labeling-2(아키텍처) (2/2)

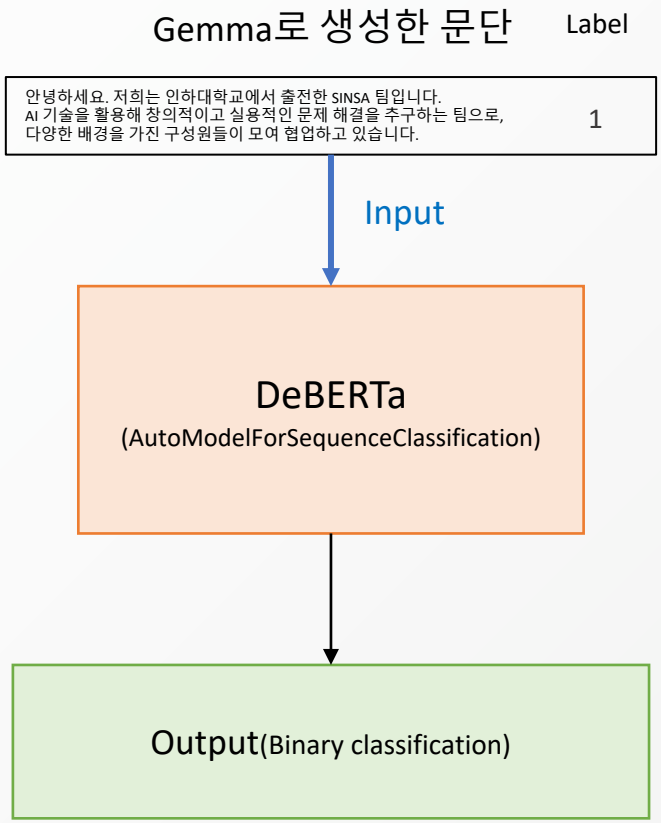


Train-Augmentation(개요) (1/2)

- **사용데이터 및 전처리**
  - llama, gemma augmentation data
    - ✓ 앞서 **증강한 LLaMA와 Gemma의 데이터를 사용**하고 위 데이터는 **문단 수준**으로 쪼개져 있음
- **사용 모델**
  - Backbone : team-lucid/**deberta-v3-base**-Korean
  - AutoModelForSequenceClassification으로 **classifier 자동 구성**
- **학습 전략**
  - Batch_size = 4 / lr_scheduler_type = “cosine” / lr = 1e-5 / Epochs = 3 / weight_decay = 1e-2
  - **Validation Set 20% / Test Set 80%**

Train-Augmentation(아키텍처) (2/2)





Ensemble

Ensemble (1/2)

- **앙상블 input**

Source	Score (AUC-ROC)
LLaMA 증강 기반 train augmentation	0.888
Gemma 증강 기반 train augmentation	0.901
train-pseudo-labeling-1	0.906
train-pseudo-labeling-2	0.909
train-full-text	0.910

Ensemble (2/2)

- **앙상블 전략**
  - 모델별 예측 확률 (generated) 통합
  - 각 **샘플별 confidence (|p-0.5|)가 높은 상위 2개** 예측값 추출
    - ✓ 두 값의 차이가 임계 값 이하 → 상위 2개 평균
    - ✓ 두 값의 차이가 임계 값 이상 → 전체 평균
  - 최종 결과 (generated)만 추출
- **특징**
  - 불확실성이 큰 샘플에선 더 안정적인 예측
  - Soft-voting의 일반적 평균보다 신뢰도 기반으로 더 정교하게 반영

Result

Result

- 리더보드 점수

PUBLIC	PRIVATE
ROC-AUC	ROC-AUC
0.93353	0.92415
RANK	RANK
5/271	7/271

확장 가능성

확장 가능성

- **끊임 없이 진화하는 생성형 모델 대응**
  - GPT / GEMINI 등 지속적으로 발달 되고 있음
    - ✓ 베이스라인 코드를 기반으로 현재 기술 상황에 맞춘 데이터 증강 적용
    - ✓ 최신 경향을 기반으로 지속 학습 통한 유지 보수 가능

- **사용자 친화적인 개발**
  - 실사용 친밀한 적용
    - ✓ 브라우저 확장 프로그램 (chrome extension) 통한 사용자 미디어 리터러시 보조 수단으로의 역량 극대화

확장 가능성

인간과 AI가 신뢰 속에서 공존하는

지속 가능한 미래를 만드는 주축이 됨

Thank You For Listening

INHA AI