KoELECTRA 를 활용한 실시간 온라인 성희롱 댓글 검열 시스템

김민환 O, 박상근
경희대학교 소프트웨어융합학과
minhwan0514@khu.ac.kr, sk.park@khu.ac.kr

Real-time Online Sexual Harassment Comments Filtering Using KoELECTRA

Minhwan Kim O, Sangkeun Park
Department of Software Convergence, Kyung Hee University

요 약

댓글을 통한 온라인 성희롱 문제가 심각한 사회적 이슈로 대두되고 있다. 기존 연구들은 성희롱 표현을 일반 혐오 표현의 일부로 다루어 왔으나, 성희롱 특유의 표현 방식을 효과적으로 탐지하는 데 한계가 있었다. 본 연구는 암묵적 성희롱 문장까지 탐지하는 모델을 개발했으며, 이를 활용해 유튜브에서 성희롱 댓글을 검열할 수 있는 크롬 플러그인을 개발하여 온라인 환경에서 성희롱 문제를 적극적으로 대응할 수 있는 가능성을 제시한다.

(주의) 해당 논문에는 선정적인 표현이 포함되어 있습니다

1. 서 론 *

온라인 시장이 활성화되고 유튜브가 대중화되면서, 많은 사용자들이 댓글을 통해 활발히 소통하고 있다. 온라인 댓글은 게시자와 시청자 간 소통의 창구로서 긍정적인 효과를 가져오는 반면, 동시에 악성 댓글로 인한 부정적 피해를 초래하기도 한다. 특히, 온라인 공간에서 성적(Sexual) 발언을 통해 특정인을 모욕하거나 희롱하는 문제가 심각한 사회적 이슈로 대두되고 있다. 특히 온라인 성희롱 문제는 주로 여아이돌이나 여성 방송인 등 여성을 대상으로 발생하며[1], 통신매체이용음란행위, 사이버 명예훼손 및 모욕에 관한 피해로 이어지기도 한다[2].

유튜브를 비롯한 온라인 영상 스트리밍 서비스에서는 성희롱을 포함한 악성 댓글을 차단하기 위해 특정 키워드 기반의 필터링, 사용자 신고를 통한 수동 차단 등 다양한 방법을 시도하고 있다. 그러나 직접적인 성적 단어가 아닌 용어, 또는 줄임말이나 초성 변환(예: '검정 스타킹' → '검스', '가슴' → 'ㄱㅅ')을 활용한 악성 성적 댓글은 기존 댓글 차단 메커니즘으로는 즉각적으로 탐지하고 차단하기 어렵다는 한계를 가진다.

성희롱 표현이 포함된 혐오 표현, 악성 댓글, 온라인 괴롭힘(Cyberbullying) 탐지를 위한 다양한 연구들이 수행되어 왔다[3, 4, 5, 6, 7]. 그러나 기존 혐오 표현 탐지 모델은 일반적인 혐오 단어(예: '꺼져라', '한남')의 특성을 잘 파악하는 반면, 성희롱 표현과 관련된 미세한 어휘적 맥락(예: '검스', ''맛있다')은 반영하지 못한다는 한계가 있다. 이러한 어휘적 맥락을 반영하여 소셜 미디어의 댓글에서 성희롱 표현을 탐지한 연구도 있으나[8, 9], 댓글 자체만으로는 직접적으로 성적 용어를 사용하지 않은 성희롱 표현까지 정확하게 탐지하는 데 한계가 존재한다.

본 연구에서는 온라인 영상 스트리밍 서비스에 게시되는 한국어 성희롱 댓글을 효과적으로 검열하기 위한 새로운 방법을 제안한다. KoELECTRA 모델[10]을 활용하여 유튜브 내 성희롱 댓글을 분류하며, 짧은 문장 내 암묵적인 성희롱 표현을 보다 정확히 탐지하기 위해 영상 제목 정보를 추가로 고려한다. 제안한 방법의 실용성을 검증하기 위해 성희롱 댓글 검열 기능을 갖춘 크롬 플러그인을 개발했다.

2. 관련 연구

2.1. 혐오 표현 탐지 연구

자연어처리를 이용해 혐오 표현 또는 온라인 괴롭힘을 탐지한 다양한 연구가 존재한다. 온라인에서 발생하는 인종, 지역, 성별 등의 혐오 표현 문장으로부터 그 유형을 분류하거나[3, 4], 감정 분석, LLM 을 활용한 성차별 발언의 유형을 분류한 연구 등이 있다[5, 6].

성희롱 표현은 기존 혐오 표현에 함께 포함되어 탐지 연구가 수행되었으나, 성희롱 고유의 맥락을 파악하는 데 한계를 보인다. Park et al. [7]은 감정 분석을 활용한 혐오 표현 문장 예측에서, ("여고생이 맛있나요, 여대생이 맛있나요?")의 False Negative 사례를 보이며 혐오 표현 탐지 연구로부터

* "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2025 년도 SW 중심대학사업의 결과로 수행되었음"(2023-0-00042)

성희롱 문장 분류의 한계점을 지적했다. 이러한 한계를 극복하기 위해, 포괄적인 혐오 표현이 아닌 성희롱 표현에 집중한 연구[8, 9]가 있으나, 댓글 자체만을 고려했기 때문에, 짧은 문장(예: “와, (저 여자) 우유 맛있겠다”)에서 그 성적 의미까지 탐지하기는 힘들다는 한계가 있다.

2.2. 혐오 표현 필터링 기법

온라인에서 실시간으로 혐오 표현을 필터링하는 다양한 기법이 연구되었다. 악성 댓글 또는 혐오 표현 문장으로부터 의심 단어를 강조하여 표시하거나[11], 대응 단어를 생성하여 추천하는 방식[12]이 있다. 또한, 작성된 문장을 분석하여 혐오 표현의 유형 및 위험률을 제시[13]하거나, 혐오 표현 문장 자체를 특정 문구로 대체하는 방법[14]이 제안되기도 하였다. 본 연구는 이를 활용하여 유튜브 영상 시청 시 영상 하단의 보이는 성희롱 댓글을 실시간으로 필터링할 수 있는 크롬 플러그인을 제안한다.

3. 성희롱 문장 탐지

3.1 데이터 수집

성희롱 문장 데이터를 수집하기 위해 유튜브 내 ‘여성 BJ 댄스’ 및 ‘아이돌 직캠’ 키워드를 검색해서 나오는 111 개의 영상에서 2,035 개의 데이터를 수집했다. 실제 수집한 데이터의 평균 댓글 길이는 23.99 자로 짧고, 이로부터 성희롱 문장의 맥락을 파악하기 어렵다는 특성을 갖는다. 따라서 본 연구는 유튜브 영상 댓글과 함께 제목 정보를 고려하여 성희롱 문장의 맥락을 파악하고자 하였다. 예를 들어 “섹시 댄스, 코카인 교차 편집”이라는 영상 제목으로부터 “와, 우유 맛있겠다”, “ㄱㅅ ㅈㄴ 크다”, “들박 마렵다”라는 댓글을 암묵적 성희롱 문장으로 유추할 수 있고, 이로부터 성희롱 댓글에 관한 최적화된 분류를 기대할 수 있다.

3.2 성희롱 댓글 데이터 구축 및 라벨링

수집된 댓글이 성희롱 관련 댓글인지 여부를 판단하기 위한 라벨링 작업은 수작업으로 진행하였다. 라벨링 기준은 ‘성폭력범죄의 처벌 등에 관한 특례법 제 13 조(통신매체를 이용한 음란행위)를 참고하여 판단하였다. 특히, 한국의 문화적 맥락을 고려하여 ‘은유적 또는 비유적 표현을 사용해 여성의 신체나 성적 행위를 묘사하며, 영상 속 특정 인물을 성적으로 괴롭히거나 다른 시청자에게 불쾌감을 조성하는 문장’을 성희롱 댓글로 판정하고, 이외의 댓글은 일반 문장으로 구분하였다. 라벨링 작업 결과, 성희롱 문장은 426 개, 일반 문장은 1,609 개로 분류되었다.

3.3 성희롱 문장 탐지 모델 개발

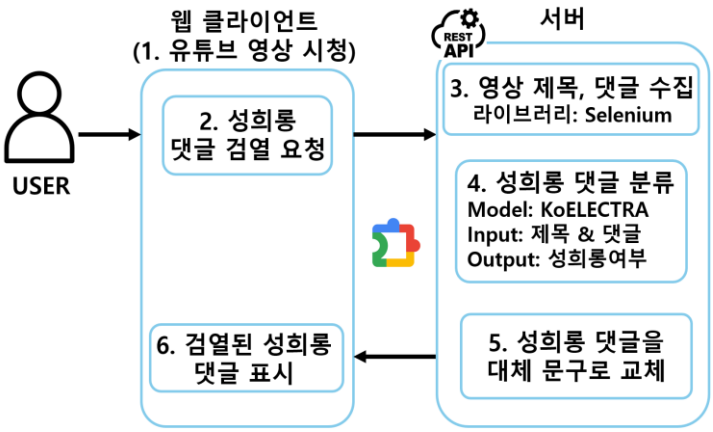
성희롱 문장 분류를 위해 대용량 한국어 데이터셋으로 사전학습된 KoELECTRA[10] 모델을 사용하였다. KoELECTRA 모델에 수집한 데이터셋을 전이학습하여 성희롱 문장 탐지 모델을 개발하였다. ELECTRA[15] 모델은 Transformer 기반의 생성기와 판별기로 구성된다. 생성기를 통해 일부 단어를 가짜 단어로 바꾸고 판별기가 문장으로부터 단어의 가짜 여부를 판단함으로써 문장 전체의 맥락을 학습한다. 이를 통해 "저 여자 맛있게 생겼다"나 ‘ㄱㅅ’과 같이 비유적 문장 또는 초성 변환을 통한 암묵적 성희롱 표현의 맥락을 효과적으로 학습할 수 있다.

또한, SentencePiece Tokenizer 를 사용하여 문장을 서브워드 단위로 토큰화를 수행하여 어휘적 맥락을 파악한다. 예를 들어 '검스' 같은 줄임말을 '검', '스'로 분리하여 토큰화를 진행한다. 기존 사전학습된 모델은 ‘검’을 ‘Sword’로 해석할 여지가 있으나, 전이학습을 통해 성희롱 문장에서 반복적으로 나타나는 ‘검’을 성적인 맥락으로 새로 파악한다. 이를 통해 성적인 의미가 내포된 줄임말의 어휘적 맥락을 효과적으로 파악할 수 있다. 수집된 데이터셋은 학습(67%), 검증(22%), 테스트(11%)로 분할하였다. 학습 파라미터는 Epoch 20, Batch Size 8 로 설정했다. Accuracy, Precision, Recall, F1 score 를 통해 성능 평가를 수행하였고, [표 1]과 같이 높은 성능으로 성희롱 표현 분류를 수행한 것을 확인할 수 있다.

[표 1] 성희롱 문장 인식 모델 성능 평가

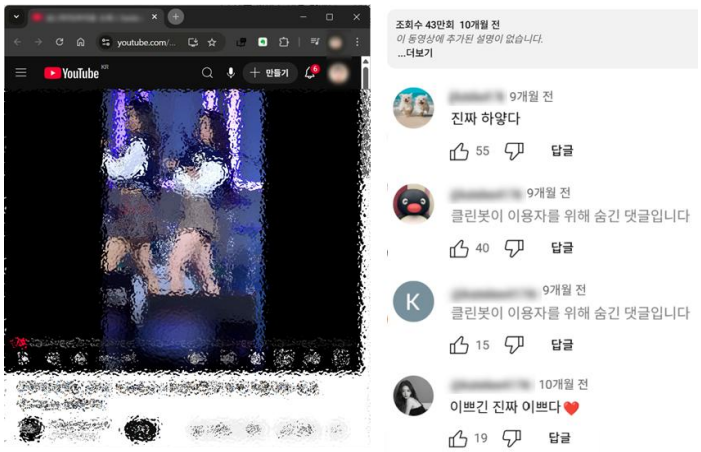
Accuracy	Precision	Recall	F1 score
0.895	0.740	0.771	0.755

3.4 성희롱 댓글 검열 크롬 플러그인 개발



[그림 1] 시스템 아키텍처

모델의 실용성을 검증하기 위해 유튜브 영상 시청 시 보이는 성희롱 댓글을 실시간으로 검열하는 크롬 플러그인을 개발했다. 전체 시스템 아키텍처는 [그림 1]과 같다. 크롬 플러그인을 실행할 경우, 사용자가 유튜브 영상 클릭 시 자동으로 영상 제목 및 댓글을 수집한다. 이후 수집한 데이터를 JSON 형식으로 웹 서버에 전송한 후 모델을 이용해 성희롱 댓글을 감지한다. 마지막으로, 인식된 성희롱 문장 댓글을 특정 문구로 대체한다.



[그림 2] 성희롱 댓글 필터링 크롬 플러그인 동작 화면

본 크롬 플러그인이 성희롱을 인식하고 해당 댓글을 필터링 메시지로 대체한 결과는 [그림 2]와 같다. '하 금딸 3 일찬데 ㅅㅂ 허벅지 진짜 **야 [1]''와 '탱탱허이 좋구나'라는 성희롱 댓글이 있었으나, 본 서비스가 이를 감지하고 '클린봇이 이용자를 위해 숨긴 댓글입니다'라는 문구로 대체하였다.

4. 결 론

본 연구는 유튜브 대중화 및 온라인 커뮤니티 활성화에 따른 성희롱 문제를 해소하기 위해 실시간 온라인 성희롱 댓글 검열 서비스를 제안했다. 이를 위해, 유튜브의 영상 제목 및 성희롱 댓글을 수집하였으며 KoELECTRA 모델로 전이학습을 수행하였다. 성희롱 문장 분류 모델에 대한 성능 실험 결과, 적은 크기의 데이터셋으로도 높은 분류 성능을 기록하였다. 특히 비유적인 성적 표현, 줄임말, 초성 변환 등이 포함된 암묵적 성희롱 문장의 맥락을 높은 인식률로 분류할 수 있음을 보였다. 이 모델을 활용한 성희롱 댓글 검열 크롬 플러그인을 개발하고 그 활용 가능성을 확인했다.

본 연구에 관한 한계점 및 향후 연구는 다음과 같다. 1) 유튜브 플랫폼에서의 암묵적인 성희롱 문장의 맥락을 파악하기 위해 영상 제목 정보를 활용했으나, 추후 영상 속 신체 노출 빈도, 해시태그 정보 등 영상의 실제 내용 정보를 추가적으로 활용할 수 있다. 2) 실제 배포 시 신고 기능을 추가하여 실사용자를 통한 추가적인 데이터셋을 수집하고 자동으로 모델을 재학습하는 파이프라인을 구축하여 성능을 개선할 예정이다.

5. 참고 문헌

[1] 이주원. "스토킹·성희롱·악플…사선으로 떠밀리는 BJ 들." *서울신문*, 2021.10.06.

[2] 송고. "중고생 45% "온라인 성희롱 경험"…외모평가·음란물도 만연" *연합뉴스*, 2019.09.27.

[3] Shin et al. "Hate Speech Detection in Chatbot Data Using KoELECTRA." *Annual Conference on Human and Language Technology*, Human and Language Technology, 518-523, 2021.

[4] Karatsalos and Panagiotakis "Attention-Based Method for Categorizing Different Types of Online Harassment Language." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 321-330, 2019.

[5] Alawneh et al. "Sentiment Analysis-Based Sexual Harassment Detection Using Machine Learning Techniques." *International Symposium on Electronics and Smart Devices (ISESD)*, IEEE, 2021.

[6] Yan and Luo. "BERT-Based Detection of Sexual Harassment in Dialogues." *Proceedings of the International Conference on Computer Science and Artificial Intelligence*, 359-364, 2021.

[7] Park et al. "Why Do I Feel Offended? Korean Dataset for Offensive Language Identification." *Findings of the Association for Computational Linguistics: EACL*, 1142-1153, 2023.

[8] Basu et al. "CyberPolice: Classification of Cyber Sexual Harassment." *Progress in Artificial Intelligence: EPIA Conference Proceedings*, 701-714, 2021.

[9] Islam et al. "Sexual Harassment Detection Using Machine Learning and Deep Learning Techniques for Bangla Text." *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, 2023.

[10] Park. "KoELECTRA: Pretrained ELECTRA Model for Korean." *Github Repository*, https://github.com/monologg/KoELECTRA, 2020.

[11] Pavlopoulos et al. "Deeper Attention to Abusive User Content Moderation." *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[12] Zhu and Bhat. "Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech." *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics (ACL), 134-149, 2021.

[13] 전계원 and 이상원. "악성 비율과 키워드에 근거한 악성 댓글 완화 설명 구조 제안." *한국 HCI 학회 학술대회*, 371-377, 2022.

[14] 이재은 et al. "자연어처리 기술을 활용한 유튜브 악성 댓글 자동 블라인드 시스템." *한국 HCI 학회 학술대회*, 696-699, 2022.

[15] Clark et al. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.1055,* 2020.

1 **은 특정 연예인을 지칭하여 자체 필터링함