인공지능의 언어 이해에 관한 맥락적 분석

「인공지능」과 「과학기술과 윤리」강의페어링

사이버보안학과, 최경주, 201920689, 이진희 교수님 지도

연구 목적

- 혐오 표현을 인공지능으로 생성하는 것은 도덕적으로 올바르지 못한 행위이다. 가령 특정 집단에 대하여 혐오를 일으키는 표현을 생성형 AI로 작성하는 것은 자동화 시스템에 의해 비방하는 글을 대량으로 생산하여 악의적인 활동에 사용될 수 있다. 이러한 혐오 표현은 **단어 중심으로 필터링**하는 것으로 무분별한 혐오를 방지할 수 있지만, 혐오 키워드가 없음에도 혐오적인 의미를 내포하는 간접적인 문장은 탐지하는데 어려울 수 있다.

- 본 연구의 목적은 맥락적인 판단만을 가지고 AI가 혐오 의도를 분류하는 능력을 평가하고 그 유형을 분석하여 생성형 AI의 언어 이해에 대한 한계를 탐색하고자 한다.

연구 방법 및 내용

데이터 수집 및 결과 산출

혐오 의도 분류를 위해 대표 커뮤니티 사이트 5개에 대해 댓글 데이터를 크롤링했으며, 데이터는 혐오 키워드가 포함된 문장과 포함되지 않은 문장을 동일한 비율로 각 사이트 당 600개의 텍스트를 수집했다.

dcinside	fmkorea	ruliweb	네이버카페	오늘의유머
600	600	600	600	600

혐오 키워드를 포함하는 문장의 경우, 인종과 정치, 지역, 젠더, 연령 카테고리에 해당하는 표현들을 기반으로 필터링을 수행했으며 혐오 단어가 없지만, 의도가 존재하는 문장의 경우 Human Labeling을 통해 처리하여 약 600개의 문장들이 존재한다. 따라서 혐오 의도가 있는 문장들은 키워드로 혐오인 900개와 혐오 단어가 없으나 맥락적으로 혐오인 문장 600개로 구성된다.
수집된 문장들 중 혐오 단어가 있으나 맥락적으로 혐오가 아닌 데이터에 대해서는 혐오 표현 중 약 117개가 존재했으며, 그 유형으로는 정의, 반대 의견, 인용 등이 있다. 즉, 혐오 의도가 없는 데이터에는 키워드가 있지만 맥락적으로 혐오가 아닌 117개의 데이터와 혐오 키워드가 없으면서 혐오 의도가 없는 1,383개의 데이터로 구성된다. 위와 같은 전처리를 통해 혐오 의도에 대한 데이터 분포는 다음과 같다.

혐오 의도 O	혐오 의도 X
1,500	1,500

ChatGPT의 혐오 판단 능력을 평가하기 위해 OpenAI의 API를 사용했으며, 결과는 정확도, 정밀도, 재현율, F1-Score로 계산하였다.

평가 유형	비율
정확도	0.86
정밀도	0.382
재현율	0.814
F1-Score	0.698

본 연구에서 주목한 부분은 정밀도와 재현율의 관계로, 낮은 정밀도는 예측한 데이터 중 실제로 양성인 데이터의 비율이 적음을 의미한다. 높은 재현율은 실제로 양성인 데이터 중 모델이 양성으로 올바르게 예측한 비율을 의미한다. 즉, ChatGPT는 혐오로 분류한 데이터 중 실제로 혐오인 데이터의 비율이 적지만, 혐오 의도가 있는 데이터에 대해서는 옳게 판단한 것이다.

혐오 의도가 있는 데이터의 대부분은 키워드가 혐오 표현이기에 재현율이 높은 이유는 키워드를 중심으로 혐오 의도를 분류했기 때문이다. 반면에 ChatGPT가 예측한 혐오 중 실제로 혐오 의도가 있는 데이터가 적은 이유는 키워드가 아닌 맥락적으로 혐오인 데이터를 혐오로 분류하지 못한 것으로 볼 수 있다.

결과 분석

ChatGPT의 혐오 의도 판단 작업에서 정확도를 기준으로 맞추지 못한 데이터는 총 420개이다. 이 데이터의 유형은 다음과 같다.

- Label 1: 혐오 단어가 있으나, AI가 학습하지 않은 문장
- Label 2: 혐오 단어가 있으나, 일상적인 표현인 문장
- Label 3: 혐오 단어가 없으나 맥락적으로 혐오인 문장

Label 1	Label 2	Label 3
120	88	212

Label 1의 경우, 재추론 시 해당 혐오 키워드에 대한 정보를 제공할 경우 올바르게 혐오로 판단할 수 있다. 정보를 제공하지 않고 판단하기 어려운 이유는 새롭게 생겨난 혐오 표현들은 주로 합성어로 표기되며, 어떤 개인이나 집단의 속성과 부정적인 의미를 내포할 수 있는 단어를 결합한다. 가령, 특정 정치인이나 방송인의 이름에 비하적인 표현을 사용하여 소수 집단의 품위 또는 사회적 인식을 깎아내리려고 한다.

Label 2의 경우, 데이터 수집 과정에서 맥락적으로 혐오가 아닌 117개의 데이터 중 88개를 ChatGPT는 예측하지 못한 것이다. 예측에 실패한 사례 중에 대부분은 혐오 감정을 가지고 있는 집단을 지칭하면서 혐오 표현을 자주 언급하지만, 해당 표현을 사용하는데 있어 타인의 생각을 물어보는 아래와 같은 문장이 존재한다.

"저런생각이 잘못된건맞는대 쟈는 XXX 더럽고 악마같고 XX은 천사다 이런이야기하고있잖아요. 본인건강은 왜 생각안하는지 캐나다가 더했으면더했지 XXX소리로밖에 안들리는데요 XX 프레임에 XXX 프레임까지 안보이시나요"

Label 3의 경우, 데이터 수집 과정에서 혐오 키워드가 없으나 맥락적으로 혐오인 600개의 데이터 중 212개는 예측에 실패한 사례들로, 예측에 실패한 420개 중 과반수 이상을 차지하고 있다.
예측에 실패한 사례 중 일부는 혐오 키워드를 언급하지는 않지만, 지칭대명사를 사용함으로써 특정 집단을 간접적으로 표현하는 의도를 가지고 있다.

"XX이 그렇게원하는 XX한테 빌붙어 살던가 ㅋㅋㅋ"

또 다른 경우로는, 두 문장들을 접속사로 연결하면서 앞의 문장이 가진 맥락적 의미를 뒷문장과 동등하게 간주하거나 역으로 보는 것이다.

1. "지금 전세계 어느나라도 난민 안받고있잖아 XXX는 난민과도 같긔"
2. "모든 XXX이 테러리스트가 아니지만 모든 테러리스트는 XXX"

따라서 예측 실패 사례들을 통해 알 수 있는 사실은 생성형 AI가 혐오 의도를 판단할 때 언어 자체를 깊이 이해하지 않고 단어 중심으로만 평가하고 있으며, 한국어가 간접 표현이나 비유, 대조 등의 문장을 다양한 방법으로 나타낼 수 있기에 맥락적인 이해가 부족하다고 볼 수 있다.

결론 및 제안

본 연구는 생성형 AI가 혐오 표현이라는 카테고리에서 언어 이해가 충분하여 맥락적인 분석 능력이 확보되었는지를 평가하였다. 혐오 판단의 정확도는 0.86으로 높은 수치를 보여줬으나, 정밀도 부분에서 낮은 성능은 곧 키워드만으로 혐오를 판단하는 것으로 볼 수 있어 맥락적인 해석이 부족한 것이다. ChatGPT가 예측에 실패한 사례들을 분석하여 절반 이상이 Label3에 해당되며 여러 문장에 걸쳐 혐오 의도를 간접적으로 드러냄으로써 탐지를 회피할 수 있었다.

Label2에 대해서도 117의 문장 중 88개를 혐오 의도가 존재한다는 잘못된 판단을 내린 이유가 혐오 키워드가 있지만, 혐오 표현의 사용을 묻거나 반박, 약한 부정, 혐오 표현을 사용하는 집단에 대한 부정적 의견 제시 등을 맥락적으로 이해하지 못하고 단어 중심으로만 판단했기 때문이라고 볼 수 있다.

생성형 AI를 활용하는 사람은 지속적으로 증가하는 추세인 반면, 언어 이해가 충분하지 못하는 인공지능을 전면적으로 수용하는 것은 잠재적으로 위험하다. 특히 대부분의 생성형 AI는 한국어가 아닌 영어권 데이터를 기반으로 학습된 모델이기에, 한국어 자체에 대한 특성이나 문장을 표현할 수 있는 다양한 방법을 이해하지 못할 수 있다. 이러한 한계점을 바탕으로, 언어 이해 능력을 높이기 위한 추가적인 모델의 학습 과정이 필요하며, 특히 한국어를 사용하는 경우에는 단어의 변형이 무수히 많으므로 인력을 투자하여 한국어 이해를 충분히 수행할 수 있는 생성형 AI 모델을 만드는 노력이 필수적이다.

참고자료

조헌국. (2024). 물리교육에서 생성형 인공지능의 활용과 문제 해결 방안: 거대 언어 모델의 환각 문제를 중심으로. *새물리*, *74*(8), 812-823.
박서윤, 강예지, 강조은, 김유진, 이재원, 정가연, ... & 김한샘. (2024). GPT-4 를 활용한 인간과 인공지능의 한국어 사용 양상 비교 연구. *국어국문학*, (206), 5-47
최유숙. (2024). '혐오표현'의 국어학적 논의 고찰. *어문론집*, *97*, 39-64..