KcELECTRA 모델을 활용한 악성 댓글 탐지 및 선한 댓글 우선 정렬

이준행, 오현영, 박상우, 강영명*
성결대학교 컴퓨터공학과
e-mail : {hkhk9495, ohy414, sgpswoo, ykang}@sungkyul.ac.kr

Malicious Comment Detection and Positive Comment Prioritization Using the KcELECTRA Model

Junhaeng Lee, HyunYoung Oh, SangWoo Park, Young-myoung Kang*
Dept of Computer Engineering, Sungkyul Univ.

요 약

부정적인 표현을 선별하는 데 초점을 맞춘 기존 악성 댓글 탐지 기법과 달리 본 연구에서는 선한 댓글을 식별하여 정렬이 가능하도록 KcELECTRA 를 파인튜닝하여 다중 클래스 분류 모델을 구축하였다. UnSmile 데이터셋, YouTube 댓글 데이터셋을 대상으로 실험한 결과 91.05%의 높은 정확도를 기록하며 악성 댓글과 선한 댓글을 효과적으로 구별할 수 있음을 보였다. 제안하는 모델을 실제 온라인 커뮤니티 환경에 적용하면 악성 댓글 노출로 인한 부정적 효과를 최소화하고 선한 댓글을 우선적으로 선택할 수 있는 기회를 제공하여 건강한 댓글 문화형성에 크게 기여할 것으로 기대된다.

1. 서론

온라인 커뮤니케이션은 사용자가 쉽게 접할 수 있는 다양한 플랫폼을 통해 이루어지며 특히 댓글 기능은 이용자 간 의견 교환과 토론을 촉진하는 중요한 역할을 한다[1]. 그러나 댓글을 통한 자유로운 의견 교환이 활성화될수록 악성 댓글 문제도 점점 심화되고 있으며 이는 단순한 언어폭력을 넘어 심각한 정신적 피해를 초래할 수 있다. 악성 댓글이 노출될 경우 이는 단순한 개인 간의 문제를 넘어 사회 전체에 부정적인 영향을 미칠 수 있다[2]. 혐오 콘텐츠에 반복적으로 노출될 경우 심리적 불안과 우울감을 유발할 뿐만 아니라 집단 간 불신을 조장하고 사회적 편견을 강화하며 나아가 폭력적인 행동을 부추길 가능성이 높아진다[3]. 또한 익명성이 보장되는 온라인 환경에서는 공격적인 댓글이 용인되는 분위기가 형성될 가능성이 크며 이는 모방 심리를 자극해 악성 댓글의 확산을 더욱 가속화한다[4].

악성 댓글 문제를 해결하기 위해 주요 온라인 플랫폼들은 AI 기반 악성 댓글 탐지 모델을 도입하여 유해 댓글을 자동으로 감지하여 삭제하고 있다. 그러나 악성 댓글 작성 방식이 점차 교묘해짐에 따라 AI 기반 필터링 기술은 이에 대한 효과적인 대응을 하지 못하고 있다. 따라서 악성 댓글 문제를 해결하기 위해서는 AI 를 활용한 탐지 기술의 고도화 뿐만 아니라 탐지되지 않은 악성 댓글의 노출을 최소화할 수 있는 추가적인 대응책이 필요하다.

기존의 악성 댓글 탐지 기술은 주로 부정적인 표현에만 초점을 맞추어 왔으나 악성 댓글뿐만 아니라 선한 댓글을 인식할 수 있는 새로운 댓글 분류 모델을 통해 악성 댓글 노출로 인한 부작용을 최소화하고 댓글에 대한 사용자 선택권을 확장할 필요가 있다.

이러한 요구를 반영하여 본 연구에서는 KcELECTRA[5] 모델을 다중 클래스 분류 모델로 파인튜닝(fine-tuning) 하여 악성 댓글뿐만 아니라 선한 댓글까지 탐지할 수 있는 방안을 제안한다. 또한 선한 댓글을 우선적으로 정렬하는 새로운 필터링 기법을 개발하여 악성 댓글 탐지 실패 시에도 사용자에게 긍정적인 댓글을 우선적으로 노출하여 부정적 효과를 최소화하고 사용자의 댓글 정렬 선택권을 부여하였다.

본 논문에서 제안하는 KcELECTRA 기반 댓글 분류 모델은 UnSmile 데이터셋[6], YouTube 댓글 데이터셋을 대상으로 실험한 결과 91.05%의 높은 정확도를 기록하며 악성 댓글과 선한 댓글을 효과적으로 구별할 수 있음을 보였다. 특히 부정(Negative) 및 선한(Positive) 댓글의 분류 성능은 정확도(Accuracy) 91.05%, 정밀도(Precision) 0.9091, 재현율(Recall) 0.9113, F1-score 0.9099 를 보여주었다. 이러한 우수한 검증결과를 바탕으로 본 연구는 실제 온라인 커뮤니티 환경에서 악성 댓글 노출을 최소화하고 긍정적인 소통을 활성화하는 데 크게 기여할 것으로 기대된다.

2. 관련연구

최근 자연어 처리(Natural Language Processing, NLP) 분야에서는 한국어 특화 사전학습 모델들이 활발히 연구되고 있다. 대표적인 모델로는 KoBERT[7], KoELECTRA[8], KcBERT[9], KcELECTRA 등이 있으며 각 모델의 성능을 비교하는 다양한 연구가 진행되었다.

KcELECTRA 와 GPT-3.5 의 성능을 비교한 결과 숫자 및 특수문자 등을 조합하여 변형된 혐오 표현에 대해 KcELECTRA 가 GPT-3.5 보다 더 높은 탐지 성능을 보였다[10]. 또한 KcBERT, KoELECTRA 등 여러 한국어 모델을 비교한 연구에서도 KcELECTRA 가 정확도, F1 score, 정밀도, 재현율 등 모든 평가척도에서 다른 모델들보다 우수한 성능을 기록하였다[11].

앞서 소개한 기존 연구들은 주로 혐오 표현 댓글 탐지에 초점을 맞추었으나 선한 댓글에 대한 심층적인 분석 및 활용 방안에 대한 연구는 제한적이었다. 본 연구에서는 이러한 선행 연구의 한계를 보완하기위해 KcELECTRA 를 활용하여 비정형 데이터를 분석하고 악성 댓글 탐지 뿐만 아니라 선한 댓글도 탐지하는 다중 클래스 분류 모델을 제안한다.

3. 본론

3.1 모델 선정

기존 BERT[12] 기반 모델은 Masked Language Model (MLM) 방식으로 학습된다. 입력 문장에서 15%의 토큰을 [Mask]로 대체한 후 이를 원래 값으로 예측하는 방식으로 학습된다. 그러나 이 방식은 Masking 된 일부 토큰만 학습에 활용되므로 모델이 문맥을 충분히 학습하려면 대량의 데이터가 필요하다.

ELECTRA[13],[14]는 이를 보완하기 위해 Replaced Token Detection (RTD) 방식을 활용한다. 그림 1 에서 확인할 수 있듯이 ELECTRA 는 Generator 와 Discriminator 두 개의 모듈로 구성된다. Generator 는 Masking 된 입력 문장으로부터 원래의 토큰을 예측하며 그 확률 분포는 (1)과 같이 정의된다.

$$P_G(x_t \mid x) = \frac{\exp(e(x_t)^\top h_G(x)_t)}{\Sigma_{x'} \exp\left(e(x')^\top h_G(x)_t\right)} \quad (1)$$

$e_{(x_t)}$는 정답 토큰 x_t의 임베딩 벡터이며 $h_G(x)_t$는 Generator 가 출력한 해당 위치의 문맥 벡터를 의미한다. 이후 Softmax 연산을 통해 확률을 계산하며 주어진 문맥에서 각 단어가 등장할 확률을 구한다. Discriminator 는 입력 문장에서 특정 토큰이 원본인지 또는 Generator 가 생성한 것을 판별하는 역할을 수행한다.

$$D(x,t) = \text{sigmoid}(w^\top h_D(x)_t) \quad (2)$$

w는 학습 가능한 가중치 벡터이며 $h_D(x)_t$는 문맥 정보를 포함한 벡터이다. 시그모이드(sigmoid) 연산을 통해 이진 분류를 수행한다.

최종 손실 함수는 Generator 의 Masked Language Model 손실(L_{MLM})과 Discriminator 의 Replaced Token Detection 손실(λL_{Disc})의 가중 합으로 정의되며 다음과 같이 표현한다.

$$\min_{\theta_G,\theta_D} \Sigma_{x\in X} \mathcal{L}_{\mathcal{MLM}}(x,\theta_G) + \lambda \mathcal{L}_{\mathcal{D}isc}(x,\theta_D) \quad (3)$$

λ는 두 손실 간의 균형을 조정하는 하이퍼파라미터로, 일반적으로 Discriminator 의 학습을 강화하기 위해 $\lambda > 1$ 로 설정한다. 예를 들면, 그림 1 에서 “the”가 [Mask]로 대체된 후 Generator 가 다시 “the”로 복원하면 “original”로 판별한다. 반면, cooked”가 [Mask]로 대체된 후 Generator 가 “ate”로 복원하면, Discriminator 는 이를 “replaced”로 판별하는 방식이다. 이러한 RTD 방식은 문장의 모든 토큰을 학습 과정에 활용할 수 있어 높은 효율성과 성능을 확보할 수 있다.

ELECTRA 를 기반으로 확장한 KcELECTRA 모델은 비정형 데이터를 활용하여 학습된 사전 학습 언어 모델이다. 해당 데이터는 온라인 뉴스 댓글이나 대댓글에서 정제되지 않은 형태로 수집되어 구어체 표현이 많고 다양한 신조어가 포함되어 있어 실제 온라인 환경에서 사용되는 한국어 텍스트를 효과적으로 처리할 수 있도록 학습되었다.

본 연구에서는 한글 문장 처리에 탁월한 KcELECTRA 를 파인 튜닝하여 악성 댓글, 일반 댓글, 선한 댓글을 분류하는 다중 클래스 분류 모델을 구축한다.

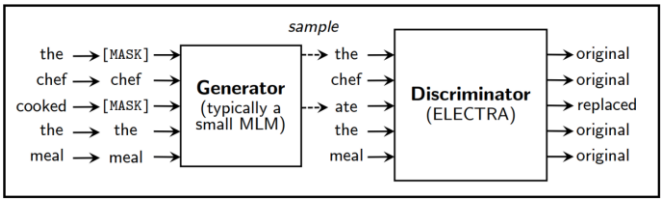

그림 1. ELECTRA 모델의 구조

3.2 데이터셋 수집

3.2.1 UnSmile 데이터셋

Unsmile 데이터셋은 Smilegate AI 에서 공개하는 한국어 혐오 표현 데이터셋으로 특정 사회적 집단에 대한 적대적 발언, 조롱, 희화화, 편견을 재생산하는 표현을 혐오 표현으로 정의한다. 기존의 데이터셋은 혐오 표현을 다양한 세부 카테고리로 분류하고 있으나 본 연구에서는 10,139 개의 혐오표현 중 2,472 개의 데이터를 기존 레이블 비율을 유지한 상태에서 무작위로 선정하여 활용하였다. 선택된 데이터는 하나의 부정적인 댓글 클래스로 통합하였다.

3.2.2 Youtube 댓글 데이터셋

일반 댓글과 선한 댓글은 직접 YouTube 에서 크롤링 하여 수집하였으며 최근 업로드 된 영상의 댓글을 대상으로 하였다. 데이터의 다양성을 확보하기 위해 뉴스, 예능, 스포츠 등 다양한 카테고리의 영상을 포

함하였으며 총 14,772 개의 댓글을 수집하였다. 수집한 댓글 데이터에서 이모티콘, HTML 태그 및 불필요한 공백을 제거하는 전처리를 수행하여, 한국어 외의 다른 언어로 작성된 댓글은 제외하였다. 라벨링은 [15]를 참고하였으며 기준은 다음과 같다. 중립적이거나 정보 전달 목적의 댓글은 일반 댓글로, 특정 이슈에 대한 찬반 여부와 상관없이 논리적이고 설득력 있는 의견을 제시하거나 긍정적 희망적 메시지를 담은 격려·사랑의 댓글은 선한 댓글로 구분하였다. 총 14,772 개의 댓글에서 2,043 개의 선한 댓글을 분류하였으며 분류가 어려운 중립적인 댓글을 제외한 후 선한 댓글의 비율을 고려하여 2,413 개의 일반 댓글을 선별하였다.

최종적으로 4,456 개의 댓글 데이터를 선정하였으며 이를 UnSmile 데이터셋과 결합하여 악성 댓글(0), 일반 댓글(1), 선한 댓글(2)로 구분된 다중 클래스 분류 문제로 정의하였다.

3.3 모델 학습

본 연구에서는 다중 클래스 분류 모델을 구축하기 위해 KcELECTRA 모델을 기반으로 파인튜닝을 수행하였다. 학습에 사용된 모델은 KcELECTRA-base-v2022 로 ElectraForSequenceClassification 을 활용하여 3 개의 클래스(악성 댓글, 일반 댓글, 선한 댓글)를 분류하였다.

학습에 사용된 하이퍼파라미터는 배치 크기 16, 학습률 2e-5, 에포크(epoch) 수 5 로 설정되었으며, AdamW 최적화 알고리즘과 EarlyStoppingCallback 을 적용하여 과적합을 방지하였다.

데이터셋은 선한 댓글 2,043 개, 일반 댓글 2,413 개, 악성 댓글 2,472 개로 구성되었으며 이를 8:2 비율로 훈련 데이터와 검증 데이터로 나누었다. 훈련 데이터는 5,542 개, 검증 데이터는 1,386 개로 구성되었으며 모델 학습은 Google Colab 환경에서 NVIDIA Tesla T4 GPU 를 활용하여 수행되었다.

3.4 성능 평가

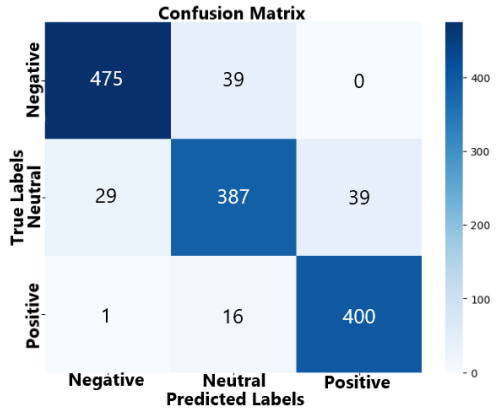

그림 2. 혼동행렬 시각화 결과

표 1. 모델 예측과 실제 라벨 및 소프트맥스 확률

-	댓글	실제	예측	소프트맥스 확률		
				악성 댓글	일반 댓글	선한 댓글
탐지성공	젠신병자는 뒤져야된다 이기야	0	0	0.993	0.005	0.001
	요즘은 석유난로 거의 안쓰니 다들 환기시켜야되는거 생각도 못했나보네	1	1	0.002	0.028	0.969
	앞으로도 맛있는요리 꿀팁많이가르쳐주십시요 감사합니다	2	2	0.002	0.023	0.974
탐지실패	아니 얼마나 살이 쪘으면 자리 돼지냐	0	1	0.375	0.616	0.007
	울 신랑이 보배회원이라 가끔 곁다리로 보는데 아재들 화력 짱	1	0	0.833	0.160	0.006
	천사분들이 여기 계셨네요	2	1	0.016	0.570	0.412

실험 결과 제안된 다중 클래스 분류 모델은 91.05%의 정확도를 기록하였으며, 정밀도(Precision) 0.9091, 재현율(Recall) 0.9113, F1-score 0.9099 를 달성하여 그 효과를 입증하였다.

그림 2 는 실험에 대한 혼동행렬 결과를 보여준다. 부정적인 댓글과 선한 댓글의 경우 높은 정밀도와 재현율을 유지하며 효과적인 분류가 이루어졌음을 알 수 있다. 그러나 중립(neutral) 댓글에서 일부 오분류가 발생하였으며 부정적 표현이 포함된 중립 댓글이 부정으로, 선한 표현이 포함된 중립 댓글이 긍정으로 분류되는 경향이 관찰되었다. 예를 들어, 표 1 에서 볼 수 있듯이 "얼마나 살이 쪘으면 저리 돼지냐"와 같은 문장은 명확한 비난의 의미를 가지지만 모델이 이를 일반 댓글(중립)로 예측하는 오류가 발생하였다. 이는 모델이 특정 단어(예: "돼지")를 단순한 명사로 해석하고 문맥 속의 조롱적 의미를 충분히 반영하지 못한 결과로 분석된다. 일반 댓글과 선한 댓글 간의 일부 혼동은 향후 추가적인 데이터 보강 및 정제 과정을 통해 개선할 것이다.

3.5 선한 댓글 정렬 기법

본 연구에서는 기존의 최신순 및 인기 댓글순 방식에 추가하여 악성 댓글의 노출을 방지할 수 있는 선한 댓글순 정렬 방식을 제안한다. 기존의 댓글 정렬 방식인 인기 댓글순과 최신순에서는 모델이 악성 댓글을 정확히 탐지하지 못할 경우 그림 3 과 같이 악성 댓글이 사용자에게 여과 없이 노출되는 문제가 발생할 수 있다. 반면, 본 연구에서 제안하는 선한 댓글순 방식을 선택하면 그림 4 와 같이 모델이 탐지한 선한 댓글만을 사용자에게 제공하여 AI 탐지를 우회한 악성 댓글의 노출을 효과적으로 차단할 수 있다. 이러한 방식은 건강한 온라인 댓글 문화를 촉진하는 데 기여할 것으로 기대된다.

그림 3. 기존 댓글 정렬 방식

그림 4. 선한 댓글순 정렬 방식

4. 결론

본 연구에서는 악성 댓글 필터링과 선한 댓글 우선 정렬 기능을 제안하였다. UnSmile 데이터셋과 YouTube 댓글 데이터셋을 활용하여 KcELECTRA 모델을 파인튜닝 하였으며 악성 댓글, 일반 댓글, 선한 댓글을 효과적으로 분류하는 다중 클래스 분류 모델을 구현하였다. 실험결과 중립적인 표현이면서도 문맥에 따라 부정적으로 해석될 가능성이 있는 댓글이 일부 오분류되는 경우가 발생한 것을 제외하면 제안하는 모델은 높은 분류 성능을 보여주었고 실질적인 악성 댓글 탐지 및 예방에 도움이 될 것으로 기대한다.

향후 연구에서는 특정 표현이 문맥에 따라 조롱 또는 일반적 의미로 해석되는 경우를 정밀하게 분류할 수 있도록 다양한 맥락 정보를 포함한 방대한 분량의 데이터셋을 활용하여 모델의 문맥 이해 능력 향상 및 정렬 시스템을 고도화 할 예정이다.

참고문헌

[1] 한국지능정보사회진흥원, 2024 년 인터넷 이용 실태조사 요약보고서, 2024, https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=99870&bcIdx=27858&parentSeq=27858.

[2] 한국정보화진흥원, 2023 년 사이버 폭력 실태조사, 2024. https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=68302&bcIdx=26483

[3] P. Madriaza, G. Hassan, S. Brouillette-Alarie, A. N. Mounchingam, L. Durocher-Corfa, E. Borokhovski, D. Pickup, and S. Paillé, “Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities,” *Campbell Systematic Reviews*, Vol.17, No.4, p. e1178, 2021. **https://doi.org/10.1002/cl2.70018**

[4] L. Rösner and N. C. Krämer, “Verbal venting in the social media age: Effects of anonymity and group norms on aggressive language use in online comments,” *Social Media + Society*, vol. 6, no. 1, pp. 1–12, 2020. [온라인]. Available: https://doi.org/10.1177/2056305116664

[5] Beomi, "KcELECTRA: Korean comments ELECTRA," GitHub, 2022. [Online]. Available: https://github.com/Beomi/KcELECTRA

[6] Smilegate AI, "Korean UnSmile Dataset," GitHub, 2021. [Online]. Available: https://github.com/smilegate-ai/korean_unsmile_dataset

[7] SKTBrain, "KoBERT: BERT for Korean," GitHub repository, 2019. [Online]. Available: https://github.com/SKTBrain/KoBERT

[8] monoloqq, "KoELECTRA: Korean ELECTRA Model," GitHub repository, 2020. [Online]. Available: https://github.com/monologg/KoELECTRA

[9] Beomi, "KcBERT: Korean Comments BERT," GitHub repository, 2020. [Online]. Available: https://github.com/Beomi/KcBERT

[10] 우지은 외, “KcELECTRA 와 GPT-3.5 의 한국어 혐오 표현 탐지 성능 비교 분석,” 한국컴퓨터정보학회 학술대회논문집, 2023. [온라인]. 이용 가능: https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11705519

[11] 이영준 외, “KcELECTRA 를 활용한 혐오성 댓글 분류,” 한국컴퓨터정보학회 학술대회논문집, 2023. [온라인]. 이용 가능: https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11705144

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint, arXiv:1810.04805, 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[13] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, 2020. [Online]. Available: https://arxiv.org/abs/2003.10555

[14] Google Research, "ELECTRA," GitHub repository, 2020. [Online].Available: https://github.com/google-research/electra

[15] 정영라, "인터넷 뉴스 댓글이 원문과 태도에 미치는 영향: 정치적·윤리적 요인의 선플, 악플 유형을 중심으로," RISS, pp. 21-54, Available: https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=DIKO0013540063#;