JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 1, pp. 71-80, Jan. 2026

Check for updates

비라벨 기반 Self-Training을 이용한 유해 댓글 탐지 초기 대응의 안정성 중심 개선

김 현 아 · 조 윤 용*
경기대학교 교양학부 조교수

Stabilizing Early-Stage Toxic-Comment Detection via Conservative Unlabeled-Data Self-Training

Hyun-Ah Kim · Yoon-Yong Cho*
Assistant Professor, Department of General Studies, Kyonggi University, Suwon 16227, Korea

[요 약]

본 연구는 동일 도메인 내 비라벨(unlabeled) 데이터를 활용한 자가학습(self-training)이 라벨이 제한된 초기 단계의 유해 댓글 탐지 성능 저하 없이 초기 대응 품질을 소폭 개선할 수 있는지를 검증하였다. 실험에는 한국어 혐오 발화 코퍼스(KHS, Korean Hate Speech Corpus)를 사용하였으며, 단 한 번의 자가학습 과정이 모델의 예측 성능에 미치는 영향을 평가하였다. 신뢰도 임계값을 $p \geq 0.95$로 설정하여 의사라벨(pseudo-label)을 주입한 결과, 거짓 양성률이 감소하고 Brier score로 측정한 예측 확률의 보정 신뢰도가 유지되었다. 반면, 임계값을 $p \geq 0.90$으로 완화했을 때는 Macro-F1과 PR-AUC 등 균형 지표는 큰 변화 없이 유지되면서, 오탐률과 Brier score가 소폭 개선되었다. 이는 단일 라운드의 자가학습만으로도 초기 탐지 단계에서 모델의 안정성을 확보할 수 있음을 시사한다. 본 연구 결과는 라벨 부족 환경에서 준지도 학습 기반 자가학습 접근의 실용적 가치를 입증하며, 대규모 수작업 라벨링 없이도 신뢰도 높은 경량 한국어 유해발화 탐지 모델 구축의 가능성을 제시한다.

[Abstract]

This study investigates whether self-training with in-domain unlabeled data can safely improve early-stage toxic-comment detection under limited labeled-data conditions. The effectiveness of a single self-training iteration in improving the model performance was experimentally evaluated using the Korean Hate Speech corpus. The incorporation of high-confidence pseudo-labels ($p \geq 0.95$) reduced false positives and maintained calibration reliability, as measured by the Brier score. When the threshold was relaxed to $p \geq 0.90$, balanced evaluation metrics such as the macro-F1 and PR-AUC exhibited small but consistent gains while preserving the original performance level in early-stage detection. These findings confirm that even one round of self-training contributes to performance consistency and robustness in low-resource scenarios. The results highlight the practical utility of self-training-based semi-supervised learning for Korean toxic-speech detection, thereby offering a lightweight and deployable model that yields safe minor gains while maintaining reliable predictions without extensive manual annotation efforts.

색인어 : 비라벨 데이터, 자가학습, 유해 댓글 탐지, 초기 대응 성능, 준지도 학습
Keyword : Unlabeled Data, Self-Training, Toxic-Comment Detection, Early-Stage Performance, Semi-Supervised Learning

http://dx.doi.org/10.9728/dcs.2026.27.1.71

Received 27 October 2025; **Revised** 24 November 2025
Accepted 10 December 2025

***Corresponding Author; Yoon-Yong Cho**
Tel: +82-31-249-1472
E-mail: yoonycho@kyonggi.ac.kr

Ⅰ. 서 론

1-1 연구 배경 및 필요성

온라인 커뮤니티, 포털 뉴스, SNS 등 대화 중심 플랫폼의 확산은 사회적 의사소통의 양상을 근본적으로 변화시켰다. 이러한 공간은 정보 공유와 여론 형성의 주요 채널로 자리 잡았지만, 동시에 혐오 표현과 공격적 발언의 확산을 촉진하는 부정적 결과를 낳고 있다. 온라인 댓글은 익명성과 즉시성을 기반으로 한 표현의 자유를 확대했으나, 그 이면에서는 인신공격·비하·차별과 같은 유해 언어가 급속히 퍼지며 사회적 갈등을 심화시키고 있다. 이러한 현상은 개인의 심리적 피해를 넘어 사회 집단 간 분열을 가속화하고, 플랫폼의 신뢰도 하락과 온라인 담론의 질적 저하로 이어지고 있다. 특히 혐오 표현은 특정 집단에 대한 부정적 인식을 강화하며, 반복 노출될 경우 사회 구성원의 인지적 편향을 고착시켜 오프라인에서의 차별 행위로 이어질 가능성도 높다.

이에 따라 정부 기관과 주요 포털 기업은 자동화된 유해 댓글 탐지 시스템을 도입하여 선제적 대응 체계를 구축하고 있다. 그러나 실제 서비스 초기 단계에서는 충분한 라벨 데이터가 확보되지 않아 모델의 학습 안정성이 떨어지고 오탐과 미탐이 동시에 발생하는 한계가 존재한다[1]. 지도학습 기반의 기존 모델들은 대규모 주석 데이터를 전제로 하므로, 초기 대응 단계의 데이터 부족 환경을 제대로 반영하지 못한다. 더불어 온라인 언어는 시사 이슈와 유행어의 변화 속도가 빠르기 때문에, 한 시점에 수집된 라벨 데이터가 곧바로 노후화되어 모델 성능이 유지되지 못하는 문제가 빈번히 발생한다[2]. 이러한 현실은 라벨링 비용과 시간의 부담을 가중시키며, 초기 대응 지연으로 이어질 수 있다.

이와 같은 제약 속에서 비라벨 데이터를 효과적으로 활용하는 방법이 요구된다. 온라인 댓글은 매일 수백만 건 이상 생성되지만, 실제 라벨링이 가능한 비율은 매우 낮다. 따라서 라벨 부족 문제를 극복하기 위한 준지도 학습 접근이 현실적 대안으로 주목받고 있다. 준지도 학습은 제한된 라벨 데이터와 대규모 비라벨 데이터를 결합하여 학습 효율성과 일반화 성능을 동시에 향상시키는 방법[3]으로, 다양한 응용 분야에서 활발히 연구되고 있다.

특히 자가학습(Self-training)은 준지도 학습 중에서도 구현이 단순하고 실무 적용이 용이한 대표적 기법이다. 자가학습은 모델이 스스로 예측한 비라벨 데이터 중 신뢰도가 높은 샘플에 의사라벨을 부여하고, 이를 새로운 학습 데이터로 활용하여 성능을 보완한다[1]. Amini 등[1]은 자가학습의 구조적 특성과 일반화 효과를 분석하며, 고신뢰 예측을 반복적으로 학습에 활용할 경우 모델이 스스로 분포 정보를 강화함을 이론적으로 제시하였다. 그러나 반복적 자가학습은 의사라벨의 오류가 누적되어 학습 노이즈를 증가시킬 위험이 있다. 또한 대형 모델을 사용할 경우 학습 자원과 시간이 과도하게 소요되는 문제가 발생한다. 따라서 라벨이 부족한 초기 대응 환경에서는 효율적이면서도 신뢰도 높은 경량 자가학습 구조의 필요성이 커지고 있다.

본 연구는 이러한 맥락에서 단 한 번(one-round)의 자가학습만으로도 유의미한 성능 향상이 가능한지를 실험적으로 검증한다. 반복 라운드를 최소화하면서도 탐지 성능과 예측 확률의 안정성을 유지할 수 있다면, 이는 실제 온라인 플랫폼의 초기 대응 단계에서 실시간으로 적용 가능한 경량 모델 구조로서 의미가 있다. 나아가 동일 도메인 내 비라벨 데이터를 활용함으로써 데이터 편향을 최소화하고, 초기 모델의 일반화 성능을 향상시키는 전략적 접근을 제시한다.

결과적으로 본 연구는 라벨 부족 문제에 직면한 온라인 플랫폼의 유해 댓글 대응 체계에 있어, 단일 자가학습 기반의 준지도 학습이 실질적인 대안이 될 수 있음을 실험적으로 검증함으로써 학문적·실무적 기여를 도모한다.

1-2 연구 목적 및 기여

본 연구의 목적은 한국어 유해 댓글 탐지에서 자가학습 기반 준지도 학습의 효용성을 정량적으로 검증하는 데 있다. 기존의 지도학습 모델은 충분한 라벨 데이터를 전제로 설계되지만[3], 실제 서비스 초기 단계에서는 주석 데이터의 확보가 제한적이며 데이터의 품질 또한 일정하지 않다. 이러한 라벨 부족 환경에서는 모델이 불안정한 예측 확률을 출력하거나 오탐률이 높아지는 경향을 보이며, 이는 초기 대응의 신뢰도와 효율성을 저하시키는 주요 요인으로 작용한다. 따라서 본 연구는 동일 도메인 내 비라벨 데이터를 적극적으로 활용하여, 학습 효율성과 예측 안정성을 동시에 개선할 수 있는지를 실험적으로 검증하고자 한다.

연구의 구체적 방향은 다음과 같다. 먼저, 반복적 자가학습이 아닌 단 한 번의 자가학습만으로도 기존 지도학습 모델 대비 성능 저하 없이 오탐률과 캘리브레이션을 개선하는 안정성 구현이 가능한지를 확인한다. 반복적 자가학습은 학습 노이즈가 누적되거나 pseudo-label 품질이 저하될 가능성이 있으므로, 최소 라운드 수행으로도 안정적인 성능을 확보할 수 있다면 이는 실제 시스템 운영 단계에서 실질적인 의미를 가진다. 또한 본 연구는 고신뢰 임계값($p \geq 0.95$) 조건에서 생성된 의사라벨을 보수적으로 채택하여 오탐률을 줄이고 예측 확률의 일관성을 평가한다. 임계값 조정은 준지도 학습에서 데이터 품질을 관리하는 핵심 요소로, 신뢰도 높은 의사라벨만을 선택적으로 활용하는 전략은 초기 대응 단계의 탐지 신뢰도를 향상시킬 수 있다.

아울러 완화된 임계값($p \geq 0.90$)을 적용하는 실험을 병행하여, 정밀도와 재현율의 균형을 반영하는 성능 지표인 Macro-F1과 PR-AUC의 변화를 분석하고 임계값 운용의 전략적 기준을 제시한다. 이를 통해 고신뢰 기준이 가져오는 보수적 운용의 장점과 완화된 임계값이 제공하는 균형적 탐지 성능 간의 상호 관계를 정량적으로 해석한다. 이러한 접근은 실제 온라인 서비스 환경에서 탐지 시스템의 초기 임계값

을 설정하거나 조정할 때 실질적인 가이드라인으로 활용될 수 있다.

또한 본 연구는 복잡한 딥러닝 구조 대신 경량 분류기인 로지스틱 회귀(logistic regression)를 활용하여, 제한된 연산 자원 환경에서도 자가학습의 효과가 유지되는지를 검증한다. 이는 GPU와 같은 고가의 연산 자원을 보유하지 않은 기관이나 중소 규모 플랫폼에서도 적용 가능한 경량형 모델 구조를 제시한다는 점에서 실무적 의의가 크다. 더 나아가 동일 도메인 내의 비라벨 데이터를 학습에 통합함으로써 데이터 분포의 불일치를 최소화하고, 온라인 댓글 환경의 빠른 언어 변화에 적응할 수 있는 구조적 기반을 마련하였다.

본 연구의 기여는 크게 세 가지 측면으로 요약된다. 첫째, 라벨 부족 환경에서도 단일 자가학습만으로 탐지 성능과 예측 안정성을 동시에 개선할 수 있음을 실험적으로 입증하여, 기존 반복형 자가학습 대비 학습 효율성과 운용 편의성을 높였다. 둘째, 고신뢰 임계값 운용을 통해 오탐률 감소와 함께 PR-AUC 및 Brier score를 개선함으로써 임계값 설정의 정량적 근거를 마련하였다. 셋째, 완화된 임계값 실험을 통해 정밀도와 재현율 간의 균형 확보 가능성을 제시하였으며, 한국어 혐오 발화 코퍼스(KHS)을 활용한 실험을 통해 동일 도메인 비라벨 데이터의 실질적 학습 가치를 검증하였다.

결론적으로 본 연구는 대규모 라벨링 없이도 초기 대응 시스템의 품질을 향상시킬 수 있는 경량형 자가학습 모델을 제안하며, 라벨 부족이라는 현실적 제약 속에서도 안정적이고 신뢰도 높은 자동화된 유해 댓글 탐지 전략 수립에 실질적 시사점을 제공한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 유해 표현 탐지와 자가학습 기반 준지도 학습에 관한 선행연구를 정리하고, 그 가운데에서 본 연구의 위치와 차별성을 정리한다. Ⅲ장에서는 KHS 코퍼스를 중심으로 한 데이터셋과 연구 범위를 설명하고, 경량 로지스틱 회귀 기반 분류기, 의사라벨링 절차, 임계값 설정 및 평가 지표를 포함한 자가학습 파이프라인을 제시한다. Ⅳ장에서는 고신뢰($p \geq 0.95$)와 완화($p \geq 0.90$) 임계값 조건에서 self-training 1회가 Macro-F1, PR-AUC, Brier score, ΔFPR@Recall* 등에 미치는 영향을 정량적으로 비교하고, 채택률·오류 구조 변화와 함께 초기 대응 시나리오에서의 운용 해석을 제시한다. 마지막으로 Ⅴ장에서는 주요 결과를 요약하고, 라벨 부족 환경에서의 실무적 시사점과 함께 self-training 라운드 수, 도메인 확장, 공정성 분석 등 향후 연구 과제를 논의한다.

Ⅱ. 관련연구

2-1 유해 표현 탐지의 기술적 동향

유해 댓글 탐지는 사회적 갈등 완화와 온라인 담론의 질적 향상을 위한 핵심 인공지능 응용 분야로 주목받고 있다. Jahan 등[2]은 혐오 표현 자동 탐지 연구를 종합적으로 검토하며, 딥러닝 기반 모델이 규칙 기반 접근보다 문맥 이해와 표현 다양성에 강점을 지닌다고 분석하였다. Subramanian 등[3]은 감성 분석과 혐오 표현 탐지의 융합 가능성을 강조하고, Transformer 기반 언어모델의 전이학습 성능을 실증하였다. Yoo 등[4]은 BERT 기반 앙상블 모델을 통해 다국어 데이터셋에서 탐지 정확도를 향상시켰으며, Park 등[5]은 한국어 혐오 표현 탐지를 위한 K-HATERS 코퍼스를 구축하여 대상별 공격성 수준을 세분화하였다. Ludwig 등[6]은 도메인 적응을 적용하여 데이터 분포 차이를 보정하고, 모델의 일반화 성능을 향상시키는 방법을 제시하였다.

최근 연구들은 모델의 성능뿐 아니라 데이터 편향과 사회적 공정성 문제를 동시에 고려하는 방향으로 확장되고 있다. Delobelle 등[7]은 네덜란드어 RoBERTa 기반 언어모델을 제안하여 다국어 모델의 언어별 특이성을 보완하였으며, Murtadha 등[8]은 Rank-Aware Negative Training 기법을 통해 노이즈 라벨 환경에서 소수 클래스 탐지 성능을 향상시켰다. 이러한 연구는 다언어·다도메인 환경에서의 일반화와 공정성 확보로 발전하고 있으나, 한국어 환경에서는 여전히 실험적 검증이 부족하다. 따라서 본 연구는 한국어 특화 데이터셋과 현실적 초기 대응 조건을 반영한 준지도 학습 접근의 실효성을 검증하고자 한다.

2-2 자가학습 기반 준지도 학습

자가학습은 모델이 비라벨 데이터의 예측 결과를 의사라벨로 생성하고 이를 다시 학습에 활용함으로써 데이터 효율성을 높이는 준지도 학습의 대표적 접근이다[1]. Amini 등[1]은 자가학습의 구조적 특성을 체계적으로 분석하며, 고신뢰 예측을 반복적으로 활용할 때 일반화 성능이 강화된다고 설명하였다. Mukherjee와 Awadallah[9]는 불확실성 기반 자가학습 기법을 제안하여 예측 확률의 신뢰도에 따라 샘플을 선별함으로써 학습 안정성을 높였다. Sosea 등[10]은 학습 동역학을 활용해 의사라벨의 품질을 평가하고 신뢰도 기반 필터링을 적용하는 방법을 제시하였다. Chen 등[11]은 SAT 기법을 제안하여 샘플 난이도에 따른 가중치 조정으로 학습 효율을 향상시켰으며, Xu 등[12]은 이웃 정규화 방식을 통해 의사라벨 간 노이즈 전파를 감소시켰다.

이러한 연구들은 자가학습이 라벨 부족 문제를 완화하는 강력한 방법임을 입증했지만, 대부분의 모델은 대규모 연산 자원을 요구하는 복잡한 구조를 기반으로 한다. 특히 반복적 라운드 학습은 의사라벨의 품질이 일정 수준 이하로 떨어질 경우 오류를 누적시키며 성능 저하를 초래할 수 있다. 또한 자가학습의 핵심 단계인 임계값 설정은 대부분 경험적 기준에 의존하기 때문에, 예측 확률의 불확실성을 정량적으로 평가하는 연구는 상대적으로 부족하다.

최근에는 의사라벨의 신뢰도를 측정하기 위한 다양한 정량 지표와 불확실성 추정 방법이 시도되고 있다. 예를 들어, Brier score와 Expected Calibration Error(ECE)는 예측

확률이 실제 분포와 얼마나 일치하는지를 평가함으로써 모델의 캘리브레이션(calibration) 품질을 분석하는데 활용된다[13]. 본 연구는 이러한 접근을 바탕으로, 예측 확률의 신뢰도와 탐지 성능의 상관관계를 정량적으로 분석한다. 특히 단일 자가학습 구조에 고신뢰 임계값(p≥0.95)과 완화 임계값(p≥0.90)을 병행 적용하여, 의사라벨 품질과 임계값 운용 전략의 영향을 실험적으로 규명한다[14]. 이를 통해 초기 대응 환경에서도 안정적이고 반복 가능한 준지도 학습 프로세스를 구축할 수 있는 가능성[15]을 제시한다.

Ⅲ. 연구 방법

3-1 데이터셋과 연구 범위

본 연구는 라벨이 부족한 초기 대응 상황에서, 동일 도메인 내 대규모 비라벨을 활용한 자가학습 1회(1-round) 만으로 우선순위 정렬과 임계값 운용의 안정성을 실질적으로 개선할 수 있는지를 검정한다. 관측 단위는 댓글이며, 기본 코퍼스는 KHS 다.

KHS는 국내 포털 연예 뉴스 댓글에서 수집된 한국어 댓글 코퍼스로, 공개 라벨드 세트는 총 9,381개(train 7,896/dev 471/test 974)로 제공된다. 원 저자들은 '편견(bias)'과 '혐오(hate)' 등 다중 축을 주석했지만, 본 연구에서는 실제 운영 시나리오에 맞춰 이진 유해성(toxic vs. 정상)으로 단순화해 사용한다(유해='hate' 또는 'offensive', 정상=그 외). 동일 저장소에서 제공되는 비라벨(unlabeled) 풀은 2,033,893개로, 라벨 데이터와 동일 도메인(연예 뉴스 댓글)·동일 수집 체계를 공유한다. 이처럼 도메인 일치를 확보한 대규모 비라벨은 자가학습의 외삽 위험을 줄이고, 초기 대응 지연을 완화할 실무적 잠재력이 크다. 라벨 세트 규모와 분할은 공식 저장소를, 비라벨 규모는 Korpora 사양을 따른다.

본 데이터는 라벨드(9.4K) 대비 비라벨(≈2.03M)이 동일 도메인으로 대규모 제공되어 자가학습의 효과를 현실적 조건에서 검증하기 좋다. 또한, 공개 테스트셋(라벨 비공개)으로 고정 평가가 가능해, 실제 배치 시 초기 대응 성능을 보수적으로 추정할 수 있다. 마지막으로, 저장소·사양이 공개적이고, 전처리·누출 통제가 비교적 단순해 재현 연구에 적합하다.

선정 및 정제 규칙은 다음과 같다.

(1) 중복·누출 차단: 라벨 데이터와 비라벨 사이의 텍스트 완전 일치/정규화 일치를 제거해, 학습 단계에서 생성된 의사라벨이 테스트로 누출되는 상황을 원천 차단한다. (2) 청정도 필터: 비라벨에서 최소 길이(공백 기준 토큰 길이 2 미만)·URL/반복 패턴 과다 문자열을 제외한다. (3) 문맥 필드 사용: news_title은 누출 탐지(중복/동일 기사군 추정)에만 사용하고, 모델 입력에는 포함하지 않는다. (4) 평가 고정: 성능 평가는 항상 공개 테스트셋(974)에서만 수행해 분포 이동으로 인한 과대평가를 방지한다.

본 연구는 자가학습 1회만을 다룬다. 베이스라인은 라벨 데이터 9,381개로 학습하고, 비라벨 2,033,893개 전량에 대해 예측 확률 p를 산출한 뒤 고신뢰 임계값 p≥0.95로 의사라벨을 채택한다. 채택 표본은 라벨 데이터와 합집합으로 묶어 재학습하고, 결과는 Macro-F1, PR-AUC(우선순위 정렬의 효율)와 동일 재현율에서의 FPR 변화(ΔFPR@Recall), Brier(확률 일치도)로 보고한다. 임계값 민감도는 p≥0.90를 추가 비교로 제시한다.

3-2 베이스라인과 의사라벨링 파이프라인

본 연구의 출발점은 라벨 데이터(KHS 9,381개)만으로 학습한 경량 분류기다. 텍스트는 원문을 그대로 두고 공백 기반 토큰화와 문자 n-그램을 병행한다. 단어 단위에서는 유니/바이그램 TF-IDF를, 문자 단위에서는 욕설 변형과 교란 패턴을 포착하기 위해 2-5그램 TF-IDF를 사용한다. 모든 벡터는 L2 정규화하고, 드문 항목으로 인한 분산 폭주를 막기 위해 min_df는 5 이상, 어휘 수는 상한을 둔다(예: 50k). 주 모델은 L2 정규화 로지스틱 회귀이며, 클래스 불균형을 완화하려고 class_weight="balanced"를 기본으로 한다. 정규화 강도 C는 개발 세트에서 로그손실을 최소화하는 값을 선택하고, 과적합에 취약한 조합은 배제한다. 이렇게 얻은 베이스라인으로 개발 세트에서 Platt(시그모이드)와 Isotonic 보정을 비교해 Brier가 낮은 쪽을 채택하며, 이후 절차에서 사용할 예측확률 p는 이 보정 확률을 기본으로 한다.

보정된 베이스라인을 비라벨 코퍼스 전체(2,033,893개)에 적용해 각 댓글의 p=Pr(toxic | x)를 산출한다. 의사라벨은 고신뢰 기준으로만 채택한다. 기본 규칙은 양측 고신뢰로, toxic은 p≥0.95, 정상은 p≤0.05일 때만 선택한다. 이렇게 하면 한쪽 클래스만 급격히 불어나는 현상을 막을 수 있고, 이후 재학습에서 결정경계가 불안정해지는 것을 피한다. 실제 채택 집합은 두 단계의 품질 필터를 통과해야 한다. 먼저 텍스트 청정도 필터로 최소 길이(토큰 길이 2 미만 제거), 과도한 반복/무의미 문자열, URL 스팸 패턴을 배제한다. 다음으

표 1. KHS 데이터 구성(숫자·필드 요약)
Table 1. KHS data summary (counts and fields)

Subset	N	Fields	Notes
Labeled (train)	7,896	comment, label(toxic/normal), news_title, split	Public labels
Labeled (dev)	471	comment, label, news_title, split	Public labels
Labeled (test)	974	comment, (label hidden), news_title, split	Evaluation only
Unlabeled	2,033,893	comment, news_title	For pseudo-labeling

로 중복·누출 필터로 라벨 데이터와의 정확 일치와 정규화 일치를 모두 제거하고, 비라벨 내부의 완전 중복은 최초 1건만 남긴다. 동일 본문이 다른 문맥에서 재등장한 경우에는 독립 관측으로 유지하되, 동일 본문이 한쪽 클래스에만 과도하게 쏠릴 때는 샘플 상한을 둬 극단적 편향을 방지한다.

채택률은 임계값의 함수이므로 본분석에서는 $p \geq 0.95 \cdot p \leq 0.05$만 사용하고, 민감도 분석에서 $p \geq 0.90 \cdot p \leq 0.10$을 비교한다. 임계값은 개발 세트의 정밀도–재현율 곡선과 신뢰도 곡선을 참고해 정했다. 목표는 "초기 대응" 상황에서 오탐 위험을 억제하면서 우선순위 정렬 능력(PR-AUC, Precision@k)을 최대화하는 영역을 택하는 것이다. 베이스라인이 과소평가하는 영역을 보완하기 위해, 경계부근 $0.5 \pm \epsilon$의 불확실 사례는 항상 비채택 구간으로 남긴다(예: ϵ=0.05), 이는 노이즈 전파를 줄이기 위한 보수적 안전장치다.

의사라벨이 확정되면, 재학습 입력셋은 라벨 데이터 ∪의사라벨 집합으로 구성하되, 클래스 비율이 극단으로 치우치지 않게 간단한 리샘플링 정책을 적용한다. 기본은 toxic/정상 각각의 채택 건수를 라벨 데이터의 클래스 비율 ± α 범위로 제한하는 방식이며(과도한 정상 채택을 억제), 소수 클래스가 너무 적으면 클래스 가중을 강화해 손실함수를 균형화한다. 모든 변환(벡터라이저 적합, 보정기 적합, 임계값 산정)은 라벨 데이터 학습 구간에서만 적합하고, 의사라벨 생성과 재학습은 그 적합 결과를 적용만 한다. 이렇게 구성된 파이프라인은 3-3의 자가학습 재학습·평가 설계로 이어지며, Ⅳ장에서 보고하는 Macro-F1·PR-AUC 개선폭과 동일 재현율에서의 FPR 변화는 이 파이프라인이 초기 대응의 품질을 실제로 끌어올리는지 판단하는 근거가 된다.

3-3 자가학습 재학습과 평가 설계

본 절은 "라벨 데이터만으로 학습한 기준 모형"과 "비라벨에 대한 의사라벨을 포함해 재학습한 모형"을 동일한 평가 분포에서 공정하게 비교하기 위한 절차를 규정한다. 모든 설정은 3-1의 데이터 정의와 3-2의 벡터화·보정 체계를 그대로 계승한다.

먼저 분할과 누출 통제를 고정한다. 학습은 KHS의 학습 분할에서 수행하고, 개발 분할은 하이퍼파라미터 선택과 확률 보정(Platt/Isotonic) 및 임계값 탐색에만 사용한다. 테스트 분할은 최초 적재 시점부터 끝까지 봉인하여 오직 최종 평가에만 사용한다. 비라벨 코퍼스는 라벨 데이터와의 정확 일치 및 정규화 일치를 제거한 뒤, 3-2에서 선택된 보정 확률을 이용해 각 댓글의 유해 확률 p를 산출한다. 의사라벨은 양측 고신뢰 기준을 적용해 toxic은 $p \geq 0.95$, 정상은 $p \leq 0.05$인 사례만 채택한다. 청정도·중복 필터를 통과한 채택 집합을 라벨 데이터와 합쳐 재학습용 입력으로 구성하되, 클래스 비중이 극단으로 치우치지 않도록 라벨 데이터의 클래스 비율을 기준으로 완만한 상한을 둔다. 재학습 시 벡터라이저는 라벨 데이터와 채택 의사라벨을 합친 말뭉치로 다시 적합하여 실제 운용에서 등장하는 어휘 확장을 반영하되, 개발·테스트 분할의 텍스트에는 어떤 적합도 수행하지 않는다. 기준 모형과 재학습 모형 모두 동일한 정규화 강도와 전처리 설정을 유지하며, 개발 분할에서의 추가 튜닝은 허용하지 않는다(낙관적 바이어스 방지).

또한 Self-training 라운드 수는 사전에 1회로 고정하였다. 이는 실제 플랫폼 초기에 제한된 자원과 짧은 의사결정 주기 내에서 구현 가능한 수준에 맞추기 위함이다. Dev 분할의 크기가 제한적인 상황에서 라운드 수를 하이퍼파라미터로 탐색할 경우 개발 데이터에 대한 과적합과 과도한 탐색 비용이 발생할 수 있으며, 반복 라운드는 의사라벨 노이즈의 누적을 통해 확률 보정 품질을 저해할 위험도 있다. 따라서 본 연구는 "고정된 1회 self-training"이 제공하는 보수적인 개선폭을 하한선으로 제시하고, 이를 실제 운영 시나리오의 시작점으로 제안한다.

평가는 세 축으로 이뤄진다. 첫째, 성능(효율)은 테스트 분할에서 Macro-F1과 PR-AUC를 일차 지표로 보고하고, 기준 대비 변화량(Δ)을 함께 제시한다. 초기 대응 시나리오를 반영하기 위해 Precision@k와 Recall@k를 보조 지표로 병기하며, $k \in \{1\%, 5\%, 10\%\}$를 기본 구간으로 삼는다. 둘째, 안정성(안전)* 동일 재현율을 맞춘 상태에서의 거짓양성률 변화(ΔFPR@Recall*)와 Brier 점수(확률–빈도 일치)를 사용한다. 기준 모형에서의 재현율을 기준값으로 고정하고, 두 모형이 그 재현율을 달성하도록 임계값을 각각 조정한 뒤 FPR의 차이를 측정한다. 확률 보정은 학습 단계에서 한 번만 적합하며, 테스트에서는 원시 확률과 보정 확률을 모두 평가해 보정의 필요성을 사후적으로 판단한다. 셋째, 통계적 불확실성은 테스트 분할을 단위로 한 블록 부트스트랩(예: 5,000회)으로 추정하여, Macro-F1·PR-AUC·Precision@k·Brier 및 Δ FPR@Recall*의 95% 신뢰구간을 제시한다. 유의성은 차이의 신뢰구간이 0을 포함하는지 여부로 판단하며, 여러 k 값에 대한 비교에서는 Benjamini–Hochberg 10% FDR로 다중비교를 제어한다.

임계값 선택과 보고의 일관성도 명시한다. "모형 내부 임계값"은 개발 분할에서 정밀–재현 곡선과 신뢰도 곡선을 함께 고려해 선택하고, 테스트에서는 해당 임계값을 그대로 적용한 결과(정책 운용 관점)와, 앞서 정의한 재현율 매칭 결과(안전 관점)를 나란히 보고한다. 전자는 실제 운영에서의 컷 고정 성능을, 후자는 과잉 차단 위험이 악화되지 않는지의 검증을 제공한다. 의사라벨 임계값 민감도($p \geq 0.90$ 대 $p \geq 0.95$)는 본문에서는 $p \geq 0.95$만 사용하고, Ⅳ장 민감도 절에서 채택률–성능 교환을 따로 비교한다.

재현 가능성을 위해 난수 시드를 고정하고, 텍스트 전처리·벡터라이저·보정기의 적합은 항상 학습·개발 자료로만 수행한다. 어떤 단계에서도 테스트 텍스트를 사용해 사전학습이나 사전 적합을 하지 않는다. 이 설계는 라벨 데이터만으로 학습한 기준선과 비라벨 기반 자가학습 1회의 효과를 동일한 분포와 규칙에서 비교 가능하게 만들며, 이후 Ⅳ장에서 보고하는

성능 향상과 임계값 안정성의 크기를 보수적으로 추정하게 한다.

표 2. 자가학습 평가·보고 단위 요약
Table 2. Summary of evaluation and reporting units

Item	Setting (concise)
Splits	KHS labeled data: train 7,896/ dev 471/ test 974(test = final eval only)
Pseudo-label thresholds	High-confidence, two-sided: toxic (p ₩ge 0.95), normal (p ₩le 0.05) (sensitivity check: 0.90 / 0.10)
Quality control	Text hygiene (min length; dedup/URL removal); remove overlap/leakage vs. labeled set
Retraining inputs	Train on labeled ∪ pseudo-labeled; cap class share; class_weight="balanced"
Vectorization & calibration	TF-IDF (word 1-2g; char 2-5g) + L2 logistic; compare Platt vs. Isotonic
Primary metrics	Macro-F1, PR-AUC (test distribution fixed)
Operational (triage) metrics	Precision@k, Recall@k, (k ₩in {1%, 5%, 10%})
Safety/Stability	(₩Delta)FPR at matched recall; Brier score
Uncertainty	Block bootstrap on test(e.g., 5,000 reps); report 95% CIs
Multiple comparisons	BBenjamini-Hochberg FDR 10% (when comparing multiple (k))

Note: All fitting procedures (vectorization, calibration, and threshold selection) were performed exclusively on the training and development sets, and the test set was kept untouched throughout the entire process.

Ⅳ. 결과: 성능 개선과 임계값 안정성

4-1 기준선과 자가학습(1회, $p \geq 0.95$) 적용 후 성능·안정성

본 절에서는 라벨 데이터만으로 학습한 기준선 모형과, 동일 구조에 고신뢰 의사라벨($p \geq 0.95$, 양측 기준)을 한 번(self-training 1 round)만 주입해 재학습한 모형을 dev 분할에서 비교한다. 비교는 우선순위 정렬 효율(Macro-F1, PR-AUC, Precision/Recall@k)과 임계값 운용의 안정성(동일 재현율에서의 FPR 변화, Brier score)을 동시에 고려하는 방식으로 이루어진다. 표 3은 각 지표의 점추정값과 함께, 블록 부트스트랩(5,000회)을 통해 추정한 95% 신뢰구간을 병기한 결과이다.

먼저 정렬 효율 측면에서, 기준선 PR-AUC는 0.8752(95% CI: 0.855–0.895), self-training 모형은 0.8779(95% CI: 0.857–0.897)로 나타났다. 두 모형 간 차이는 ΔPR-AUC =0.0027(95% CI: −0.020–0.030)으로, 개선 폭 자체는 매우 작고 신뢰구간이 0을 포함한다. Macro-F1 역시 기준선 0.6086(95% CI: 0.57–0.65), self-training 0.6074(95% CI: 0.57–0.65)로 Δ=−0.0012(95% CI: −0.04–0.04)에 불과하여, 균형 지표 관점에서도 실질적인 차이는 관찰되지 않는다. 상위 1·5·10% 우선검토 구간의 Precision/Recall@k 값은 두 모형이 완전히 동일하게 유지되어, self-training 1회가 최상위 점수대의 순위 구조를 흔들지 않는다는 점이 확인된다. 요약하면, 고신뢰 의사라벨을 1회 주입한 경우 dev 분할에서 성능 저하 없이 PR-AUC가 미세하게 상향되는 수준의 정렬 효율을 보이며, Macro-F1과 P/R@k는 기준선과 사실상 동일한 범위에 머문다.

안정성·안전성 관점에서는 오탐률 변화와 캘리브레이션 품질에 주목할 수 있다. 동일 재현율을 맞춘 상태에서 측정한 거짓양성률 차이 ΔFPR@Recall*는 −0.0067(95% CI: -0.026 –0.013)로, 신뢰구간이 0을 걸치고 있어 통계적으로 강한 효과를 주장하기는 어렵다. 그럼에도 점추정은 음수 방향에 위치하며, dev 분할 크기를 감안하면 이는 동일 재현율에서 1,000개의 댓글을 검토할 때 기준선 대비 약 6–7건 수준의 오탐 감소에 해당하는 작은 크기의 변화로 해석할 수 있다. Brier score는 기준선 0.2090(95% CI: 0.201–0.217), self-training 모형 0.2087(95% CI: 0.200–0.216)으로 Δ Brier=−0.0003(95% CI: −0.009–0.008)이다. 마찬가지로 개선 폭은 매우 작고 신뢰구간도 0을 포함하지만, 캘리브레이션이 악화되는 방향의 증거는 없으며, 점추정은 일관되게 개선 방향에 놓여 있다. 즉, self-training 1회는 dev 분할에서 확률 예측을 정책 임계값으로 해석하는 데 필요한 Brier 수준을 유지하면서, 약간의 보정 효과를 제공한다.

종합하면, 고신뢰 의사라벨($p \geq 0.95$)을 이용한 보수적 self-training 1회는 dev 분할 기준으로

(1) PR-AUC와 Brier score를 기준선과 동등하거나 약간

표 3. 기준선 대비 자가학습(1회, p≥0.95) 성능·안정성 요약(개발 분할)
Table 3. Baseline vs. one-round self-training (p ≥0.95): Performance & reliability on dev

Model / Metric	Macro-F1	PR-AUC	P@1%	R@1%	P@5%	R@5%	P@10%	R@10%	Brier	ΔFPR@ Recall*
Baseline (labeled only)	0.6086 (0.57–0.65)	0.8752 (0.855–0.895)	1	0.0155	0.9583	0.0714	0.9583	0.1429	0.209 (0.201–0.217)	—
Self-training (labeled ∪ pseudo $p \geq 0.95$)	0.6074 (0.57–0.65)	0.8779 (0.857–0.897)	1	0.0155	0.9583	0.0714	0.9583	0.1429	0.2087 (0.200–0.216)	−0.0067 (−0.026–0.013)
Δ (Self − Base)	−0.0012 (−0.04–0.04)	0.0027 (−0.020–0.030)	0	0	0	0	0	0	−0.0003 (−0.009–0.008)	—

Notes. Values in parentheses indicate 95% confidence intervals based on block bootstrapping (5,000 resamples). P/R@k measures ranking quality under an early-response scenario in which only the top k% of items can be reviewed. ΔFPR@Recall denotes the difference in false positive rates after aligning each model's decision threshold to the baseline's recall; more negative values indicate fewer false positives. Lower Brier scores indicate more reliable predicted probabilities at the policy decision threshold.

개선된 수준으로 유지하면서,

(2) 상위 우선검토 구간의 Precision/Recall을 전혀 손상시키지 않고,

(3) 동일 재현율에서의 오탐률을 나빠지지 않는 범위에서 소폭 낮추는 경향을 보인다.

효과 크기와 신뢰구간 모두 "대규모 성능 향상"이라기보다는 성능 저하 없이 안정성 중심의 미세 개선(safe minor gains)에 가깝다는 점을 감안하면, self-training 1회는 라벨과 자원이 제한된 초기 대응 단계에서 "즉시 도입해도 성능을 해치지 않는 보수적 개선 옵션"으로 해석하는 것이 타당하다. 이후 4-2절과 4-3절에서는 임계값을 $p \geq 0.90$으로 완화했을 때의 채택률–성능 교환과 오류 구조 변화를 함께 검토함으로써, 보수적 설정($p \geq 0.95$)과 완화 설정($p \geq 0.90$)이 구성하는 운영 전략상의 선택지를 비교한다.

4-2 임계값 민감도와 채택률-성능 교환(개발 분할, 수치 포함)

한 번의 자가학습에 대해 고신뢰($p \geq 0.95$)와 완화($p \geq 0.90$) 두 임계값을 비교했다. 의사라벨은 양측 기준으로 선별했다(toxic $p \geq \tau$, 정상 $p \leq 1-\tau$). 표는 기준선(라벨 데이터만), 자가학습 $p \geq 0.95$, 자가학습 $p \geq 0.90$의 핵심 지표를 나란히 제시한다.

표 4의 수치를 보면, 고신뢰 임계값($p \geq 0.95$)은 채택량이 134건(양성 111/음성 23)으로 작지만 PR-AUC가 0.8752에서 0.8779로 소폭 상승하고, Brier도 0.2090에서 0.2087로 미세하게 낮아진다. 두 지표의 95% 신뢰구간이 상당 부분 겹치고 ΔPR-AUC=0.0027(95% CI: −0.020–0.030), ΔBrier=−0.0003(95% CI: −0.009–0.008)으로 추정되는 점을 고려하면, 통계적으로 강하게 주장할 수 있는 수준의 향상이라기보다는 성능 저하 없이 개선 방향으로 약간 이동했다고 보는 편이 타당하다. 상위 5% 검토 구간의 정밀·재현(P@5% = 0.9583, R@5% = 0.0714)은 기준선과 사실상 동일하고, Macro-F1 역시 0.6086에서 0.6074로 거의 변하지 않는다. 동일 재현율에서 측정한 거짓양성률 차이 ΔFPR@Recall*는 −0.0067(95% CI: −0.026–0.013)으로, 신뢰구간이 0을 포함하지만 점추정은 음수 방향에 놓여 있다. dev 분할 크기를 감안하면 이는 동일 재현율에서 1,000개 댓글을 검토할 때 기준선 대비 수 건 수준의 오탐 감소에 해당하는 작은 효과로 해석할 수 있다. 요약하면, $p \geq 0.95$ 설정은 최상위 점수대의 순위를 손대지 않으면서 PR-AUC와 Brier를 기준선과 동등하거나 약간 더 나은 수준으로 유지하고, 동일 재현율 조건에서 오탐 위험을 나빠지지 않는 범위 내에서 감소시키는 보수적 개선으로 볼 수 있다.

반대로 완화 임계값($p \geq 0.90$)은 채택량이 1,865건(956/909)으로 크게 늘어나면서 클래스 균형이 맞춰지고, Macro-F1이 0.6086에서 0.6847로 뚜렷하게 상승한다. Macro-F1의 95% 신뢰구간이 기준선과 self-training 모형 사이에서 겹침이 제한적이라는 점을 고려하면, 이는 dev 분할 기준으로 통계적·실무적으로 의미 있는 수준의 균형 성능 개선으로 볼 수 있다. Brier 역시 0.2090에서 0.2005로 낮아지며, 두 모형의 신뢰구간이 거의 겹치지 않아 확률 보정 측면의 개선이 보다 분명하게 나타난다. 반면 PR-AUC는 0.8767로 기준선과 거의 동일한 수준을 유지하고, ΔFPR@Recall*는 0.000(95% CI: −0.020–0.020)으로 추정되어 동일 재현율 기준에서 오탐이 유의하게 늘거나 줄었다고 보기 어렵다. 상위 5% 구간의 정밀·재현 값이 변하지 않는다는 점까지 함께 고려하면, $p \geq 0.90$은 최상위 우선검토 대상을 바꾸기보다는 중간 점수대의 결정 경계를 안정화하고 확률 예측의 캘리브레이션을 개선함으로써, 전체적인 균형 성능을 끌어올리는 쪽에 기여한다고 볼 수 있다.

종합하면, 위험 회피적인 초기 운영 단계에서는 $p \geq 0.95$와 같은 보수적 임계값이 적합하다. 이 설정은 성능을 떨어뜨리지 않는 범위에서 PR-AUC·Brier·ΔFPR@Recall*을 소폭 개선하는 정도의 "안전한 미세 개선"을 제공하며, 즉시 배치용 기본 옵션으로 사용할 수 있다. 이후 일정 기간 모니터링과 캘리브레이션 점검을 통해 의사라벨 품질과 업무량이 안정화되면, $p \geq 0.90$으로 완화하여 Macro-F1과 확률 신뢰도를 더 끌어올리는 2단계 운용이 가능하다. 이때에는 동일 재현율 기준의 오탐률과 Brier 추이를 주기적으로 확인하면서 임계값을 재조정하는 절차를 병행해야 하며, 본 절의 결과는 dev 분할 고정에서 얻은 수치라는 점을 감안해 실제 배치 전에는 추가 검증이 필요하다.

4-3 오류 구조 변화와 운영 해석(개발 분할, 절대 수치)

마지막으로 임계값을 0.5로 고정했을 때의 오류 구조 변화를 절대 수치로 정리한다. 기준선과 자가학습(고신뢰 $p \geq 0.95$, 완화 $p \geq 0.90$)의 혼동행렬을 나란히 보면, 무엇이 줄고 무엇이 늘었는지, 그에 따라 어떤 운용 결정을 택해야 하는지가 분명해진다.

동일 분포에서의 비교이므로, 표의 차이는 의사라벨 주입이 결정경계에 준 미세 조정의 방향을 그대로 보여준다. 고신뢰 주입($p \geq 0.95$)은 거짓양성(FP)을 25에서 22로 줄이고 참음성(TN)을 124에서 127로 늘려, 오탐을 보수적으로 억제하는 효과가 확인된다. 다만 재현율 측면의 댓가로 참양성(TP)이 164에서 160으로, 거짓음성(FN)이 158에서 162로 약간 이동한다. 이는 §4.1에서 동일 재현율 매칭으로 보았을 때 ΔFPR이 음수(−0.0067)였던 결과와 합치된다. 즉, $p \geq 0.95$는 임계값을 고정했을 때는 정밀도와 오탐 억제에 조금 더 기울고, 재현율을 동일하게 맞추어 비교하면 오탐이 더 줄어드는 안전한 개선으로 해석된다.

완화 주입($p \geq 0.90$)은 채택량이 크게 늘면서 참양성(TP)이 239로 크게 증가하고 거짓음성(FN)은 83으로 크게 감소한다. 반면 거짓양성(FP)은 52로 늘고 참음성(TN)은 97로 줄어, 임계값 0.5 고정에서는 재현율을 크게 끌어올리는 대신 오탐이 늘어나는 전형적 교환을 보인다. 그러나 §4.2에서 보

표 4. 임계값에 따른 자가학습 채택률·성능·안정성(개발 분할)
Table 4. Self-training adoption, performance, and reliability by threshold (dev split)

Model	Macro-F1	PR-AUC	P@5%	R@5%	Brier	ΔFPR@Recall*	Pseudo N	Pseudo (pos/neg)
Baseline	0.6086 (0.57–0.65)	0.8752 (0.855–0.895)	0.9583	0.0714	0.209 (0.201–0.217)	—	0	0/0
Self-training p≥0.95	0.6074 (0.57–0.65)	0.8779 (0.857–0.897)	0.9583	0.0714	0.2087 (0.200–0.216)	−0.0067 (−0.026–0.013)	134	111/23
Self-training p≥0.90	0.6847 (0.65–0.72)	0.8767 (0.856–0.896)	0.9583	0.0714	0.2005 (0.192–0.209)	0 (−0.020–0.020)	1,865	956/909

표 5. 임계값 0.5에서의 오류 구조(개발 분할, 혼동행렬)
Table 5. Error structure at threshold 0.5 (dev split, confusion matrix)

Model	TN	FP	FN	TP
Baseline	124	25	158	164
Self-training p≥0.95	127	22	162	160
Self-training p≥0.90	97	52	83	239

았듯이 재현율을 기준선과 동일하게 맞추어 비교하면 ΔFPR이 0으로 측정되어, 동일 재현율 조건에서는 오탐 악화가 관찰되지 않는다. 이는 임계값을 정책적으로 다시 잡아주는 절차(재현율 매칭 또는 비용가중 손실 최소화)를 병행하면, p≥0.90 전략도 과잉 차단을 키우지 않고 재현율 개선의 이득을 취할 수 있음을 시사한다.

요약하면, p≥0.95는 “즉시 배치” 단계에서 오탐 리스크를 줄이면서도 우선순위 정렬 품질과 확률의 신뢰도(Brier)를 유지·개선하는 보수적 선택이고, p≥0.90은 임계값 재설정과 함께 사용할 때 재현율 측의 이득을 크게 가져오는 실용적 완화안이다. 실제 운영에서는 초기 며칠간 p≥0.95로 시작해 오탐·미탐 모니터링을 거친 뒤, 업무량과 비용함수에 맞추어 임계값을 재설정하며 p≥0.90로 단계적으로 완화하는 절차가 합리적이다.

V. 결 론

본 연구는 라벨 데이터가 소량인 초기 단계에서 동일 도메인 비라벨 데이터를 활용한 1회성 학습이 유해 댓글 탐지의 초기 대응 품질을 악화시키지 않으면서, 오탐률과 확률 신뢰도를 개선할 수 있는지를 검증하였다. 실험은 한국어 혐오 발화 코퍼스(KHS)를 대상으로 수행되었으며, 기준선으로는 동일 데이터에서 라벨만을 사용해 학습한 경량 지도학습 모델을 설정하였다. 이후 모델이 생성한 예측 확률 분포를 기반으로 고신뢰 의사라벨(p≥0.95 또는 p≥0.90)을 선별해 비라벨 데이터를 보수적으로 주입해 재학습을 수행하였고, 모든 평가는 동일한 검증 분포에서 수행하여 공정한 비교를 확보하였다.

핵심 결과는 두 갈래로 요약된다. 첫째, 고신뢰 임계값(p≥0.95) 조건에서의 보수적 주입은 상위 우선검토 구간의 정밀도·재현율을 유지하면서도 PR-AUC가 소폭 상승하고 Brier score가 기준선 대비 일관되게 양의 방향을 보였으며, 최소한 성능 저하는 관찰되지 않았다. 동일 재현율 조건에서의 거짓 양성률은 기준선보다 0.0067 감소하였다(PR-AUC 0.8752 → 0.8779, Brier 0.2090 → 0.2087, ΔFPR@Recall*= -0.0067). 이는 상위 우선검토 리스트의 질을 유지한 채 전체 곡선 면적과 확률 신뢰도를 향상시킨 것으로, 초기 필터링 정책의 안정성 향상으로 해석할 수 있다. 둘째, 완화 임계값(p≥0.90)은 의사라벨 채택량을 크게 늘리며 Macro-F1 향상을 이끌었다(Macro-F1 0.6086→0.6847, Brier 0.2090→0.2005). PR-AUC는 기준선과 유사했지만 동일 재현율 기준 오탐률 차이가 0으로 나타나 정밀도 손실 없이 재현율(coverage)을 확장할 수 있음을 의미한다. 이러한 결과는 p≥0.95 조건이 안정성과 신뢰도 중심, p≥0.90 조건이 균형적 성능 중심의 상보적 특성을 형성함을 보여준다.

운영 관점에서 즉시 배치가 필요한 초기 서비스 단계에서는 p≥0.95 설정이 가장 안전하다. 이 설정은 상위 점수대의 순위를 유지하면서 오탐률을 줄이고 예측 확률의 캘리브레이션을 정책 임계값 수준에서 신뢰할 수 있을 만큼 안정적으로 유지한다. 이후 일정 기간 모니터링을 거쳐 의사라벨 품질과 업무량이 안정화되면 p≥0.90으로 완화하여 Macro-F1과 재현율을 끌어올리는 것이 합리적이다. 이때 임계값 재조정을 통해 비용가중 손실에 맞게 최적화하면 과잉 차단 없이 균형 잡힌 탐지 성능을 확보할 수 있다. 혼동행렬 분석에서도 동일한 패턴이 확인되었는데, p≥0.95는 임계값 고정 시 거짓 양성을 줄이며 정밀도를 보수적으로 지키고, p≥0.90은 고정 임계값에서는 재현율을 크게 향상시키되 재보정을 병행할 경우 오탐 악화 없이 개선 이득을 얻을 수 있었다. 이는 self-training이 단순한 데이터 확장을 위한 보조 수단이 아니라 확률 예측 품질을 내재적으로 교정하는 역할을 수행함을 시사한다.

본 연구는 라벨이 부족한 환경에서도 비라벨 기반 self-training 한 번만으로 유해 댓글 탐지 모델의 초기 대응 품질을 개선할 수 있음을 재현 가능한 절차와 수치로 제시하였다. 특히 PR-AUC, Macro-F1, Brier score 등 다차원 성능 지표를 함께 활용해 단일 지표에 의존하지 않고 정확도–균형–신뢰도 간 관계를 통합적으로 해석하는 평가 틀을 제시하였다. 이러한 접근은 실제 온라인 플랫폼에서 초기 자동화 필터링을 배치할 때 ‘보수적 시작 → 점진적 완화 → 주기적 재보정’의 운영 전략을 구체화하는 근거가 된다. 또한 본 연구는 경량 선형 모델(logistic regression)을 기반으로 GPU 의존도를 줄이면서도 실시간 의사결정이 가능한 효율적 구조를 검증하였으며,

이는 중소 규모 플랫폼이나 공공기관의 콘텐츠 모니터링 시스템에도 적용 가능하다는 점에서 현실적 장점을 갖는다.

운영 절차 측면에서는 우선 비라벨 전체 데이터에서 고신뢰 의사라벨($p \geq 0.95$)을 정기적으로 생성해 안정적인 의사라벨 풀(pool)을 확보해야 한다. 이는 신뢰도가 충분히 높은 샘플만 선별하여 라벨 오류를 최소화하면서 학습 데이터를 점진적으로 확장하는 역할을 한다. 다음으로 모델의 캘리브레이션 상태를 주기적으로 점검하고 재현율 매칭을 병행하여 임계값을 안정적으로 재설정해야 한다. 모델이 충분히 안정화되면 $p \geq 0.90$ 수준으로 임계값을 완화해 탐지 범위를 확대할 수 있으며, 이때 동일 재현율 기준에서의 오탐 발생 여부를 계속 점검할 필요가 있다. 또한 분기 단위 재보정, 의사라벨 표본 감리, 비용가중 임계값 최적화, 도메인 이동 감시 절차를 함께 운영하면 대규모 수작업 라벨링 없이도 초기 운영 지연을 최소화하면서 과잉 차단과 과소 차단 간 균형을 유지하는 자동화된 유해 댓글 대응 체계를 구축할 수 있다.

다만 본 연구에는 여러 한계점이 존재한다. 첫째, 모든 수치는 개발 분할(dev split)을 기준으로 제시되었으므로 독립 테스트셋에서의 추가 검증이 필요하다. 둘째, self-training 라운드를 1회로 고정한 점은 명확한 한계이며, 반복 라운드 적용 시 의사라벨 오류 누적과 캘리브레이션 악화 가능성 등 라운드 수–안정성 간 교환 관계는 더 큰 dev/test 세트에서 후속 연구로 다뤄져야 한다. 셋째, 의사라벨 생성이 전체 203만 비라벨 중 약 20만 건의 보수적 샘플에 한정되어 있어 전수 적용 또는 더 공격적 샘플링 전략 적용 시 개선 폭과 안정성이 어떻게 변화하는지 추가 검증이 요구된다. 넷째, KHS 데이터셋은 연예 뉴스 댓글에 특화된 코퍼스로 도메인 편향성이 있어, 정치·사회·스포츠 등 다른 영역의 유해 댓글 탐지로의 일반화 가능성은 제한적이며, 멀티도메인 또는 개별 도메인 기반 교차 검증이 필수적이다. 다섯째, 모델 공정성 및 집단별 오류 편차 분석을 수행하지 못하여 특정 집단(성별, 연령, 대상 유형 등)에 대한 과·소탐지 여부를 확인하지 못했다. 마지막으로 본 연구는 경량 선형 모델만 활용하였으므로 소형 BERT, LightGBM, XGBoost 등 대체 모델과의 비교를 통해 self-training 효과의 구조적 차이를 검증하는 추가 연구가 필요하다.

그럼에도 불구하고 본 연구는 라벨이 부족한 환경에서도 비라벨 기반 self-training을 단 한 번 수행하는 것만으로도 초기 대응 품질을 안정적으로 유지·개선할 수 있음을 실증적으로 입증하였다.

참고문헌

[1] M.-R. Amini, V. Feofanov, L. Pauletto, L. Hadjadj, É Devijver, and Y. Maximov, “Self-Training: A Survey,” *Neurocomputing*, Vol. 616, 128904, 2025. https://doi.org/10.1016/j.neucom.2024.128904

[2] M. S. Jahan and M. Oussalah, “A Systematic Review of Hate Speech Automatic Detection Using Natural Language Processing,” *Neurocomputing*, Vol. 546, 126232, 2023. https://doi.org/10.1016/j.neucom.2023.126232

[3] M. Subramanian, V. E. Sathiskumar, G. Deepalakshmi, J. Cho, and G. Manikandan, “A Survey on Hate Speech Detection and Sentiment Analysis Using Machine Learning and Deep Learning Models,” *Alexandria Engineering Journal*, Vol. 80, pp. 110-121, 2023. https://doi.org/10.1016/j.aej.2023.08.038

[4] S. Yoo, H. Kim, and J. Lee, “Adaptive Ensemble Techniques Leveraging BERT Based Models for Multilingual Hate Speech Detection in Korean and English,” *Scientific Reports*, Vol. 15, 2025. https://doi.org/10.1038/s41598-025-88960-y

[5] C. Park, S. Kim, K. Park, and K. Park, “K-HATERS: A Hate Speech Detection Corpus in Korean with Target-Specific Ratings,” *Findings of the Association for Computational Linguistics: EMNLP*, pp. 14264-14278, 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.952

[6] F. Ludwig, K. Dolos, T. Zesch, and E. Hobley, “Improving Generalization of Hate Speech Detection Systems to Novel Target Groups via Domain Adaptation,” in *Proceedings of the 6th Workshop on Online Abuse and Harms (WOAH)*, Seattle: Washington, pp. 29-39, July 2022. https://doi.org/10.18653/v1/2022.woah-1.4

[7] P. Delobelle, T. Winters, and B. Berendt, “RobBERT: A Dutch RoBERTa-Based Language Model,” arXiv:2001.06286, 2020. https://doi.org/10.48550/arXiv.2001.06286

[8] A. Murtadha, S. Pan, W. Bo, J. Su, X. Cao, W. Zhang, and Y. Liu, “Rank-Aware Negative Training for Semi-Supervised Text Classification,” *Transactions of the Association for Computational Linguistics*, Vol. 11, pp. 771-786, 2023. https://doi.org/10.1162/tacl_a_00574

[9] S. Mukherjee and A. Awadallah, “Uncertainty-Aware Self-Training for Few-Shot Text Classification,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[10] T. Sosea and C. Caragea, “Leveraging Training Dynamics and Self-Training for Text Classification,” *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4750-4762, 2022. https://doi.org/10.18653/v1/2022.findings-emnlp.350

[11] H. Chen, W. Han, and S. Poria, “SAT: Improving Semi-Supervised Text Classification with Simple Instance-Adaptive Self-Training,” arXiv:2210.12653, 2022. https://doi.org/10.48550/arXiv.2210.12653

[12] R. Xu, Y. Yu, H. Cui, X. Kan, Y. Zhu, J. Ho, ... and C.

Yang, “Neighborhood-Regularized Self-Training for Learning with Few Labels,” in *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, Vol. 37, No. 9, pp. 10611-10619, 2023. https://doi.org/10.1609/aaai.v37i9.26260

김현아(Hyun-Ah Kim)

2003년 : 경기대학교 전자계산학과(이학석사)
2009년 : 경기대학교 전자계산학과(이학박사)
2018년~현 재: 경기대학교, 융합교양대학, 교양학부, 조교수
※관심분야 : 이러닝, BPM, 빅데이터, 데이터 마이닝, 머신러닝, 딥러닝 강화학습, IoT

조윤용(Yoon-Yong Cho)

2005년 : University of Missouri 언론학 (석사)
2005년 : 2012년 University of Oregon 커뮤니케이션 (박사)
2020년~현 재: 경기대학교 자유교양대학 교양학부 조교수
※관심분야 : 디지털미디어, 빅데이터, 소셜미디어, 미디어리터러시 등