이신행 중앙대 교수 "온라인상 혐오 발언, 맥락 데이터 기반 AI로 잡을 수 있다"

김동원 기자 | 승인 2021.04.23 18:22

현재 기술로는 혐오 발언 방지 어려워
중의적 표현이나 뉘앙스 잡지 못해
네이버 클린봇, 여성 혐오 댓글 65%밖에 못 잡아내
"문장 맥락으로 혐오 발언 잡아내는 기술 개발 중"

(사진=셔터스톡)

온라인상에서 문제 되는 혐오 발언을 방지하기 위해선 맥락 데이터 기반 인공지능(AI) 기술이 개발돼야 한다는 주장이 제기됐다.

이신행 중앙대 미디어커뮤니케이션학부 교수는 23일 진행된 '제18회 인공지능인문학 국내학술대회'에서 "혐오 발언 댓글 등을 AI가 차단하는 데 기존 기술로는 한계가 있다"면서 "맥락 데이터를 이용해야 혐오 발언을 정확하게 방지할 수 있다"고 밝혔다.

그는 악성 댓글을 탐지하는 '네이버 클린봇'을 예로 들며, 온라인에 있는 AI 기술이 혐오 발언을 정확하게 탐지하지 못 한다고 설명했다.

클린봇은 네이버가 악성 댓글을 탐지하기 위해 2019년 11월 12일부터 모든 뉴스에 적용한 AI 시스템이다. 욕설과 저속한 표현, 선정적 표현, 폭력적 표현, 차별적 표현, 비하적 표현 등 약 35만여 건의 데이터를 분석·학습해 악성 댓글을 탐지한다. 클린봇에 의해 탐지된 악성 댓글은 사용자 선택에 따라 숨길 수 있다.

이 교수 연구팀은 악성 댓글 탐지에 뛰어난 성능을 보이는 클린봇이 혐오 표현 탐지에도 적용할 수 있는지를 시험했다. 댓글에 도출된 맥락 기반 의미 정보로 클린봇이 탐지한 악플을 예측하는 지도 기계학습(SYM)을 실시했다. 대상은 사회적 소수 집단에 대한 혐오가 조직적으로 표출되는 여성과 성 소수자, 외국인으로 했다.

그 결과 클린봇은 성 소수자 혐오 댓글을 86% 정확도로 잡아낸 반면, 여성혐오 댓글은 65% 정확도로 잡아냈다. 흔히 기준점으로 잡는 70%보다 낮은 정확도였다.

또 네이버 클린봇은 커뮤니티 이름이나 벌레, 조선족, 여포 등 다른 의미가 포함돼있는 단어는 혐오 발언으로 잡아내지 못했다.

이 교수는 "성 소수자 악성 댓글은 혐오 발언이 두드러져 높은 정확도로 잡아낼 수 있었지만, 여성의 경우 혐오 발언이 두드러지지 않아 낮은 정확도를 기록했다"고 설명했다. 이어 "현재 사용하고 있는 악성 댓글 탐지 기능은 클린봇처럼 혐오 발언을 정확히 찾아내기 힘들다"고 덧붙였다.

이신행 중앙대 교수는 "혐오 발언 댓글은 맥락 데이터 기반 AI가 정확하게 잡아낼 수 있다"고 설명했다.

현재 온라인상에서 혐오 발언 댓글을 숨기거나 차단하는 기술로는 ▲키워드 기반 분류 ▲인간코더 기반 분류 ▲커뮤니티 기반 분류가 쓰인다.

키워드 기반 분류는 욕설과 모욕적 용어들이 포함된 표현을 탐지한다. 키워드로 정확하게 혐오 표현을 탐지할 수 있지만, 중의적인 표현이나 뉘앙스를 파악하는데 제한이 있다.

인간코더 기반 분류는 한국어 이해 능력은 물론 혐오 의미를 포착할 수 있는 역량을 갖춘 인간 코더가 분류한 자료로 기계학습 후 혐오 표현을 판별하는 방법이다. 사람이 해석한 혐오 표현을 탐지할 수 있는 장점이 있지만, 비용이 많이 들고 주관적으로 해석할 수 있는 위험이 있다.

커뮤니티 기반 분류는 온라인 커뮤니티로부터 추출된 자료로 기계학습 해 혐오 표현을 탐지하는 기술이다. 새로운 용어와 언어 뉘앙스를 고려해 혐오 표현을 탐지할 수 있지만, 커뮤니티 내에서만 혐오 표현인 말을 일반화할 수 있는 오류를 범할 수 있다.

이 교수는 "혐오 발언을 정확하게 찾기 위해서는 기존 기술과 달리, 문장의 맥락을 이해해야 한다"면서 "이를 위해 연구팀에서는 맥락 데이터 기반 AI 기술을 개발하고 있다"고 밝혔다.

AI타임스 김동원 기자 goodtuna@aitimes.com

[관련기사] 이루다 사태 막으려면...“설명가능한AI·자연어 데이터 연구가 답”

[관련기사] 정부 “올해 AI 윤리 체크리스트 만든다”...제2의 이루다 사건 막을 것

김동원 기자 goodtuna@aitimes.com