대화형 인공지능을 위한 메신저 대화의 비윤리적 표현 연구

고예린○, 남길임, 송현주
경북대학교
goyelin08@naver.com, nki@knu.ac.kr, songhj@knu.ac.kr

Unethical Expressions in Messenger Talks for Interactive Artificial Intelligence

Yelin Go○, Kilim Nam, Hyunju Song
Kyungpook National University

요 약

본 연구는 대화형 인공지능이 비윤리적 표현을 학습하거나 생성하는 것을 방지하기 위한 기초적 연구로, 메신저 대화에 나타나는 단어 단위, 구 단위 이상의 비윤리적 표현을 수집하고 그 특성을 분석하였다. 비윤리적 표현은 '욕설, 혐오 및 차별 표현, 공격적 표현, 성적 표현'이 해당된다. 메신저 대화에 나타난 비윤리적 표현은 욕설이 가장 많은 비중을 차지했는데, 욕설에서는 비표준형뿐만 아니라 '존-', '미치다' 등과 같이 맥락을 고려하여 판단해야 하는 경우가 있다. 가장 높은 빈도로 나타난 욕설 '존나류, 씨발류, 새끼류'의 타입-토큰 비율(TTR)을 확인한 결과 '새끼류'의 TTR이 가장 높게 나타났다. 다음으로 메신저 대화에서는 공격적 표현이나 성적인 표현에 비해 혐오 및 차별 표현의 비중이 높았는데, '국적/인종'과 '젠더' 관련된 혐오 및 차별 표현이 특히 높게 나타났다. 혐오 및 차별 표현은 단어 단위보다는 구 단위 이상의 표현의 비중이 높았고 문장 단위로 떨어지기 보다는 대화 전체에 걸쳐 나타나는 것을 확인하였다. 따라서 혐오 및 차별 표현을 탐지하기 위해서는 단어 단위보다는 구 단위 이상 표현의 탐지에 대한 필요성이 있음을 확인하였다.

주제어: 메신저 대화, 비윤리적 표현, 사용역, 학습용 데이터

1. 서론

2016년 마이크로소프트사에서 출시한 '테이'와 2020년 한국의 스캐터랩에서 출시한 '이루다'는 모두 딥러닝 방식 기반의 대화형 인공지능 서비스로, 사용자가 많아질수록 더 많은 대화를 학습할 수 있다는 장점이 있다. 하지만 사용자들이 입력한 비윤리적인 표현까지 모두 학습해버린 탓에, 두 서비스 모두 얼마 가지 못하고 서비스를 중단해야 했다. 대화형 인공지능이 사용자와 대화하면서 비윤리적 표현을 학습하는 것을 방지하고, 대화체 학습용 데이터의 비윤리적 표현을 자동적으로 처리하는 등의 기술을 개발하기 위해서는 '대화'라는 사용역을 우선적으로 고려할 필요가 있다. 그중에서도 대화형 인공지능이 대부분 메신저의 형태로 서비스된다는 점에서 본 연구에서는 메신저 대화에 나타난 비윤리적 표현의 유형 및 그 특성을 살펴보고자 한다.

지금까지 비윤리적 표현을 자동적으로 탐지하고자 하는 시도가 계속되어 왔지만 그 대상은 주로 댓글로 한정되어 왔다. 댓글은 수집이 용이하여 다양한 주제, 대상의 데이터를 구축하는 것이 가능한 반면 대화에서 나타나는 비윤리적 표현을 수집하는 것은 개인정보, 저작권 등의 어려움이 따른다. 따라서 본 연구는 2019년 국립국어원에서 구축한 메신저 대화(약 140만 어절)에 나타난 비윤리적 표현을 수집하여 이를 유형화하고 그 특성을 살피는 것을 목적으로 한다.

조태린(2018)에서는 대화형 인공지능이 비윤리적 표현을 생성하지 않도록 하기 위해 비윤리적 표현의 기초적인 연구를 수행하였는데, 사전, 논문, 웹, SNS에서 수집된 단어 단위의 비윤리적 표현을 대상으로 삼았다. 다만, 이는 사용역에 의한 차이를 고려하지 못하였으며, 여전히 단어 단위의 금칙어 목록 기반 접근이라는 점에서 한계가 있다.[1)][1] 본 연구는 실제 메신저 대화를 대상으로 하였다는 점, 단어 단위는 물론이고 구 단위까지 확장하여 정량적 분석을 함께 시도한다는 점에서 이전 연구과 차별성이 있다.

2. 관련 연구

조태린(2018)에서는 비윤리적 표현의 유형에 '욕설, 비어, 속어' 외에 '맥락'과 '내용' 유형을 추가적으로 설정하였다는 점을 주목할 만하다. 맥락 유형은 동물이나 사물이 아닌 사람에게 사용될 때에만 비윤리적인 표현인 것(대가리, 주둥이), 지역이나 상황에 따라 비윤리적 표현인 것(계집, 마누라)이 포함되고 내용 유형에는 불법이나 범죄에 관련된 내용(강간, 로리타)이나 미성년자에게 제한할 필요가 있는 표현(성인용품, 오르가즘), 비윤리적 사건이나 행위를 가리키는 표현(자살, 참수) 등이 있다. 남길

1) 단어 단위의 금칙어 목록을 기반으로 하는 비윤리적 표현에 대한 탐지 방식은 네이버의 클린봇1.0에서 활용되었는데 그 효과가 미미하였다. 이에 네이버에서는 최근 맥락을 고려하여 공격적, 선정적, 폭력적 표현까지 판단할 수 있도록 업데이트된 클린봇2.0을 공개하였다.

임 외(2021)에서는 보다 실용적 관점에서 비윤리적 표현을 '욕설, 혐오 및 차별 표현, 공격적 표현, 성적 표현'으로 구분하였고, '미치다, '존-'과 같은 경우는 맥락을 고려하여 비윤리성을 판단하였다. 즉, '미친 연기력, 날씨 미쳤다'나 '존맛, 존예' 등은 강조의 표현으로 보고 비윤리적 표현으로 판단하지 않았다.[2] 박미은 · 정유남(2022)은 비윤리적 어휘의 형식적, 의미기능적 양상을 분석하였다. 형식적 양상은 줄임말(한남, ㅅㄲ), 변이형(능지, 머가리), 유사 파생(-충, -슬아치)로 분류하고 의미기능적 양상은 금칙어 회피(샛기, 새1끼), 사태의 부호화(틀딱, 보확찢), 특정 대상 비하(짱개, 쪽바리), 맥락 비유(개돼지, 홍어), 복합적 의미 기능으로 나누었다.[3] 하지만 역시 명사형 어휘만을 다룬다는 한계가 있다. 본 연구는 남길임 외(2021)의 분류를 따르고 조태린(2018)에서 언급한 맥락과 내용 유형에 해당하는 경우를 추가적으로 살펴본다.

국가인권위원회에서 조사에서 혐오표현을 가장 많이 경험한 장소로 온 · 오프라인을 통틀어 뉴스 기사 댓글 영역이 1위를 차지한 것으로 미루어 보아, '댓글'에서는 기사에 나타난 대상을 향해 공격적 의도를 가지고 발화하는 비윤리적 표현이 많이 나타나는 듯하다. 반면, '개인 메신저'는 온라인 영역에서 최하위를 기록하였는데 남길임 · 고예린 · 송현주(2022)에 따르면 오프라인에서 1위를 차지한 방송매체보다 개인 메신저에서 더 많은 비윤리적 표현이 나타났다.[4]

이는 메신저 대화에서는 사용자들이 발화한 표현이 비윤리적인 표현임을 스스로 인식하지 못 하는 것으로 해석할 수 있다, 즉, 메신저 대화에서는 '비의도적 용법(이정복, 2016)'의 비윤리적 표현이 두드러진다.[5]

3. 메신저 대화에서 나타난 비윤리적 표현

3.1 욕설

다음은 메신저 대화에서 수집한 비윤리적 표현의 네 가지 유형에 대한 개수와 어절 빈도를 나타낸 것이다. 본 연구에서는 비윤리적 표현을 '욕설, 혐오 및 차별 표현, 성적 표현, 공격적 표현'의 네 가지 유형으로 분류하였다. 여기서는 메신저 대화에 나타난 비윤리적 표현의 대표적인 유형인 '욕설'과 '혐오 및 차별 표현' 중심으로 유형별 특성을 살펴보고자 한다.

표 1 비윤리적 표현의 개수/어절 비율

	개수	어절	개수/어절
욕설	2,769	2,842	0.974
성적 표현	29	146	0.198
공격적 표현	74	136	0.544
혐오 · 차별 표현	172	1,347	0.127
합계	**3,044**	**4,471**	-

메신저 대화의 비윤리적 표현의 네 가지 유형 중에서 욕설이 가장 큰 비중을 차지한다. 욕설은 단어 단위로 나타나는 경우가 대부분이지만 다양한 변이형과 비표준형이 사용되기 때문에 자동 탐지에는 한계가 있다. 댓글에서는 욕설 사이에 숫자나 기호 등을 추가하거나 영어로 바꾸어 타자하는 등 다양한 방식이 동원되었기 때문에 금칙어 기반 탐지 방식에서는 새로운 비표준형이 나타날 때마다 이를 목록에 추가해왔다. 아래는 가장 높은 빈도를 보이는 욕설 세 가지의 타입과 토큰 비율과 예시를 나타낸 것이다.

표 2 욕설의 타입 토큰 비율

	TTR
'존나' 류	0.076
'씨발' 류	0.190
'새끼' 류	0.505

(1) ㄱ. ㅈㄴ, 쥐언나, 조오온나
ㄴ. 씨발, 시발, 시바, ㅅㅂ, ㅅㅂㅅㅂ
ㄷ. 시키, 쉐끼, 쉐이, 새이, ㅅㄲ

'새끼' 류가 가장 다양한 비표준형을 보이고 '씨발' 류와 '존나' 류는 비교적 덜 다양한 비표준형이 사용된다. 이는 '씨발' 류와 '존나' 류가 자소형으로 사용되는 비율이 많기 때문으로 보인다. 한편, '씨발' 류는 모든 경우에 비윤리적 표현인 의미로 사용된 반면 '존나' 류와 '미치다' 류는 비윤리적 의미로 사용되는 경우도 있지만, 단순히 강조의 의미를 갖는 경우도 있어, 비윤리성 여부를 판단하기 위해 맥락을 고려할 필요가 있다.

표 3 맥락을 고려해야 하는 유형

유형	비윤리적 의미	강조의 의미	**합계**
'존나' 류	808	593	**1,401**
'미치다' 류	252	426	**678**

'존나' 류 중에서 일반적 맥락으로 판단된 표현은 '맛있다, 재미있다, 멋있다, 좋다, 예쁘다' 등의 긍정적 용언에 '존-' 이 덧붙어 강조의 의미를 더하는 경우이다. 이 중 '존맛' 이 345건으로 가장 많이 나타났다. 하지만 '빡치다, 못생기다, 싫다, 구리다'와 같이 부정적 용언과 어울리는 경우에는 비윤리적 표현으로 사용된 것인데, 이 경우가 200건 가량 더 많이 나타났다. '미치다' 류는 총 678건 출현하였는데, 비윤리적인 사용이 273건, 강조의 의미로 사용된 경우가 426건으로 일반적 맥락에서 훨씬 더 많이 사용되는 것으로 나타났다. '미치다' 가 비하의 의미 없이 강조의 의미로 더 많이 사용된다는 점에서 '미치다' 를 일괄적으로 욕설로 판단하는 것은 실제 언어 생활을 고려하지 못한다는 한계가 있다.

3.2 혐오 및 차별 표현

<표 1>에서 보듯이 성적 표현과 차별 표현[2]은 구 단위 이상의 표현이 매우 높게 나타났다. 혐오 및 차별 표현은 성적인 표현이나 공격적 표현에 비해 어절 기준으로 10배 이상 더 많이 나타났다. 메신저 대화는 서로 친밀한 사이의 사적 대화로, 주로 음식, 여행, 쇼핑 등의 일상적 주제로 이루어지는데 자신의 경험이나 감정 등을 전달하는 과정에서 다른 취향, 문화, 상품 등에 대한 부정적 내용이 여과없이 드러나게 된다. 또한, 대화 참여자가 대체로 비슷한 연령, 성별, 지역의 배경을 가진 동질 집단인 경우가 많아 차별 표현 사용에 대한 부담이 상대적으로 적다. 아래 [그림 1]은 차별 표현 중에서 전체의 92%를 차지하는 상위 4가지의 유형의 어절 수를 나타낸 것이다. 나머지 유형에는 장애/질병, 종교, 정치, 사회, 직업, 외모 차별표현이 해당된다.

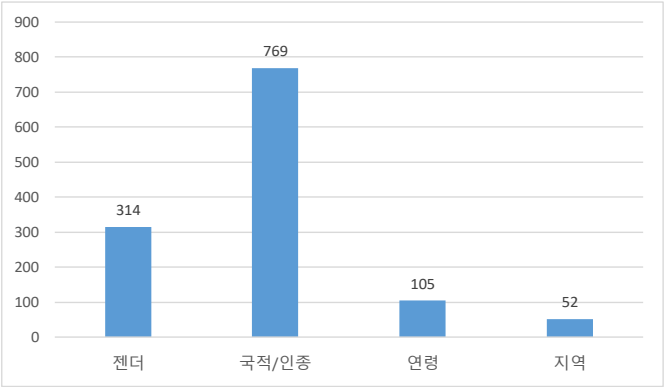

그림 1 상위 유형의 어절 수 비교

국적이나 인종에 대한 표현에서는 구 단위 이상의 차별 표현이 높은 비중을 차지하였기 때문에 어절 수의 비중이 두드러지게 높게 나타났다. 차별 표현은 '중국인은 더럽다' 와 같이 하나의 완성된 문장으로 발화되는 것이 아니라 아래와 같이 문장 이상의 단위에 걸쳐 출현한다는 특징이 있다.

(2) 1: 그리고 중국은
절대 가면 안대
2: 중국왜여?
1: 진짜더럽데
더럽대
물도진짜 쉽게먹기힘들고

(3) 1: ㅎ... 나는 설거지 시키면 개대충할거임 ㅡㅡ
짱시르네 시키는것도 민망하지않냐
2: ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ뭐시바 어차피 여기는
'여'직원이면 무조건...ㅋ
1: 남초는 그래서 안돼
여기는 극여초라
그런거 전혀 음슴

대상이 되는 국적은 외국이나 외국인 전반을 향한 표부터 중국 캄보디아, 중국, 인도, 이집트, 동남아, 유럽, 일본, 프랑스, 필리핀, 태국, 영국, 스페인, 독일, 남미, 아프리카, 미국, 소말리아, 백인, 동양인 등이 있었고 그중 일본이 32건, 중국이 18건으로 가장 많은 비중을 차지했다. 일본이나 중국에 대해서는 '원숭이, 쪽바리, 왜놈, 짱개' 등 단어 단위의 표현도 두드러졌다. 국가에 따라서 '더럽다, 냄새나다, 시끄럽다, 위험하다, 미개하다' 등의 용언과 특별히 어울리는 경우가 있으며, 'N에는 가면 안 된다, 꼭 N만 그러하다, N에 대해 편견을 안 가질 수가 없다, 일부 N이겠지만, N을 나쁘게 보려는 건 아니지만' 등과 같은 패턴으로 차별 표현을 사용하기도 하였다.

젠더 차별 표현이 높은 빈도로 나타난 것은 최근 젠더 갈등이 심각해진 상황을 반영한 것으로 보이며, 여기에는 '한남, 냄져, 김치남' 과 같은 단어 단위의 차별 표현이 매우 높은 비율을 차지한다는 특징이 있다. 특히, '한남' 이 총 19회 출현하여 가장 높은 빈도를 보였는데 이는 대화 참여자의 성별 비율이 여성이 높은 데서 기인한 것으로 보인다. 한편, '게이' 와 같은 단어는 그 자체로는 비윤리적이라 할 수 없지만 특정 맥락에서 차별표현으로 사용된 경우가 있었다.

또한, 메신저 대화에서는 대화 참여자에게 직접적으로 차별 표현을 발화한 경우는 한 건도 발견되지 않았는데, 이는 메신저에서 '비의도적 용법(이정복, 2016)' 이 두드러지는 것을 확인할 수 있다.

4. 결론

본 연구는 대화형 인공지능이 비윤리적 표현을 학습하거나 생성하는 것을 방지하기 위한 기초적 연구이다. 기존의 비윤리적 표현에 대한 논의는 주로 댓글을 대상으로 이루어져 왔지만, 대화형 인공지능이 대체로 메신저 대화의 형태로 제공된다는 점에서 메신저 대화에 나타나는 비윤리적 표현을 분석할 필요가 있다.

메신저 대화에 나타난 비윤리적 표현의 특징은 다음과 같다. 첫째, 댓글과 다르게 별도의 금칙어 설정이 없어 단어 단위의 비윤리적 표현의 사용이 자유로워 비표준형이 비교적 한정적으로 나타난다. 둘째, 실제 언어사용을 고려한다면 다소 비속한 표현 중에서도 중에서도 맥락에 따라 판단해야 하는 경우가 있다. 셋째, 연령, 성별, 지역 등의 배경이 비슷한 친밀한 관계에서 이루어지기 때문에 다른 배경을 가진 대상을 향한 부정적 표현이 가감없이 드러나 혐오 및 차별 표현의 비중이 높다. 넷째, 차별 표현의 경우 구 단위 이상의 표현이 많아 단어 이상의 단위에 대한 탐지가 필요하다.

참고문헌

[1] 조태린·김신각·유희재·김예지·이주희, <대화형 인공

2) 혐오 표현보다 차별 표현이 더 포괄적 개념이므로(이정복 2017[6]) 이하 '차별 표현'이라 한다.

지능의 윤리적 언어 표현을 위한 기초 연구>, ≪語文學≫ 140, 65-96, 2018
[2] 박미은 · 정유남, <대화형 인공지능의 윤리적 언어 표현을 위한 기초 연구>, ≪한국어학≫ 95, 241-276, 2022
[3] 남길임, 곽용진, 안미애, 김진웅, 송현주, 안의정, 황은하, 심난희, 이후영, 최지선, 강신아, 강윤희, 이갑진, 백미경, 강현아, 안진산, 황지윤, 고예린, 성민규, 장희선, 이지혜, 김수지, 정나현, 전현진, 박정혁, <2021년 맞춤법 교정 말뭉치 연구 분석>, 국립국어원, 2021
[4] 남길임, 고예린, 송현주, <사용역에 따른 비윤리적 표현의 분포와 사용 양상 연구>, 사회언어학, 30(3), 1-29, 2022
[5] 이정복, <누리꾼들의 비의도적 차별 언어 사용 연구>, 사회언어학, 24(3), 345-377, 2016
[6] 이정복, <한국어와 한국 사회의 혐오, 차별 표현>, ≪새국어생활≫, 27(3), 2017