

창작자 보호를 위한 SNS 정서 리스크 지표화: KcBERT-GPT 파이프라인과 Shannon 다양성 기반 평가

임예은¹, 이주현¹, 김정훈¹, 성유빈¹, 송미화²

¹한성대학교 컴퓨터공학부 학부연수생

²한성대학교 컴퓨터공학부 교수

2271514@hansung.ac.kr, 2271183@hansung.ac.kr, Bandlagom0927@hansung.ac.kr,
mimolulu@hansung.ac.kr, mhsong@hansung.ac.kr

Operationalizing Affective Risk Metrics for Creator Protection on Social Platforms: A KcBERT - GPT Pipeline with Shannon Diversity

Ye-Eun Im¹, Ju-Hyun Lee¹, Jeong-Hun Kim¹

Yu-Bin Seong¹, Mi-Hwa Song²,

¹Research Student, Division of Computer Engineering, Hansung University

²Professor, Division of Computer Engineering, Hansung University

요 약

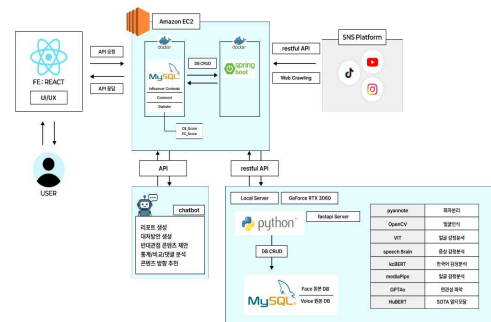
본 연구는 인플루언서를 둘러싼 정서적 위험 요소인 사이버불링 위험과 에코체임버 정도의 위험을 정량화하여 진단하는 멀티모달 분석 프레임워크 SCOPE-X를 제안한다. SNS 콘텐츠의 음성, 얼굴, 텍스트 데이터를 기반으로 감정 흐름을 분석하고, 사용자 댓글의 감정 및 주제를 GPT-4 및 KcBERT 기반으로 분류한 뒤, 사용자 클러스터링을 통해 공격성·편향성 수준을 수치화한다. 최종적으로 콘텐츠-댓글 간 정서적 정합성과 다양성 지표를 바탕으로 CB Score, EC Score를 산출한다. SCOPE-X는 인플루언서의 정서적 리스크를 조기에 식별하고, 개인 맞춤형 정서 회복 전략 수립에 기여할 수 있다.

1. 서론

SNS의 확산과 함께 인플루언서는 여론 형성의 중심으로 부상했지만, 악성 댓글과 편향된 반응 구조로 인해 정서적 위험에 노출되고 있다. 사이버불링은 반복적이고 조직화된 공격으로 진화하고 있으며, 에코체임버는 이질적 의견을 배제하고 감정 편향을 강화시킨다. 이를 해결하기 위해 멀티모달 감정 분석 기술이 주목받고 있으며, 최근 연구로는 Cross-Modal Transformer 기반 구조에 Neural Architecture Search를 적용한 융합 프레임워크가 제안된 바 있다[1]. 제안한 프레임워크는 음성, 얼굴, 자막 간 감정 정보를 효율적으로 통합하지만, 댓글과 같은 외부 반응 분석은 포함하지 않는다. 본 연구는 인플루언서 콘텐츠의 멀티모달 정서 흐름과 댓글 기반 감정·의견 편향성을 통합 분석하여, 사이버불링 (Cyberbullying, CB) 위험도와 에코체임버 (Echo Chamber, EC) 정도를 정량화하는 프레임워크 SCOPE-X를 제안한다.

2. SCOPE-X: 정서 리스크 진단 프레임워크

그림 1은 SCOPE-X 프레임워크의 전체 시스템 구조를 나타낸다. 본 시스템은 SNS 플랫폼에서 인플루언서 콘텐츠를 수집하고, 멀티모달 정서 분석을 수행한 후, 댓글 기반의 사이버불링 위험도 및 에코체임버 정도를 산출한다.



(그림 1) SCOPE-X Framework

분석 결과는 챗봇 인터페이스를 통해 리포트로 제공되며, 사용자에게 대처 방안 및 콘텐츠 전략을 제안할 수 있도록 설계되어 있다. 본 연구는 선행 연구를 확장한 것이다. 이전 연구에서는 KoBERT를 이용해 댓글을 감정 범주로 분류하고 GPT 기반 주제 분류 및 클러스터링을 수행하였으나, 본 연구에서는 KcBERT[3]를 적용하여 약 23만 건의 유튜브 댓글을 7가지 감정으로 분류하였다. 이후 GPT-4 기반 주제 분류와 사용자 간 의견 유사도 분석을 통해 댓글을 다수의 클러스터로 군집화하고, 각 클러스터의 감정 및 주제 분포를 분석하였다. 이를 기반으로 공격적 클러스터의 비율과 분노·혐오 감정의 밀도, 비속어 사용 빈도 등을 종합하여 CB Score를 산출하였다. CB Score는 인플루언서가 직면한 정서적 위험을 수치화하는 지표로, 그림 2에 계산 구조를 보인다.

$$CB_Score = 0.2 \times Fear + 0.2 \times Anger + 0.2 \times Disgust + 0.1 \times Issue/Controversy + 0.3 \times Aggressive Cluster$$

(그림 2) CB Score 산출구조도

한편, EC Score는 인플루언서 영상 댓글에 나타난 감정, 주제, 사용자 클러스터의 Shannon 다양성 지수를 기반으로, 여론 편향의 위험을 평가하는 정량 지표이다. 본 연구에서는 감정, 주제, 사용자 군집 각각의 다양성을 계산하고, 이를 0.4 : 0.2 : 0.4의 가중 평균으로 통합하여 최종 EC Score를 산출하였다[4]. 감정이 한쪽으로 치우치고 특정 사용자 군이 지배적일수록 EC Score는 낮아지며, 반대 의견이 차단된 Echo Chamber 형성이 우려된다. 반대로, 다양한 감정·주제·의견이 고르게 분포된 경우 EC Score는 높아지며, 이는 건강한 커뮤니케이션 구조를 의미한다. 해당 구조는 그림 3에 도식화하였다.

$$EC_Score = 1 - \left(0.4 \cdot \frac{-\sum p_i \ln p_i}{1.9459} + 0.2 \cdot \frac{-\sum p_j \ln p_j}{2.3026} + 0.4 \cdot \frac{-\sum p_k \ln p_k}{1.9459} \right)$$

(그림 3) EC Score 산출 구조도

3. 실험 결과

본 연구는 실제 유튜브 콘텐츠를 대상으로 SCOPE-X를 적용하여 사이버 정서 리스크를 평가하였다. 아래 표는 5가지 영상에 대해 산출된 CB Score와 EC Score를 나타낸다.

<표 1> 영상별 CB/EC Score 결과 비교표

영상 ID	CB Score	EC Score	진단 결과
V1D001	82.3	33.1	위험
V1D002	21.7	78.9	안정
V1D003	64.5	41.0	경계
V1D004	47.8	59.3	준 위험
V1D005	35.6	70.1	비교적 안정

표 1은 5개의 인플루언서 영상에 대한 CB Score와 EC Score를 기반으로 한 진단 결과를 나타낸다. V1D001은 CB Score가 82.3으로 매우 높고, EC Score는 33.1로 낮아, 공격적 감정이 집중되고 여론 다양성이 부족한 고위험 콘텐츠로 분류되었다. 반면 V1D002는 CB Score가 21.7로 매우 낮고, EC Score는 78.9로 높아, 정서적으로 안정되고 다양한 의견이 활발히 형성되는 콘텐츠였다. V1D003은 CB와 EC 점수가 모두 중간 수준으로 나타나 경계 영역으로 판단되었으며, V1D004는 상대적으로 높은 CB Score(47.8)와 보통 수준의 EC Score(59.3)를 보이며 잠재적 정서 위험이 존재하는 ‘준위험’ 콘텐츠로 분류되었다. 마지막으로 V1D005는 CB Score가 다소 낮고 EC Score가 높아, 비교적 안전한 정서 반응이 형성된 것으로 분석되었다.

그림 4는 SCOPE-X 시스템이 생성한 CB/EC 리스크 리포트의 예시이다. 리포트는 챗봇 기반 인터페이스를 통해 자동 생성되며, 콘텐츠별 사이버불링 점수, 에코체임버 점수, 감정 클러스터 분석 결과, 요약 진단문 및 대응 전략을 포함한다.



(그림 4) SCOPE-X 기반 CB/EC 리스크 리포트 생성

예를 들어, 해당 콘텐츠는 CB Score가 12, EC Score가 37.88로 산출되었으며, 이는 상대적으로 높은 사이버 정서 리스크를 의미한다. 특히 CB Score는 공격성 표현, 분노·혐오 감정의 밀도, 클러스터 집중도 등을 기반으로 도출된 수치이며, EC Score는 감정·주제·사용자 분포의 다양성을 종합적으로 반영한다. 사용자는 리포트를 통해 콘텐츠의 정서적 위험 수준을 직관적으로 파악하고, 적절한 대응 방안을 수립할 수 있다.

4. 결론

본 연구에서는 SNS 창작자 콘텐츠에 대한 사이버 정서 리스크를 정량적으로 진단하기 위한 멀티모달 분석 프레임워크 SCOPE-X를 제안하였다. 본 시스템은 영상의 음성, 얼굴, 자막으로부터 감정 흐름을 분석하고, 댓글의 감정·주제·사용자 클러스터링 결과를 기반으로 사이버불링 위험도와 에코체임버 정도를 산출한다. 실험 결과, 해당 지표들은 인플루언서를 둘러싼 정서적 위험 신호를 조기에 감지하고, 콘텐츠의 리스크 수준을 효과적으로 구분하는 데 유효함을 입증하였다. 향후 SCOPE-X는 실시간 리스크 모니터링, 브랜드-인플루언서 매칭, 커뮤니티 갈등 예측의 응용으로 확장 가능하다. 특히 SNS 플랫폼과 통합될 경우, 디지털 공간의 감정 생태계를 관리하는 공공적 정서 인프라로 발전할 수 있을 것으로 기대된다.

참고문헌

- [1] Jeong-Hun Kim, Mi-Hwa Song, "MRO: Multimodal Routing Optimization via Neural Architecture Search of Fusion Paths," *International Journal of Internet, Broadcasting and Communication*, Vol. 17, No. 3, pp. 144-154, 2025
- [2] Jeonghun Kim, Seok Kee Lee, Mi-Hwa Song, "SCOPE: A Transformer-Based System for Predicting Influencer Popularity via Comment Sentiment and Topic Analysis," *Proceedings of the 1st International Conference on Artificial Intelligence Computing and Systems (AICompS 2024)*, Jeju, Korea, Dec. 16-18, 2024, pp. 9-13.
- [3] 이준범 외, "KcBERT: 한국어 댓글 데이터를 활용한 BERT 기반 감정 분류 모델" 한글 및 한국어 정보처리 학술대회 논문집, 2020.
- [4] C. R. Sunstein, *#Republic: Divided Democracy in the Age of Social Media*, Princeton, NJ, Princeton University Press, 2017.