온라인 댓글에서의 정제된 피드백 제공 서비스, CleansedFeedback

한지훈 ○ 박상근
경희대학교 소프트웨어융합학과
jhhan0208@khu.ac.kr, sk.park@khu.ac.kr

A Service for Providing Refined Feedback in Online Comments: CleansedFeedback

Jihoon Han○ Sangkeun Park
Department of Software Convergence, Kyung Hee University

요 약

온라인 예술 콘텐츠 플랫폼의 댓글 공간은 감상 공유의 수단이자 창작자에게 중요한 피드백 수단이 된다. 그러나 악성 댓글은 창작자와 사용자 모두에게 심리적 피해를 초래하며, 건전한 소통을 방해하는 요소로 작용하고 있다. 본 연구에서는 KcELECTRA 기반 모델을 활용하여 댓글을 네 가지 유형(중립, 순수 피드백, 순수 비난, 비난 섞인 피드백)으로 분류하고, 피드백은 살리고 비난은 제거 또는 순화하는 웹 기반 정제 시스템을 개발하였다. 피드백 여부에 대한 분류 모델은 LLM 기반 라벨링과 수작업 검증을 거쳐 학습하였으며, 웹툰 플랫폼 실사용 시나리오에 적용 가능한 형태로 구현하였다.

1. 서 론

최근 영화, 드라마, 웹툰 등 다양한 콘텐츠에 대해 사람들이 댓글을 통해 자유롭게 의견을 공유하는 문화가 확산되었다. 그러나 이 댓글들 중 악성 댓글은 공격적인 표현, 욕설, 비난 등으로 구성되어 창작자와 다른 이용자에게 정신적 피해를 주고 있다. 이에 따라 온라인 플랫폼에서는 욕설 필터링, 댓글 제한 등의 방식으로 대응하고 있지만, 이러한 방법은 건설적인 피드백까지 제거할 위험이 있으며, 창작자가 필요한 피드백을 놓치는 문제가 발생한다.

본 연구는 '비난과 피드백이 혼재된 댓글'에 주목하여, 비난은 순화하고 피드백은 유지하는 방향으로 정제된 댓글을 사용자에게 제공하는 CleansedFeedback 시스템을 제안한다.

2. 관련 연구

악성 댓글 문제를 해결하기 위해, 비난 표현을 분류/제거하는 연구들이 존재한다. 대표적으로 박채원 et al.[1]은 192,158개 한국어 뉴스 댓글에 대해 3단계 혐오 레이블을 적용한 대규모 Corpus 'K-HATERS'를 구축하고, BERT 기반 모델을 통해 정치·종교 등 특정 타겟에 대한 혐오 표현을 효과적으로 분류·제거했다. 그러나 이러한 방식은 비난 섞인 피드백 댓글의 경우 피드백까지 제거된다는 한계가 있다.

문제 해결을 위한 또다른 방법으로 비난 표현을 순화하는 연구들도 존재한다. 이주필 et al.[2]는 해당 연구는 비난 표현을 완곡한 표현으로 바꾸는 자연어 생성 접근 방식으로 순화된 댓글을 생성하는 모델을 제안했다. 하지만 비난만 존재하는 댓글의 경우에는 순화를 하더라도 여전히 부정적인 내용만 담고 있다는 한계가 있다.

기존 연구와 비교	비난 분류/제거 연구	비난 순화 연구	본 프로젝트
중립 댓글 Ex) "재밌게 보고 있어요"		원본 제공	
순수 피드백 댓글 Ex) "전개 속도가 좀 빨라요"		원본 제공	
순수 비난 댓글 Ex) "XX 재미없네"	제거	순화 의미 X	제거
비난 섞인 피드백 댓글 Ex) "맞춤법 검사기도 안돌리나"	피드백도 사라짐	순화해서 제공	순화해서 제공

그림 1. 본 프로젝트와 기존 연구 비교

본 프로젝트는 댓글을 4가지 유형으로 분류하고, 상황에 따라 다르게 처리한다 [그림 1]. 특히 피드백 유무를 추가로 고려하여, 정말 비난만 있는 경우에는 기존의 비난 분류/제거 연구처럼 제거한다. 하지만 비난이 존재하는 경우에도 피드백이 포함되어 있다면 비난은 순화하고, 피드백은 살려서 제공한다. 이를 통해 비난으로 인한 마음의 상처는 최소화하면서, 피드백은 받아들여 더 좋은 작품을 만드는 데에 기여하는 것이 본 프로젝트의 목표이다.

3. 댓글 분류를 위한 모델 구축

3.1 비난 여부 분류 모델

비난 여부 판단에는 공개된 한국어 혐오 표현 분류 모델인 beomi/korean-hatespeech-classifier [1] 를 사용하였다. 해당 모델은 한국어 댓글에 대해 좋은 성능을 보이는 KcELECTRA[2] 모델 기반으로 Fine-tuning된 모델이며, 여러 댓글에 대해서 비난 표현이 존재하는지 여부를 세가지 라벨(None, Hate, Offensive)로 분류한다.

[1] https://huggingface.co/beomi/korean-hatespeech-classifier
[2] https://github.com/Beomi/KcELECTRA

3.2 피드백 여부 분류 모델

피드백 여부 분류는 공개된 기존의 데이터나 모델이 존재하지 않아, 자체적인 데이터 수집, 라벨링, 모델 학습이 필요하였다. 먼저 다양한 평점대의 네이버 웹툰들의 댓글들을 수집하였으며, 수집 댓글들에 대해 피드백 여부 라벨링을 진행하였다.

1. 피드백 여부를 판단하는 주요 기준은 "문제점 제기", "보완 방법 제안" 등을 통해 개선에 도움이 될 수 있는 내용이면 피드백, 그렇지 않으면 피드백이 아니다.
   - "문제점 제기": EX) "아우 캐릭터들 죄다 말 디게 많네.." -> 0
   - "보완 방법 제안": EX) "네비게이터들의 워프장치 정도만 소개하고 다른 과정은 적절히 스킵하는 것이 좋아보입니다." -> 0
   - "단순 비난" EX) "이따구로 그려도 작가소리 듣고 돈받는구나ㅋㅋㅋ" -> 1
   - "단순 감상" EX) "인간이 됐을까??" -> 1
2. 피드백이 "얼마나 피드백인지"와는 관계 없이, 조금이라도 1의 요소가 있으면 피드백 O(0)으로 라벨링, 아니라면 피드백 X(1)로 라벨링한다.
3. 구체적인 내용이 없는 텍스트의 경우 기본적으로 피드백 X(1)로 라벨링한다.
   - EX) "12345", "ㅋㅋㅋㅋ", "??" 등
4. 피드백 여부는 비난 여부와는 관계가 없다.
   - EX) "진짜 맞춤법 검사기도 안 돌리는건가 ㅋㅋ": 비난이 포함되어 있음에도 맞춤법이 맞지 않다는 피드백이 존재 -> 0
   - EX) "응원합니다"는 비난은 없지만 피드백이 존재하지 않음 -> 1

그림 2. LLM 라벨링 프롬프트

라벨링에는 LLM(GPT-4o)를 사용하였으며, [그림 2]의 프롬프트를 바탕으로 진행하였다. 라벨링 결과 피드백인 댓글 수 1427개, 피드백이 아닌 댓글 수 4404개로 클래스 불균형 문제가 발생하였다. 불균형 문제를 해결하기 위해 피드백이 아닌 댓글 중 2000개를 downsampling 하여 총 3427개의 데이터를 확보하였다.

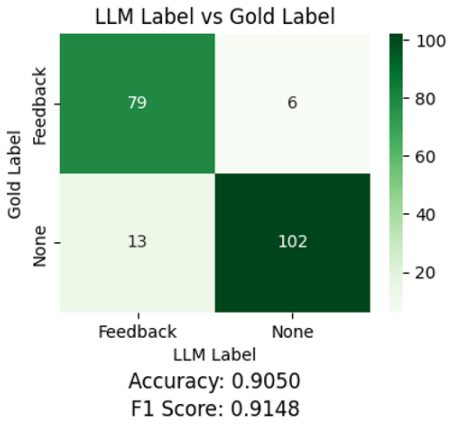

그림 3. LLM 라벨링 성능 평가

LLM 라벨링 성능을 평가하기 위해 라벨링된 데이터 중무작위 200개 샘플을 추출하여 사람(본인과 동료)이 직접 라벨링을 진행하였다. 본인과 동료의 라벨링 중 불일치한 라벨에 대해서는 상의 후 Gold Label을 생성하였으며, Gold Label과 LLM 라벨을 비교하여[그림 3] LLM 라벨링의 신뢰도를 평가하였다. 평가 결과 Accuracy는 약 0.9, F1-Score은 약 0.91임을 확인할 수 있었다.

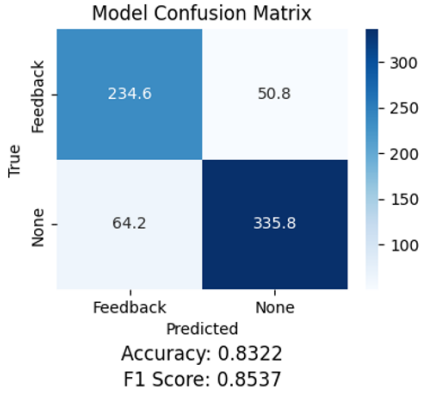

그림 4. 피드백 분류 모델 성능 평가

라벨링된 수집 데이터로 KcELECTRA 모델에 대한 Fine-tuning을 진행하였고, 사용을 위해 Huggingface에 업로드하였다[3]. 모델 평가를 위해 80%의 Train set와 20%의 Test set로 5-Fold Cross Validation을 진행하였다[그림 4]. 평가 결과, 5-Fold에 대한 평균 F1 Score는 약 0.85, Accuracy는 약 0.83으로 준수한 성능을 보였다.

4. 웹서비스 개발

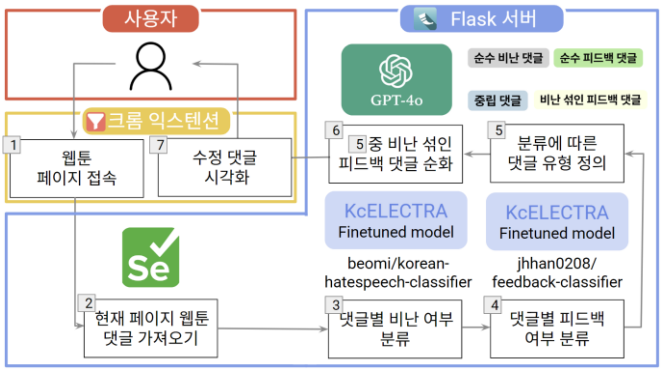

그림 5. CleansedFeedback 아키텍처

본 연구에서는 비난 여부 분류 모델과 위에서 학습한 피드백 여부 분류 모델을 활용해서, 사용자들이 보는 네이버 웹툰 댓글에 대해 정제된 피드백을 보여주는 웹서비스 CleansedFeedback [4] 를 개발했다. CleansedFeedback의 아키텍처는 [그림 5]와 같다.

[3] https://huggingface.co/jhhan0208/feedback-classification-kcelectra-v1
[4] https://www.youtube.com/watch?v=EaX9uLS-Ggk ((Demo Video)

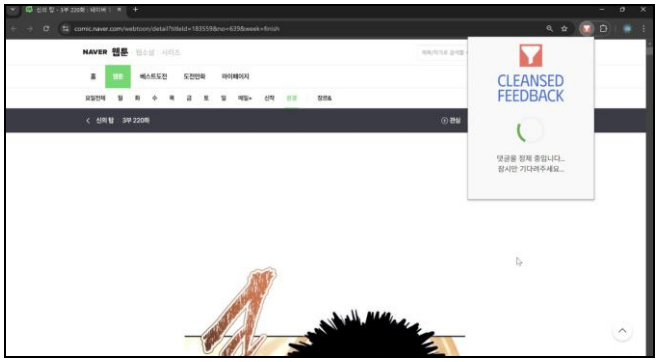

그림 6. CleansedFeedback 동작 과정

본 서비스는 ON 상태에서 사용자가 웹툰 댓글 페이지에 접속하면 자동으로 현재 페이지의 URL이 Flask 서버로 전달되어 분류/정제가 시작된다 [그림 6].

그림 7. 댓글 분류/순화 결과

분류/정제가 완료된 댓글들은 웹페이지상 기존 댓글들을 덮어쓰는 방식으로 사용자에게 표시된다[그림 7]. 이 중 중립 댓글과 순수 피드백 댓글은 원문을 그대로 유지하여 보여주는 것을 볼 수 있다. 순수 비난 댓글은 숨김 처리를 하여 사용자가 볼 수 없도록 하며, 비난 섞인 피드백 댓글은 LLM을 통해 비난 표현을 제거한 후 피드백만 남긴 순화 댓글로 대체하여 사용자에게 보여준다.

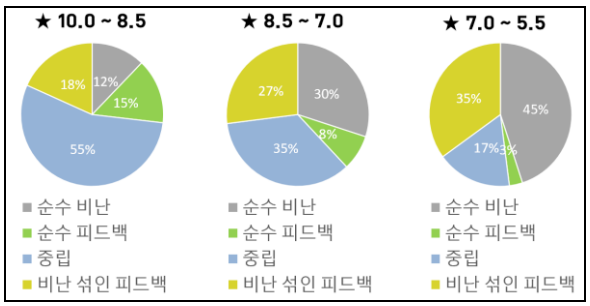

그림 8. 평점대별 댓글 유형 분포

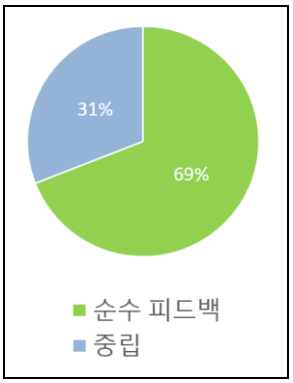

그림 9. CleansedFeedback 적용 시 댓글 유형 분포 예시

평점대별 댓글 유형 분포[그림 8]를 봤을 때, 평점이 낮은 웹툰일수록 순수 피드백의 비중은 줄고, 비난 섞인 피드백의 비중은 늘어나 해당 서비스의 효용성이 더 높아진다고 볼 수 있다. CleansedFeedback 적용 시 순수 비난은 제거하고, 비난 섞인 피드백은 순수 피드백으로 변환하여[그림 9] 더 건전하면서 도움이 되는 환경을 조성하는 데에 도움이 될 것이다.

5. 결론

본 연구에서는 단순한 혐오 표현 제거를 넘어, 피드백의 의미와 가치를 보존하면서 부정적 표현을 제거하는 서비스 CleansedFeedback를 구현하였다. 특히 비난과 피드백이 혼재된 댓글에 대한 차별화된 처리 방식을 통해 창작자와 이용자 모두에게 유익한 피드백 환경을 조성할 수 있다.

향후 네이버 웹툰 이외에도 다양한 플랫폼으로의 확장 가능성이 있으며, 피드백 정제 품질 향상을 위한 모델 보완 및 사용자 맞춤형 정제 강도 조절 기능 도입 등을 통해 활용성을 더욱 확장할 수 있을 것으로 기대된다.

6. 참고 문헌

[1] 박채원, et al. "K-HATERS: A Hate Speech Detection Corpus in Korean with Target-Specific Ratings", arXiv.2310.15439, 2023

[2] 이주필, et al. "거대언어모델과 설명가능 인공지능을 활용한 한국어 유해 댓글 순화 연구", 한국정보과학회 학술발표논문집, 2024