PROJECT OVERVIEW

한국어 Hate Speech

멀티라벨 분류

모델 비교 & 로컬 파인튜닝 실전 프로젝트

성능 비교 뿐만 아니라, 모델의 성격을 분석하고
실제 서비스 적용 시의 정책 설계도 고민해보았습니다.

Multi-Backbone | Local Tuning | Deep Analysis

Task 정의 & 데이터셋 (KMHaS)

Multi-label Classification

- **입력:** 한국어 댓글 및 문장
- **출력:** 8개 혐오 유형에 대한 멀티라벨 예측
- **데이터셋:** KMHaS (Korean Multi-label Hate Speech)
- **평가 지표:** F1-micro, F1-macro, Accuracy

> 한 문장에 **여러 혐오 유형**이 동시에 존재할 수 있음을 고려하여 모델을 설계합니다.

Label	Meaning
origin	출신/이주민 비하
physical	외모/신체/장애 비하
politics	정치 성향/이념 혐오
profanity	일반 욕설/비속어
age	특정 세대/나이 비하
gender	성별(여성/남성) 혐오
race	인종/국가 혐오
religion	종교 혐오

실험 구조 개요: Two-Track Strategy

Track 1: Colab Multi-Backbone

- **Environment:** Colab Pro (T4/A100)
- **Target:** 3 Models Comparison (KcELECTRA, KoELECTRA, KR-Medium)
- **Goal:** 최적의 백본 모델 탐색 및 하이퍼파라미터 (v1~v3) 튜닝

Track 2: Local KcELECTRA

- **Environment:** Local GPU (Windows)
- **Target:** Deep Dive into KcELECTRA
- **Goal:** Epoch, LR, Threshold 변화에 따른 모델의 '성격' 변화 관찰 및 심층 분석

◎하나의 태스크를 **넓게(Comparison)**, 그리고 **깊게(Local Tuning)** 탐구

실험 환경: Colab Multi-Backbone

Colab Pro

Tesla T4 / A100 GPU
High-RAM Runtime

Hugging Face

Transformers
Datasets
Evaluate

Multi-label Setup

BCEWithLogitsLoss
Sigmoid Activation
Threshold 0.5

실험 환경: 로컬 KcELECTRA 파인튜닝

Local Workstation

Windows OS
NVIDIA CUDA
Support

Python Environment

Conda Virtual Env
PyTorch
Scikit-learn

Experiment Focus

LR Scheduler
Threshold
Batch
Epoch Variations

백본 모델 3종 & 튜닝 전략

Target Backbones

KcELECTRA-base

한국어 댓글 구어체 특화 (beomi)

KoELECTRA-base

다양한 한국어 코퍼스 학습 (monologg)

KR-Medium

BERT 계열 중형 모델 (snunlp)

Tuning Versions (Sweet Spot Search)

Version	Learning Rate	Epoch	Objective
v1	2e-5	3	Baseline Setting
v2	3e-5	3	LR Increase (Aggressive)
v3	2e-5	4	Epoch Increase (Check Overfitting)

Colab 정량 결과: F1 Score 비교

🏆 Best: **KcELECTRA v2**

LR을 살짝 조정한 v2에서 대체로 최고 성능.
Epoch를 늘린 v3는 개선폭이 미미하거나 과적합 경향.

정성 평가 1: 명확한 문장 (Consensus)

"누구나 동의하는 명확한 케이스에서는 3개 모델 모두 안정적"

입력 문장	KcELECTRA	KR-Medium	KoELECTRA
"오늘 날씨가 좋아서 산책하러 나갔다 왔어요."	Clean	Clean	Clean
"와 진짜 오늘 일 개빡셌다, 녹초 됐어." (단순 욕설)	Clean	Clean	Clean
"여자들은 감정적이라 맡기면 안 된다."	Gender	Gender	Gender
"이민자들은 다 쫓아내야 나라가 산다."	Origin	Origin	Origin

정성 평가 2: 애매한 문장 (Divergence)

Case A: 인용/충격

"댓글에 '이민자들은 다 쫓아내야 한다'라는 글이 올라와서 충격 받았다."

- **KcELECTRA:** Origin=1 (공격적)
- **KoELECTRA:** Clean (보수적)
- **KR-Medium:** Clean (보수적)

Case B: 은근한 성 역할

"여자는 남자보다 감성적 성향이 강한 편이다."

- **KcELECTRA:** Clean (미탐 가능성)
- **KoELECTRA:** Clean (미탐 가능성)
- **KR-Medium:** Gender=1 (민감함)

모델별로 **공격형(Kc)**, **보수형(Ko)**, **젠더 민감형(KR)** 성격이 드러남.

로컬 KcELECTRA 실험: "동일 모델, 다른 성격"

Goal

Epoch, LR, Threshold 변화가
실제 예측 결과(성격)에 미치는 영향
관찰

v1

라벨별 Threshold 개별 적용 &
안정화 실험
(lr=2e-5, ep=3, batch=32)

v2 ~ v3

Epoch 증가 & lr 변경
(lr=3e-5, ep=4, batch=32)

v4 ~ v5

Batch size 증가 & LR Scheduler
(lr=3e-5, ep=4, batch=64)

KcELECTRA v1~v5: 학습 조건 vs 성능

Ver	Description	F1 Micro	F1 Macro
Baseline	Base (lr=2e-5, ep=3, batch=32)	0.979	0.915
v1	Threshold Adj.	0.980	0.920
v2	Epoch Increase(ep=4)	0.978	0.918
v3	Learning Rate(lr=3e-5)	0.977	0.918
v4	Batch Size(batch=64)	0.979	0.921
v5	LR Scheduler	0.979	0.979

💡정량 지표는 비슷하지만, **Threshold와 세팅**에 따라 모델의 성격이 달라짐.

정성 평가 3: 명확한 문장 (Consistency)

"버전을 바꿔도 명확한 케이스에 대한 판단은 흔들리지 않음"

Clean Case

"오늘 날씨가 좋아서 산책하러 나갔다 왔어요."

v1: Clean v3: Clean v5: Clean

Hate Case

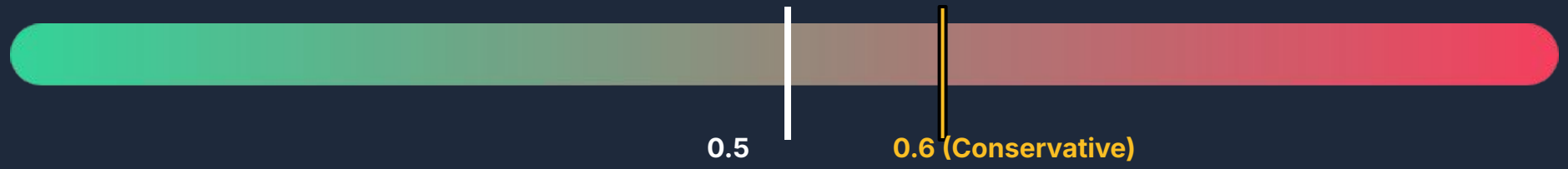
"이민자들은 다 쫓아내야 나라가 산다."

v1: Origin v3: Origin v5: Origin

하이퍼파라미터 튜닝은 주로 "경계선(Borderline)" 케이스에 영향을 미침.

Threshold & 정책의 중요성

"같은 모델도 Threshold에 따라 완전히 다른 필터가 됩니다"



안전 중시 (Safety)

혐오를 놓치면 안 됨

오탐↑ 미탐↓ 허용

Low Threshold

자유 중시 (Freedom)

과한 차단 방지

오탐↓ 미탐↑ 허용

High Threshold

오탐(FP) vs 미탐(FN): 모델보다 중요한 서비스 정책

핵심 질문: 어떤 실수가 더 위험한가?

- 오탐 **(False Positive)**
  혐오가 아닌데 혐오라고 차단
  (예: 인용, 비판, 보고 문장)

- 미탐 **(False Negative)**
  혐오인데 혐오가 아니라고 통과
  (예: 은근한 차별, 교묘한 패턴)

Insight: 모델 성능이 비슷해도,
어떤 종류의 실수를 자주 하느냐는 서로 다르다.

정책 시나리오: 안전 우선형 vs 표현의 자유 우선형

"같은 KMHaS 데이터, 비슷한 F1 점수라도 **어떤 정책**을 택하느냐에 따라 백본이 달라진다"

Scenario 1: 안전 우선형
커뮤니티 보호 중심 (미탐 최소화)

- 혐오 발화를 절대 놓치지 않는 것이 목표
- Threshold를 **낮게 설정** (예: 0.4)
- 소수자(Race/Gender) 라벨 민감도 ↑
- 인용/비판 문장도 일단 잡고, 사람이 복구

Best Fit Model
KcELECTRA (공격형/민감형)

Scenario 2: 표현의 자유
토론·연구 공간 (오탐 최소화)

- 인용·비판·보고는 최대한 막지 않음
- Threshold를 **높게 설정** (예: 0.6~0.7)
- 명백한 욕설만 차단, 나머지는 신고 기반
- "충격 받았다 / 동의하지 않는다" 허용

Best Fit Model
KoELECTRA / KR-Medium (보수형/관찰자형)

✅ 모델 튜닝만큼이나 **정책 설계(서비스 성격 정의)**가 중요하다는 점을 확인.

통합 인사이트: 모델별 '성격' 프로파일

KcELECTRA

공격형 필터

정량 성능 최고.
인용문까지 민감하게 잡음.

KoELECTRA

보수형 관찰자

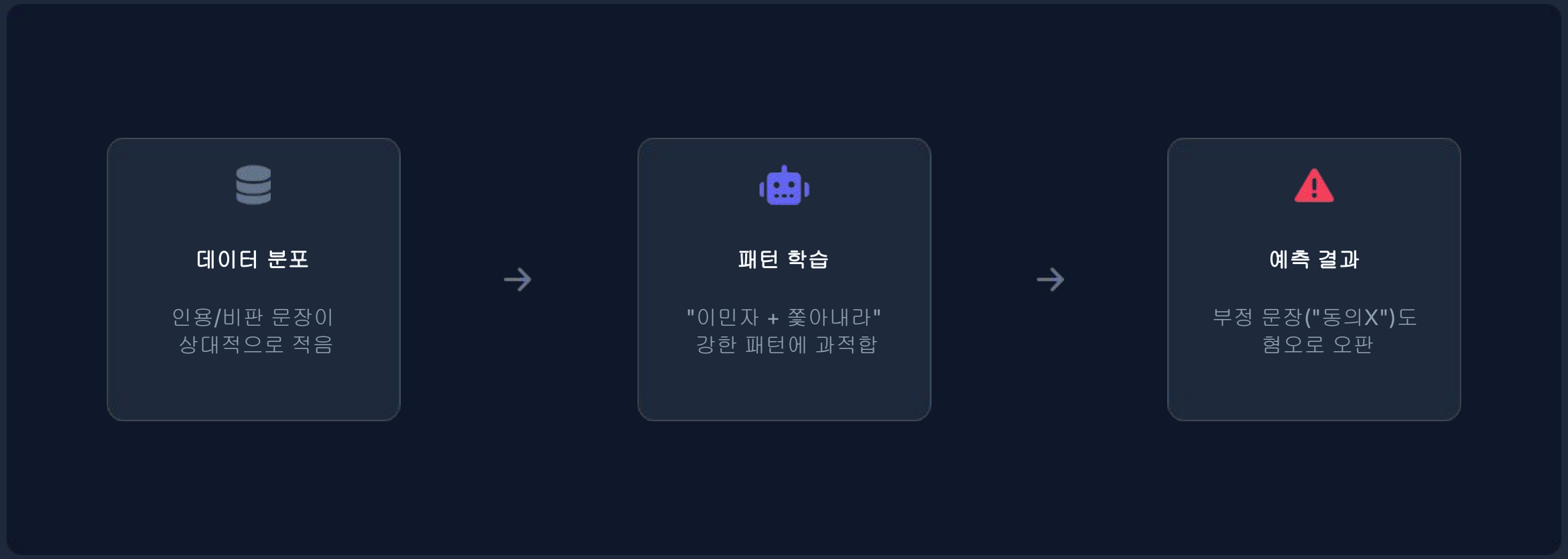
인용/보고에 관대함.
은근한 혐오 놓칠 수 있음.

KR-Medium

젠더 민감형

성 역할 고정관념에
특히 예민하게 반응.

통합 인사이트: 데이터와 라벨의 한계



모델은 **윤리적 진실**이 아니라, **데이터의 편향**을 학습합니다.

에러 분석: 왜 모순적인 결과가 나올까?

Case 1: 부정 표현 실패

"나는 이민자들은 다 쫓아내야 한다는 말에 동의하지 않는다."

Origin=1 (오탐)

모델이 '동의하지 않는다'는 부정어보다
'쫓아내야 한다'는 혐오 패턴에 압도됨.

Case 2: 애매한 보고

"여자는 감성적 성향이 남자보다 강한 편이라는 보고가 있다."

Gender=1 vs 0 (모델별 상이)

데이터 내에서 '여자는 ~' 패턴이 주로 혐오로
라벨링 되었을 가능성이 큼.

개선 방향: Data, Rule & Model

Data Augmentation

- 인용/부정/비판 문장 데이터 대폭 추가
- "동의하지 않는다" 등 Non-hate 케이스 명시적 학습

Rule-based Filter

- "~라는 글이", "충격 받았다" 패턴 감지
- 해당 패턴 시 혐오 점수 패널티 부여

Hybrid Model

- 공격형(1차) + 보수형(2차) 모델 조합

결론

"같은 태스크라도, **백본·튜닝·정책**에

따라

전혀 다른 성격의 필터가 만들어진다."