감정 분석을 활용한 개인 맞춤형 AI 클린봇

이우열[1], 송희준[2], 장한빈[1], 서승현[3]
[1]한양대학교 ERICA 전자공학부 학부생
[2]한양대학교 ERICA 건설환경공학과 학부생
[3]한양대학교 ERICA 전자공학부 지도교수
wooimpact@naver.com, huijunsong@naver.com, roni053@naver.com, seosh77@hanyang.ac.kr

A Sentiment Analysis-Driven Personalized AI Cleaning Bot

Woo-Yeol Lee[1], Hui-Jun Song[2], Han-Been Jang[1], Seung-Hyun Seo[3]
[1]School of Electrical Engineering, Hanyang University ERICA
[2]Dept. of Civil and Environmental Engineering, Hanyang University ERICA
[3]School of Electrical Engineering, Hanyang University ERICA

요 약

본 논문에서는 기존 클린봇의 한계를 극복하고자, 시선추적 및 얼굴 분석 기반 감정 인식 기술을 활용한 개인 맞춤형 AI 클린봇을 제안하였다. 사용자의 불쾌 반응을 사전에 감지하고, 이를 바탕으로 맞춤형 필터링이 가능한 데이터셋을 구성하여 웹 상 혐오 표현을 효과적으로 차단하였다. 제안한 시스템은 OpenCV와 dlib 기반의 시선 추적, DeepFace 기반의 감정 분석, 맞춤형 NLP 모델 학습 및 웹 크롤링 기반 필터링 모듈로 구성되었다. 제안한 시스템은 개인의 민감성에 기반하여 정밀한 차단이 가능하며, 향후 크리에이터 중심 플랫폼에도 효과적으로 적용될 수 있다.

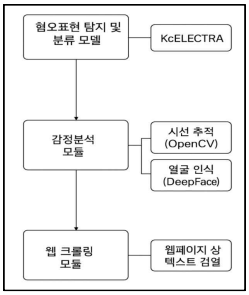
1. 서론

최근 인터넷 환경에서 사용자가 마주하는 댓글 및 텍스트 콘텐츠의 양은 급격히 증가하고 있으며, 이 중 일부 콘텐츠는 혐오 표현이나 공격적인 어휘를 포함하여 사용자의 정신적 스트레스와 불쾌감을 유발하였다. 그러나 기존의 댓글 필터링 시스템은 사전 정의된 비속어나 키워드를 기반으로 콘텐츠를 일괄 차단하는 단순 규칙 기반 방식을 사용하여, 사용자 개인의 정서적 반응이나 콘텐츠의 맥락적 특성을 반영하지 못하였다는 한계가 존재하였다. [1]

따라서 본 연구에서는 시선 추적 기술과 감정 분석 모델을 결합하여 사용자의 감정 상태 및 개인적 특성을 분석하고, 이를 바탕으로 혐오 표현을 맞춤형으로 필터링하는 개인화된 AI 클린봇 시스템을 제안하였다.

2. 감정분석을 활용한 개인 맞춤형 AI 클린봇

본 논문에서 제안하는 AI 클린봇 시스템은 혐오 표현 탐지 및 분류 모델, 감정 분석 모듈, 웹크롤링 모듈로 구성되며, 사용자의 감정 민감도에 기반한 혐오 표현 필터링을 목표로 한다. 초기 단계에서 대중적으로 민감하게 반응하는 키워드를 포함한 문장을 사용자에게 제시하고, 시선 데이터 및 표정 변화를 수집하여 사용자 개인의 불쾌 반응 민감도를 측정하였다. 수집된 민감도 데이터는 혐오 표현 탐지 모델의 가중치로 반영되어 사용자 맞춤형 모델을 생성하였다. 이후 실시간 웹 크롤링을 통해 댓글 및 게시글을 수집하고, 수집한 데이터를 사용자 맞춤형 모델에 입력하여 사용자가 해당 데이터를 얼마나 불쾌하게 받아들일지를 예측하였다. 이후 모델에서 예측한 결과를 기반으로 웹 콘텐츠에 적용하여 필터링하였다. (그림 1)은 시스템 구성을 나타내는 블록 다이어그램이다.



(그림 1) 시스템 구성 블록 다이어그램

2.1. 혐오 표현 탐지 및 분류 모델

본 연구에서는 KcELECTRA 모델을 기반으로 혐오 표현을 탐지하였다. KcELECTRA는

ELECTRA 아키텍처를 기반으로 한국어 뉴스와 댓글 데이터로 학습된 모델로, 한국어의 특성과 온라인 표현의 노이즈를 효과적으로 반영하였다. 학습에는 Selectstar 혐오 표현 데이터셋을 사용하였으며, 문장마다 혐오 표현 여부와 심각도를 라벨링하여 다단계 필터링이 가능하도록 구성하였다. [2][3]

2.2. 감정 분석 모듈

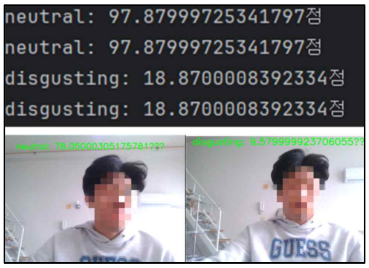
OpenCV와 dlib을 이용해 사용자의 얼굴 및 눈 영역을 실시간으로 추출하고, DeepFace를 기반으로 한 얼굴 분석 모델을 활용하여, 추가로 감정 분류기를 적용하여 행복, 분노, 혐오, 슬픔 등 주요 감정 상태를 분석하였다. 특히 특정 문장 제시 시, 주시 시간 감소, 눈 깜빡임 빈도 증가, 부정적 표정 변화 등의 지표를 종합하여 불쾌감 수치를 정량화하고, 이를 사용자 개인의 민감도 설정 데이터로 저장하였다. 이후 해당 과정에서 수집한 사용자의 민감도 프로파일을 혐오 표현 탐지 모델에 반영하고, 특정 키워드나 표현에 대해 가중치를 조정하여 맞춤형 필터링을 구현하였다. [4]

2.3. 웹 크롤링 모듈

TamperMonkey 확장 프로그램을 활용하여 웹페이지 상의 댓글과 게시글을 문장 단위로 실시간 수집하였다. 수집된 텍스트는 혐오 표현 탐지 모델에 입력되어, 각 문장에 대해 사용자 민감도 기반 필터링 여부를 판단하였다. 필터링이 필요한 문장은 블라인드 처리하거나 대체 문구로 변환하여 사용자에게 노출을 차단하였다. 이는 사용자 경험을 저해하지 않도록 설계되었다.

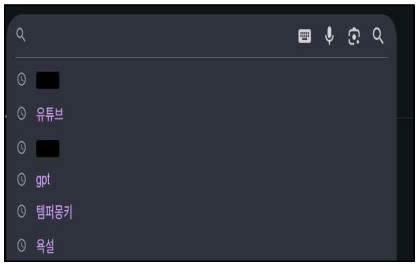
3. 실험 결과

(그림 2)은 감정 분석 모듈이 작동하는 화면을 나타낸 것이다.



(그림 2) 감정 분석 모듈 작동 예시

(그림 3)은 웹 크롤링 모듈이 웹페이지에서 텍스트를 블라인드 처리하는 과정을 보여준다.



(그림 3) 웹 크롤링 모듈 블라인드 처리 예시

4. 결론 및 기대 효과

본 시스템은 기존의 일괄적이고 획일적인 클린봇 필터링 방식을 탈피하여, 사용자 개인의 감정 반응에 따라 동적으로 반응하는 차세대 콘텐츠 필터링 방식의 가능성을 제시하였다. 사용자는 자신의 감정 반응을 기반으로 한 데이터셋을 지속적으로 개선해 나가며, 정밀한 혐오 표현 필터링이 가능해진다. 이로 인해 단순한 욕설 차단을 넘어, 사용자 본인이 민감하게 반응하는 주제나 표현에 대해 사전에 방어할 수 있어 보다 쾌적한 인터넷 사용 환경을 조성할 수 있다.

또한 본 시스템은 일반 사용자뿐만 아니라 유튜버, 인터넷 방송인, SNS 인플루언서와 같은 크리에이터에게도 적용 가능성이 높다. 이들이 이용하는 플랫폼의 댓글이나 채팅창에 본 시스템을 적용하면, 각자의 성향에 따라 악성 표현을 미리 걸러낼 수 있어 정신적 피로도를 줄이고 창작 활동에 더욱 집중할 수 있는 환경을 제공한다. 향후에는 텍스트뿐만 아니라 이미지 기반 혐오 콘텐츠로 확장하거나, 성능이 낮은 모바일 기기나 노트북 환경에서도 동작할 수 있도록 최적화하여, 전반적인 온라인 이용자층에게 효과적인 혐오 표현 차단 솔루션으로 발전시킬 수 있을 것이다.

참고문헌

[1] 욕설 막아도 혐오표현 못 걸러내... 결국 댓글창 폐지뿐?, 파이낸셜뉴스
https://www.fnnews.com/news/202307040602175282
[2] https://huggingface.co/beomi/KcELECTRA-base-v2022
[3] https://github.com/Beom0i/KcELECTRA
[4] Joseph Bardeen, et al. (2017). “An Eye-Tracking Examination of Emotion Regulation...”