국립국어원 2021-01-07

발 간 등 록 번 호
11-1371028-000858-01

말뭉치 언어의 사회적 인식 조사·분류

사업 책임자

이 찬 규

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘말뭉치 언어의 사회적 인식 조사·분류 사업’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2021년 06월 ~ 2021년 09월

2021년 10월 13일

사업 책임자: 이찬규 (중앙대학교 인문콘텐츠연구소)

사업 수행자 중앙대학교 산학협력단
(주)미디어 코퍼스
사업 책임자 이찬규
사업 참여자 김민수, 박일섭,
김보현, 안윤, 이은재, 조재윤, 현재홍

<사업 수행자> 중앙대학교 산학협력단·(주)미디어 코퍼스

사업 책임자	이찬규(중앙대학교 인문콘텐츠연구소 소장)
사업 참여자	김민수(동서울대학교)
	김보현(중앙대학교 인문콘텐츠연구소)
	이은재(중앙대학교 인문콘텐츠연구소)
	박일섭(미디어코퍼스)
	조재원(미디어코퍼스)
	안윤(미디어코퍼스)
	현재홍(미디어코퍼스)

<국문 초록>

말뭉치 언어의 사회적 인식 조사·분류

본 연구의 목적은 국립국어원이 기구축한 메신저, 웹, 구어 말뭉치를 대상으로 부적절한 표현과 내용의 포함 여부에 대해 일반 언어 사용자의 인식을 조사하고, 이를 통해 향후 말뭉치 구축의 참조가 될 수 있도록 비윤리적 표현을 판단하는 유형을 분류하고 배포 가능한 말뭉치의 정제 수준을 제시하는 것이다.

우리는 말뭉치 내에서 정제가 필요한 표현의 분류 범주로 '혐오 표현', '성적 표현', '욕설 표현', '차별적 표현'과 같은 비윤리적 표현과 이에 귀속되지 않는 '기타 비윤리적 표현'을 제시하였다. 아울러 개인을 식별할 수 있어 대중에게 공개할 수 없는 '개인정보'까지 총 6개를 설정하였다. 이를 근거로 하여 성별, 연령, 지역, 직업 등을 고려한 100명의 평가자를 모집하고 국립국어원이 제공한 검토용 말뭉치 총 25,190,902 어절, 407,498 문서의 비윤리적 표현에 대한 사회적 인식 조사를 실시하였다.

조사 진행 과정에서 평가자를 위한 평가 가이드라인 제시와 평가자 교육을 실시하였으며 조사 과정을 1차와 2차로 구분하여 평가자의 조사 결과를 검토하고 전문가 자문을 통해 데이터 정확성의 고도화를 시도하였다. 조사의 수월성을 위해 평가자들은 참여 기관인 미디어 코퍼스에서 제작한 평가 도구를 활용하였다. 25,190,902 어절, 407,498 문서에 대한 비윤리성 유형 분류와 문서별 정제 수준은 각각 JSON 파일과 엑셀 파일 형태로 제출하였고 아울러 변인별 통계 분석 내용과 이를 시각화한 그래프를 이 보고서에 제시하였다. 끝으로 본 연구진이 생각한 보고서의 활용방안과 관련 정책에 관한 제언을 약술하였다.

주요어: 말뭉치, 비윤리적 표현, 비윤리적 표현 유형, 비윤리적 표현 민감도 조사, 비윤리적 표현 분류 지침

차 례

표 차례

그림 차례

그림 차례

그림 차례

제 1 장

연구의 개요

1. 연구의 개요

1.1 연구의 목적

본 연구의 목적은 국립국어원이 기구축한 메신저, 웹, 구어 말뭉치를 대상으로 부적절한 표현과 내용의 포함 여부에 대해 일반 언어 사용자의 인식을 조사하고 이를 통해 향후 말뭉치 구축의 참조가 될 수 있도록 비윤리적 표현을 판단하는 지침을 수립한 후 이에 기반하여 표준화된 정제 말뭉치를 구축하는 것이다.

4차 산업혁명 시대의 도래와 함께 인공 지능 기술, 특히 인간과 자연스럽게 소통하고 공감하는 인공 지능 기술에 대한 사회적인 관심이 확산되고 있다. 이러한 기술 개발에 대한 관심과 필요가 증가함에 따라 기술 개발의 핵심 자원인 말뭉치에 대한 필요 또한 나날이 증가하고 있다.

정부에서도 이러한 수요에 대응하여 선진국 수준의 언어 능력을 갖춘 인공 지능을 개발하는 데 필요한 말뭉치를 구축하는 것을 목표로 다양한 매체로부터 실사용 언어 자료를 모으고 있으며 국립국어원 또한 4차 산업혁명에 대비한 말뭉치 구축 사업을 2018년부터 시행하고 있다.

특히 국립국어원에서는 2019년 매체 사용의 시대적 변화를 반영하여 메신저 대화와 웹을 포함한 말뭉치를 구축하였다. 구축된 말뭉치 중 2,174,506 발화 분량의 메신저 대화 4,203건과 블로그 11,521건, 게시판 9,089건, 누리소통망(SNS) 1,989,656건, 리뷰 96,810건을 국립국어원 '모두의 말뭉치'를 통해 배포하여 AI 서비스 개발과 언어 연구 등을 비롯하여 다양한 분야에서 활용할 수 있는 공공 데이터로서 공개하였다.

그런데 2021년 초 챗봇 서비스 '이루다'의 혐오와 차별 표현 논란을 시작으로 해당 서비스 개발 과정에서 개인정보를 부적절하게 취급하여 서비스가 중단되는 사태가 발생함에 따라 인공 지능 기술의 개발과 학습에 사용하는 말뭉치 언어의 윤리 문제와 공공 데이터의 개인정보 보호 문제가 제기되었다. 국립국어원에서도 사적인 상황에서 수집된 자료 내 개인의 언어 습관으로 인해 비속어나 혐오, 차별성 발언, 개인정보 등이 포함되어 있을 가능성을 고려하여 일부 말뭉치의 게시와 내려받기를 일시 중단하였다.

일련의 사건을 통해 인간과 소통하는 인공 지능에 대한 사람들의 기대에는 윤리적인 기준 또한 결코 지나칠 수 없는 중요한 고려 사항이라는 점을 알 수 있다. 언어를 통한 사회적인 소통을 위해서는 언어 표현과 내용 차원에서 시민사회의 대중이 용인할 수 있는 수준의 윤리성 또한 전제되어야 하기 때문이다.

그러나 언어를 사용하는 인공 지능 기술 개발의 역사가 그리 짧지 않음에도 불구하고 이러한 기술이 갖춰야 할 윤리적인 기준에 대한 논의와 협의가 구체화된 것은 비교적 최근의 일이다. 인간과 의사소통하는 인공 지능이 본격적으로 개발되어 다양하게 활용됨에 따라 인공 지능 언어의 윤리적 기준에 대한 사회적 요구가 증가한 것이다.

또한 매체 기술의 발전에 따라 언어 데이터의 생산량도 폭발하고 있는 최근에는 언어가 변화하는 속도 또한 빠르다. 이러한 변화 속에서 새롭게 등장하는 언어 표현과 의미 해석에 영향을 미치는 사회문화적인 맥락을 고려하여 표현과 내용의 윤리적인 적절성을 판단하기 위해서는 이에 대한 새로운 판단 기준이 마련되어야 한다.

이에 본 연구에서는 성별, 연령, 지역, 직업 변인의 대표성을 고려하여 선발한 평가자의 언어 직관을 통해 말뭉치 언어의 표현과 내용이 갖는 윤리적 적절성과 개인정보 포함 수준에 대한 판단을 내려 보고자 한다. 평가의 대상이 되는 말뭉치는 사적인 상황을 전제로 이루어져 개인의 언어 습관에 따른 자유로운 언어 사용 습관과 개인정보가 나타날 가능성이 큰 메신저 말뭉치, 웹 말뭉치, 일상 구어 말뭉치이다. 그리고 평가자의 판단을 종합하여 부적절한 표현과 내용이 포함된 말뭉치를 제외한 배포용 말뭉치를 구축하는 근거로 삼고자 한다.

본 연구의 세부 목표는 아래와 같다.

○ 말뭉치 언어에 대한 윤리성 검증 체계를 구축한다.

- 기구축된 말뭉치 내에 비윤리적으로 판단될 수 있는 표현을 구체적으로 분류하기 위해 4가지 유형(혐오 표현, 성적 표현, 욕설 표현, 차별적 표현)을 제시한다.
- 아울러 개인정보 등장 문서를 위와 같이 비윤리적 표현이 담긴 문서로 간주한다.
- 상기 내용에 명확히 포함된다고 판단할 수 없는 비윤리적 표현에 대해서는 '기타' 항목을 두어 분류한다.

○ 비윤리적 표현에 대한 사회적 인식 조사를 실시한다.

- 국립국어원에서 제공하는 말뭉치(메신저, 웹, 구어) 25,190,902 어절 내 비윤리적 표현이 포함되어 있는지에 대한 평가 조사를 실시한다.
- 이를 위하여 기존의 비윤리 언어 검출에 대한 국어학과 윤리학의 연구 성과를 참고하여 구체적인 설문 조사 방안을 설계한다.
- 크라우드 워커의 성별, 연령, 지역, 직업과 같은 다양한 결과를 도출할 수 있는 변인에 따라 인식 조사를 실시하고 그 결과를 분석적으로 제시한다.
- 성별, 연령, 지역, 직업 분류에 따른 변인을 고려하여 일반 언어 사용자 100명을 대상으로 평가 조사를 실시한다.
- 평가를 위하여 조사 대상자 누구나 접근이 용이한 작업 도구를 개발하여 활용한다.

○ 비윤리적 표현 검출 및 분류는 다음과 같은 방법으로 실시한다.

- 말뭉치를 문장 단위로 나누어 제시한다.
- 비윤리적 표현이 포함된 부분을 구체적으로 표시한다.
- 비윤리적 표현의 유형을 분류하여 표시한다.
- 문서별 비윤리적 표현 등장 빈도에 따라 정제 수준(비윤리성의 강도)을 평가한다.

○ 조사 결과를 분석하고 말뭉치의 윤리성 제고와 활용도 증진을 위한 개선 방안을 제시한다.

1.2 연구의 대상 및 범위

본 연구에서는 국립국어원이 제공하는 3종의 말뭉치를 대상으로 하여 개인정보 포함 여부와 윤리적인 적절성 여부에 대한 판단이 이루어졌다. 조사 대상인 3종의 말뭉치는 구어 말뭉치 중 사적 대화, 메신저 대화 말뭉치, 웹 말뭉치이다. 이들 말뭉치는 사적인 상황에서 이루어진 대화와 웹 게시물이라는 특성을 지니

기 때문에 개인의 언어 사용 습관과 개인을 특정하는 정보가 여과 없이 나타날 가능성이 크고 이로 인해 부적절한 표현이나 내용을 비롯해 상대방이나 제3자의 이름과 같은 개인정보가 노출될 가능성도 큰 범주이다.

구어 말뭉치, 웹 말뭉치, 메신저 대화 말뭉치 전체 구축 내용 중에서 본 연구의 조사 대상은 사적 대화 2,224건, 웹 말뭉치는 모두의 말뭉치에 배포된 웹 말뭉치 분량의 19.02%를 차지하는 400,680건, 메신저 대화 말뭉치는 모두의 말뭉치에 배포된 4,203건과 배포되지 않은 391건을 합산한 4,594건 전체 25,190,902 어절을 대상으로 평가가 이루어졌다.

	전체 구축 규모		사업 기간 조사 규모	
	유형	분량(건)	분량(건)	어절 수(어절)
구어 말뭉치	공적 독백	2,490	-	8,711,096
	공적 대화	19,104	-	
	사적 대화	2,224	2,224	
	준구어-대본	4,102	-	
소계		27,920	2,224	
웹 말뭉치	블로그	11,521	2,050	9,766,838
	게시판	9,089	517	
	누리소통망	1,989,656	395,680	
	리뷰	96,810	2,433	
소계		2,107,076	400,680	
메신저 대화 말뭉치	배포용	4,203	4,203	6,712,968
	미배포용	391	391	
소계		4,594	4,594	
합계		**2,139,590**	**407,498**	**25,190,902**

〈표 1〉 말뭉치 언어의 사회적 인식 조사 연구 대상 문서 및 어절

위 연구 대상을 기준으로 본 연구의 범위는 사업의 범위와 연구 수행 집단의 구성으로 이루어진다.

○ 사업의 범위는 아래의 표와 같다.

사업의 범위	세부 내용
말뭉치 언어 (내용 및 표현)에 대한 사회적 인식 조사	• 메신저, 웹, 구어 등 약 2천 5백만 어절의 말뭉치 내 부적절한 표현 및 내용(혐오, 차별 등), 개인정보 포함 여부 조사 • 일반 언어 사용자 100명 이상(성별, 연령, 지역 고려)을 대상으로 설문 조사 • 정제 대상(부적절한 내용 및 표현, 개인정보) 포함 여부 조사
조사용 말뭉치 제시 방법 및 질의, 평가 방법 설계	• 조사 대상자에게 주어지는 말뭉치의 효과적 제시 방안 모색 • 부적절한 표현, 부적절한 내용, 개인정보 포함 등 정제 대상의 판단 근거에 대한 유형화 • 응답자 특성에 따른 조사 결과 분석 및 추후 말뭉치 구축 및 정제 사업을 위한 윤리적 지침 제시
부적절한 표현 및 내용을 제외한 배포용 말뭉치 구성	• 조사 결과를 바탕으로 한 정제 말뭉치 문서 단위 분류 • 부적절한 표현 및 내용이 포함된 문서를 제외한 배포용 말뭉치 구성
품질 검증	• 자문 위원회 운영: 품질 목표 수립 및 자문·품질 검수 • 설문 조사 기준 검증, 시정, 품질 확보 • 설문 조사 결과 검증, 시정, 품질 확보 • 배포용 말뭉치 구축 결과 검증, 시정, 품질 확보
산출물 납품	• 중간 산출물: 조사 지침(조사 대상, 문항 등 포함), 설문 조사 샘플, 중간보고서 • 완료 시: 완료계 및 사업 결과 보고서, 배포용 말뭉치, 기타 산출물(응답자 특성별 말뭉치 문서 분류 목록, 설문 조사 결과 가공 파일, 설문 조사 통계 조사 파일, 기타 자료 일체)
프로젝트 관리	• 공정 관리/위험 관리/자원 관리/일정 관리/보안 관리/산출물 관리 수행 • 일정 계획/세부 활동 도출/중간목표 수립/자원 배분 • 작업 도구 구현 • 과업 관리: 산출물 관리, 단계별 진행 관리, 완성도/보안 관리 • 보고 관리 - 정기 및 비정기 단계별 진행 사항 보고
프로젝트 지원	• 품질 보증 계획 수립 및 품질 확보 • 교육 훈련: 교육 훈련 방법/내용/일정/조직 운영/기술지원 • 하자 보수 계획 수립 및 보증

〈표 2〉 말뭉치 언어의 사회적 인식 조사·분류 사업 범위

○ 연구 수행기관의 구성은 다음과 같다.

[그림 1] 과업 수행 조직

○ 사업 조직별 수행 내용은 다음과 같다.

사업 수행 조직	사업 조직별 수행 내용
총괄 실무 책임	• 공정 관리/위험 관리/자원 관리/일정 관리/보안 관리/산출물 관리 수행 • 일정 계획/세부 활동 도출/중간목표 수립/자원 배분 • 보고 관리 - 정기 보고 및 비정기 보고, 단계별 보고 • 위기 관리 - 위험 요소 사전 분석 및 대응 전략 수립, 모니터링 • 납품 관리 - 착수, 중간, 완료 각 단계별 산출물 납품 • 하자 보수 계획 수립 및 보증 활동
자문 위원회	• 윤리 데이터 설계 및 철학(윤리) 전문가를 자문위원으로 초빙 • 조사 설계 단계의 평가 문항 및 조사 방안 자문 • 사업 진행 단계별 조사 및 산출물 문제 자문 • 사업 완료 단계의 조사 품질 자문 • 오류 사항 점검 및 분석 • 오류 개선 방안 도출 및 시정 요청
설계/분석팀	• 말뭉치 언어 사회적 인식 조사 기획 및 설계 - 성별, 나이, 직업, 지역 등 선별 기준 수립 - 평가용 말뭉치 제시 방법, 조사 방법 및 문항 설계 - 평가 도구 설계 • 평가 결과 분석 및 문서 분류 기준 수립 - 응답자 특성별 응답 특성 분석 - 응답 특성에 따른 문서 정제 수준 분류 기준 수립 • 작업 지침 수립 / 교육 / 매뉴얼 제작 - 평가 방법에 대한 지침 수립 및 매뉴얼 제작 - 분석 방법에 대한 지침 수립 및 매뉴얼 제작 - 작업자 대상 교육 운영
조사/데이터팀 조사 분야	• 조사 대상자 모집을 위한 홍보 운영 • 참여자 현황 분석을 통한 홍보 모집 전략 수립 • 조사 대상자 모집을 위한 외부 협력 수행 • 조사 대상자 모집 및 대화 수집 • 조사 대상자 인구 통계 정보 수집 / 관리 • 평가 말뭉치 배포
조사/데이터팀 데이터 분야	• 응답자별 평가 결과 데이터 관리 • 평가용 말뭉치 생성 • 배포용 말뭉치 구축

〈표 3〉 조직별 수행 내용

○ 전문가 자문위원 위촉 및 자문 활용 내역은 다음과 같다.

자문위원	소속	성명	전문분야	직위
	중앙대학교 인문콘텐츠연구소	김형주	철학(윤리)	교수
자문활용	1차 자문: 착수보고회 이후 - 조사 평가 가이드라인 제작 및 비윤리 범주 구분에 대한 자문 활용 2차 자문: 중간보고회 이후 - 조사 방안 보완 방안 자문 활용 3차 자문: 최종보고회 이후 - 조사 결과물 분석 내용 보완 자문 활용			

〈표 4〉 자문위원 위촉 및 활용 내역

제 2 장

조사 방법

2. 조사 방법

2.1 사회적 인식 조사를 위한 비윤리적 표현 유형 분류 기준 설정

2.1.1 말뭉치 언어의 비윤리적 표현 유형 분류 기준 제시

가. 비윤리적 표현 유형 분류 기준 배경

앞에서 밝혔듯 조사 대상 말뭉치는 모두의 말뭉치 중 사적 대화, 메신저 대화 말뭉치, 웹 말뭉치로서 사적인 상황에서의 의사소통이며 개인의 언어 사용상의 습관이 드러난다는 특성상 사회적으로 용인될 수 없는 표현, 개인을 특정하는 정보가 여과 없이 나타날 가능성이 크다. 구체적으로 말하자면 말뭉치 언어 구축 사업 추진 단계에서 민감 정보를 비식별화하고 비윤리적 표현을 정제하고자 하였음에도 불구하고 다양한 경로로 수집된 데이터의 방대함과 비윤리적 표현에 대한 개별 언어 사용자의 윤리적 기준의 차이로 인해 우리가 조사 대상으로 삼은 말뭉치 안에는 사회적으로 용인되기 어려운 표현이 잔존할 수 있다. 이러한 문제를 해결하기 위한 첫 단계는 비윤리적 표현의 기준 설정이라고 판단하여 다음과 같은 선행 연구들을 참고하여 분류의 기준을 설정하였다.

이를 위하여 본 연구진이 참여하여 작성한 “윤리적 인공지능을 위한 비도덕 문장 판별 온톨로지 구축에 대한 연구(이청호 외 2021)”와 “방송심의에 관한 규정(방송통신심의위원회 규칙 제150호, 2020)” 및 국내외 도덕, 윤리 교과서를 검토하였다. 해당 내용을 근거로 하여 비윤리적 언어 사용을 가치 기준별로 유목화하였으며 비윤리적 발화 행위 유형을 아래와 같이 구성하였다.

‘사회적으로 용인되지 않는 표현’에 해당하는 비윤리적 발화 행위 유형은 총 4가지 종류이다. 특정 개인이나 집단을 향한 증오를 드러내는 ‘혐오 표현’, 노골적이고 선정적인 언어 사용으로 성적 수치심을 일으키는 ‘성적 표현’, 상대방에게 불쾌감이나 모욕감을 줄 목적으로 사용하는 비속어인 ‘욕설 표현’, 특정 성별이나 계층, 지역 등을 구분 지어 공격하거나 갈등을 조장하는 ‘차별 표현’으로 범주화할 수 있다.

여기에 더하여 언어가 가진 사회・문화・역사적 특성 때문에 위의 4가지 분류

유형 범주에는 포함되지 않음에도 여전히 ‘사회적으로 용인되지 않는 표현’으로 여겨질 수 있는 것은 ‘기타’비윤리적 표현도 분류하였다.

또한 개인정보는 비윤리적 표현은 아니지만 개인의 동의 없이 사용될 경우 사회적 차원에서 다양한 법적, 윤리적 문제가 발생할 수 있기 때문에 기타 비윤리적 표현과 마찬가지로 공적으로 배포되는 말뭉치에는 포함되지 않는 것을 원칙으로 하여야 한다. 따라서 ‘개인정보’를 비윤리 유형의 하나로 범주화하였다.

이상의 기준으로 추출된 표현을 비교・검토함과 더불어 각 표현 및 내용의 출현 빈도와 성별, 연령, 지역 등의 변인을 고려하여 정량적 지표를 다각도로 제시하였다.

나. 말뭉치 언어의 비윤리적 표현 유형 분류 개념 설명

결과적으로 말뭉치 언어의 사회적 인식 조사를 위한 비윤리적 표현의 유형은 6가지로 설정되었다. 각 유형에 대한 개념적 규정은 다음과 같다.

○ ‘혐오 표현’은 특정 개인 및 집단과 이들이 가진 속성에 대하여 적의, 혐오의 감정을 명시적으로 드러내는 표현을 말한다.

○ ‘성적 표현’은 특정 개인 및 집단에 대하여 성적(sexual)으로 묘사하거나 불필요한 맥락에서 특정 신체 부위 및 성적 행위를 적나라하게 드러내는 표현을 말한다. 대화의 소재가 성애와 관련한 연상을 불러일으킬 수 있는 신체 부위, 도구 등의 표현이 등장한 경우에도 비윤리적 성적 표현에 해당하는 것으로 본다.

○ ‘욕설 표현’은 격이 낮고 속된 말, 대상을 얕잡아 보고 경멸하는 태도를 드러내거나 타인에게 불쾌감을 주는 표현을 말한다. 또한 의도적으로 맞춤법을 변형시키거나 단어를 축약해서 새로운 뉘앙스를 부여한 신조어 또한 욕설로 간주한다. 최근에 많이 등장하는 초성만 나열한 경우도 다수의 사람들이 욕설로 해석할 수 있는 경우 도덕적 금지어로서 욕설에 해당하는 것으로 본다.

○ ‘차별적 표현’은 암묵적으로 특정 개인 및 집단을 분리하고 불평등하게 대우하는 표현을 말한다.

○ 위의 4가지 유형에 해당되지는 않지만 비윤리적이라 판단되는 경우는 ‘기타’로로 분류한다.

○ 문장 내에서 개인을 특정할 수 있는 민감 정보가 포함되면 ‘개인정보’ 유형으로 분류하여 분석한다. 이러한 정보는 다음과 같은 사항들을 지칭한다.

실명, 주민등록번호, 전화번호, 주소, 이메일 주소, 위치, 계좌 정보, 아이디/비밀번호/닉네임 등.

2.1.2 말뭉치 언어의 비윤리적 표현 유형 분류 기준의 타당성 확보

가. 비윤리적 표현 판별을 위한 온톨로지 개발 선행 연구 참조

주지하듯 ‘윤리’에는 우리 모두가 익히 잘 알고 있는 ‘일상어로서의 윤리’, 그리고 하나의 학문으로서 ‘윤리학’이라는 상이한 두 가지 의미가 상존한다. 본 연구의 문제의식은 AI 기술의 발달로 인해 최근 우리 사회에 가장 많이 이슈가 되는 단어에 속하는 ‘AI 윤리’가 주로 ‘윤리’의 첫 번째 의미에 집중하고 있다는 사실에서 출발한다. 따라서 이 연구에서 문장의 비윤리성을 판별하는 과정은 ‘윤리’와 ‘도덕’의 의미 차이나 각각에 대한 명확한 규정에 집중하지 않는다. 사회적인 상식에 의지한 판단을 통하여 사회적으로 명백히 허용할 수 없는 문장들을 가려낼 수 있도록 하는 것이 더 중요한 목적이기 때문이다.

위와 같은 맥락에서 비윤리적 표현 판별을 위한 온톨로지 개발 연구 내용을 본 연구의 맥락에 맞게 참조하였다.[1)]

비윤리적 표현 판별을 위한 온톨로지 개발 연구의 특징은 제거적 귀류법을 적용한다는 것이다. 윤리적인 문장의 조건이나 특성을 미리 규정하지 않고, 비윤리적 표현을 제거함으로써 ‘윤리적 표현’을 간접적으로 규정하는 방법이다.

---

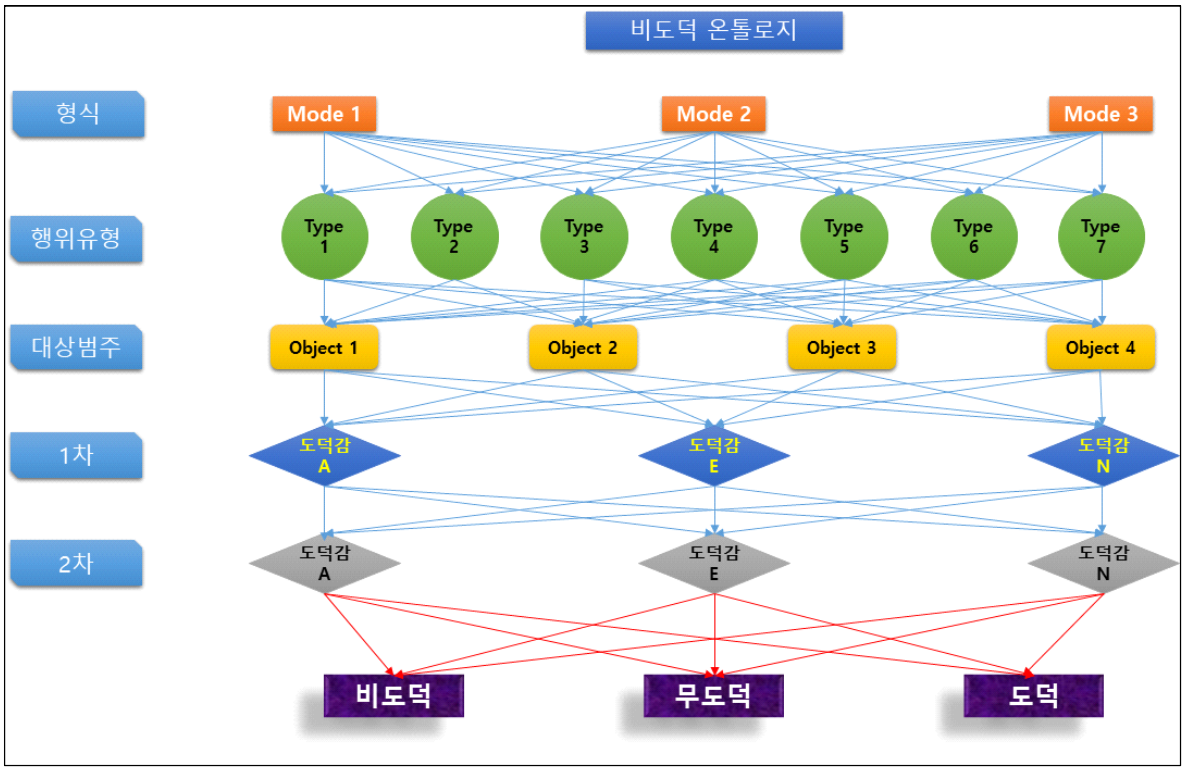
1) 온톨로지 개별 연구의 내용은 본 연구의 연구 책임자인 이찬규 교수와 자문위원인 김형주 교수가 작성한 논문인 “윤리적 인공지능을 위한 비도덕 문장 판별 온톨로지 구축에 대한 연구(인공지능인문학 7집, 2021)”를 전거로 삼았다.

또 '옳음' 혹은 '좋음'의 추구를 원칙으로 도덕과 윤리에 대해 하향식으로 접근하는 의무론과 공리주의 윤리학, 상황맥락적인 윤리 요소를 강조하는 덕 윤리의 상향식 접근을 절충하여 다양한 윤리적 언명의 상황을 고려하는 윤리 검증의 틀을 마련하고자 하였다.

다음은 우리가 참조한 비도덕 문장 판별 온톨로지 구축의 한 모델이다. 먼저 비윤리적인 문장으로 판별되는 문장의 형식(mode)을 그 문장이 어떻게 비윤리성을 가지게 되는지에 따라 두 유형(도덕적 금지어가 긍정되는 Mode 1과 도덕적 가치어가 부정되는 Mode2)으로 나눈다. 그리고 문장의 내용 요소(element) 별로 7가지의 행위 유형을 구분한다. 마지막으로 화자나 문장이 지칭하는 대상의 범주를 결정한 후 최종적으로 문장의 비윤리성을 확정한다.

<table>
<tr><th colspan="3">형식
(Mode)</th><th>Mode 1
도덕적 금지어 + 도덕적 긍정정서 표현</th><th>Mode 2
도덕적 가치어 + 도덕적 부정정서 표현</th></tr>
<tr><td rowspan="2">S1</td><td colspan="2">유형
(Type)</td><td>①차별 행위 유형
②(물리적) 폭력 행위 유형
③선정 행위 유형
④욕설 행위 유형</td><td>⑤혐오(증오) 행위 유형
⑥범죄적 행위 유형: 살인, 사기, 강간
⑦비난 행위 유형: 조롱, 모독, 비방 등</td></tr>
<tr><td colspan="2">관련 내용 요소
(Elementary)</td><td>편견, 욕설, 차별, 폭력, 증오, 살인, 학대, 절도, 유괴, 고문, 혐오, 음란, 모독, 비방, 조롱 등</td><td>정직, 자주, 성실, 절제, 책임, 용기, 효도, 예절, 협동, 민주적 대화, 준법, 정의, 배려, 애국・애족, 평화・통일, 생명존중, 자연애, 사랑 등</td></tr>
<tr><td>S2</td><td colspan="2">대상 범주
(Object)</td><td>①개인(성별, 연령, 학력, 직업, 외모, 장애)
②공동체(계층, 지역, 인종, 국가, 민족)</td><td>③문화(종교, 습속, 역사)
④자연(동물, 생명체)</td></tr>
<tr><td rowspan="2">P</td><td rowspan="2">도덕정서술어</td><td>대상 평가 서술어 (1차)</td><td colspan="2">Positive ①도덕적 긍정정서 술어: 착한, 선한, 좋은, 옳은, 즐거운, (good, right, pleasant, like)
Negative ②도덕적 부정정서 술어: 나쁜, 잘못된, 틀린, 불쾌한, 싫은(bad, wrong, unpleasant, dislike)</td></tr>
<tr><td>화자 판단 서술어 (2차)</td><td colspan="2">Positive ①도덕적 긍정정서 술어: 착한, 선한, 좋은, 옳은, 즐거운, (good, right, pleasant, like)
Negative ②도덕적 부정정서 술어: 나쁜, 잘못된, 틀린, 불쾌한, 싫은(bad, wrong, unpleasant, dislike)</td></tr>
<tr><td colspan="3">분류</td><td>비도덕적 문장</td><td>비도덕적이지 않은 문장</td></tr>
</table>

〈표 5〉 비도덕적 문장 판별 모델



[그림 2] 비도덕적 문장 판별 온톨로지

이상의 내용을 참조하여, 일반 언어 사용자의 비윤리적 표현에 대한 민감도 판별의 수월성을 높이기 위하여 위 결과를 단순화고 비윤리적 표현의 유형을 혐오 표현, 성적 표현, 욕설 표현, 차별적 표현의 4가지로 한정하였다. 아울러 4가지 범주에 해당하지 않지만 비윤리적 표현이라고 간주될 수 있는 표현들은 '기타'로 범주화하였다.

나. '방송심의에 관한 규정'에서 비윤리적 표현 유형 분류 참조

비윤리적인 문장을 판별하기 위한 기준을 마련하는 과정에서 고려하지 않을 수 없는 중요한 요소 중 하나는 바로 언어의 사용에 대한 윤리성 판단의 바탕에는 사회·관습적 인식이 자리하고 있다는 점이다.

물론 '사회적 합의'나 '사회 상규' 같은 개념은 그것이 일종의 윤리 감정이라는 점에서 그 자체로 준거가 될 수 있을 만큼 객관적이거나 구체적이지는 못하다. 하지만 그러한 윤리 감정을 가장 일반적인 형태로 정제해 반영할 수 있다면 언어의 비윤리성을 판단하기 위해 충분히 참고할만한 기준이 될 것이다.

이러한 이유에서 본 연구진은 방송통신심의위원회가 '방송심의에 관한 규정'에

서 규정하고 있는 불쾌감 유발 언행, 욕설 표현, 성적 수치심 유발 표현 혐오감 유발 표현 등 4가지 비윤리적 표현에 대한 유형 분류 기준이 우리의 연구와 유사하다고 판단하여 아래와 같은 관련 내용을 참조하였다.

제4절 윤리적 수준

제27조(품위 유지) 방송은 품위를 유지하기 위하여 시청자의 윤리적 감정이나 정서를 해치는 다음 각 호의 어느 하나에 해당하는 표현을 하여서는 아니 되며, 프로그램의 특성이나 내용전개 또는 구성상 불가피한 경우에도 그 표현에 신중을 기하여야 한다. <개정 2015.10.8., 2020.12.28.>

1. 불쾌감을 유발할 수 있는 과도한 고성·고함, 예의에 어긋나는 반말 또는 음주 출연자의 불쾌한 언행 등의 표현
2. 신체 또는 사물 등을 활용하거나 의도적으로 무음·비프음, 모자이크 등의 기법을 사용한 욕설 표현
3. 혐오감·불쾌감을 유발할 수 있는 성기·음모 등 신체의 부적절한 노출 또는 과도한 부각, 생리작용, 음식물의 사용·섭취 또는 동물사체의 과도한 노출 등의 표현
4. 불쾌감이나 성적 수치심을 유발할 수 있는 부적절한 신체 접촉, 신체 촬영, 성적 언행 등에 대한 표현
5. 그 밖에 불쾌감·혐오감 등을 유발하여 시청자의 윤리적 감정이나 정서를 해치는 표현

[전문개정 2014.12.24.]

- 이하 생략 -

이상의 내용을 토대로 하여 일반 언어 사용자의 접근 가능성, 판단의 수월성, 작업 도구의 효율성 등을 고려하여 비윤리적 표현 유형을 상기한 바와 같이 4가지로 규정하고 기타 비윤리적 표현 및 개인정보를 포함하여 총 6가지 분류 기준을 설정하였다.

2.2 사회적 인식 조사의 평가 절차

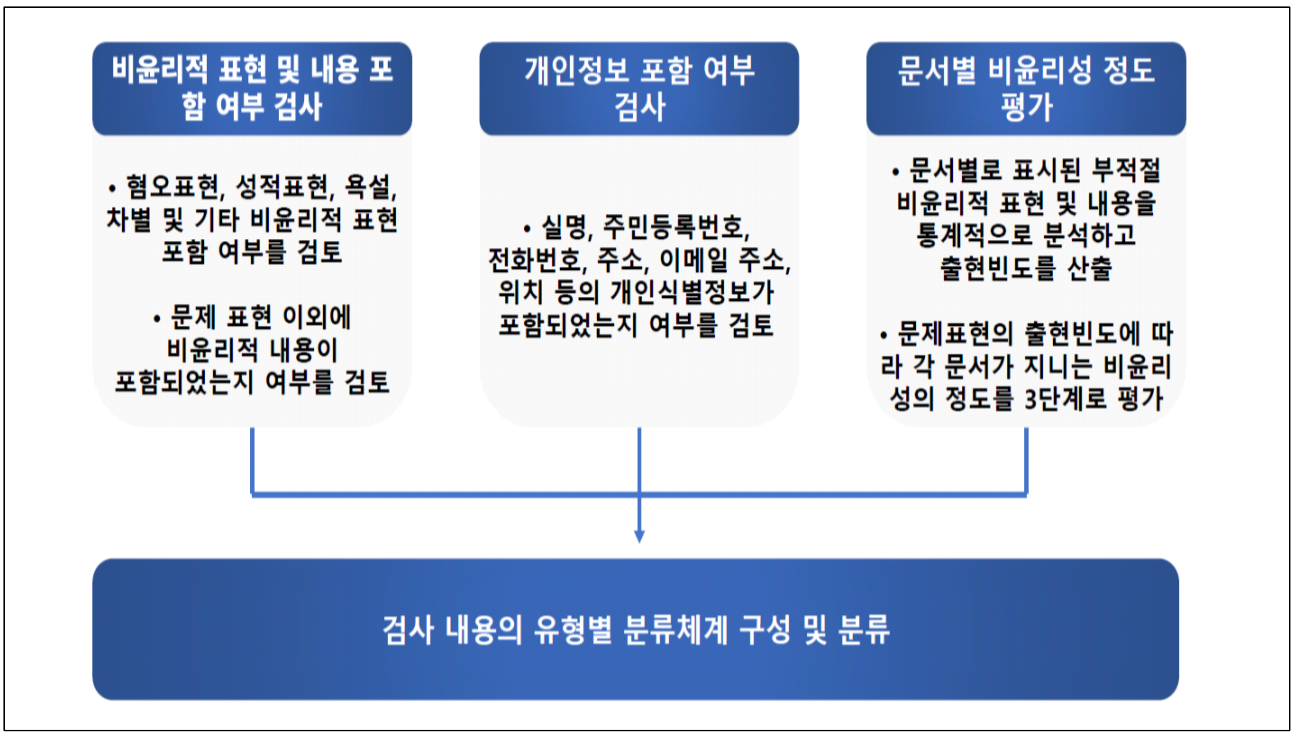
2.2.1 사회적 인식 조사를 위한 평가자용 가이드라인 및 평가 절차 제시

말뭉치 언어의 사회적 인식 조사를 실시하기에 앞서 실제 평가 작업을 어떻게 진행할 것인지에 대한 가이드라인과 평가 절차를 먼저 확립하였다. 해당 내용은 다음과 같다.

가. 평가 가이드라인 일반 사항

○ 빠른 시간 내에 평가자들이 평가 작업에 숙련될 수 있도록 하는 것이 중요하다. 따라서 직관적인 작업 도구를 직접 개발해 활용함으로써 비윤리적 표현 및 내용에 대하여 일반 언어 사용자가 효율적으로 평가할 수 있도록 한다.

○ 조사가 끝나고 나면 유형별 출현 빈도를 분석하고 표현 및 문서별 정제 수준의 강도를 효과적으로 제시한다.



[그림 3] 비윤리적 표현 평가 체계

나. 말뭉치 내 비윤리적 표현 내용 유무에 대한 평가자 인식 조사

○ 평가자는 국립국어원 메신저, 웹, 일상대화 말뭉치 내 혐오 표현, 성적 표현, 욕설 표현, 차별 및 기타 비윤리적 표현이 포함되었는지 여부를 검토한다.

○ 평가자는 위의 문제 표현 이외에 윤리적 금기어나 비속어 등이 포함되지 않았음에도 비윤리적으로 판단할 수 있는 내용이 포함되었는지 여부를 검토한다.

○ 평가자는 국립국어원 메신저, 웹, 일상대화 말뭉치 내 실명, 주민 등록 번호, 전화번호, 주소, 이메일 주소, 위치 등의 개인 식별 정보가 포함되었는지 여부를 검토한다.

○ 평가자는 문제 표현, 내용 또는 개인정보 발견 시 해당 표현 및 내용, 개인정보의 위치와 유형을 표시한다.

다. 말뭉치 내 비윤리적 표현 유형별 빈도 통계 산출

○ 말뭉치 내 각 문서별로 표시된 비윤리적 표현을 통계적으로 유형을 분석하고 출현 빈도를 산출한다.

○ 산출된 문제 표현의 출현 빈도에 따라 각 문서가 지니는 비윤리성의 정도를 3단계(상-높음・중-보통・하-낮음)로 평가한다.

2.2.2 평가자 조사 수행 지침 설명

가. 말뭉치 내 비윤리적 표현 분류 기준에 대한 설명 제시

○ 혐오 표현, 성적 표현, 욕설 표현, 차별적 표현 등 말뭉치 공개 시 일반 대중에게 윤리적 논란의 소지가 있는 항목을 선별할 수 있도록 그 분류 기준에 대해 설명한다.

○ 이름, 주소, 연락처, 금융 거래 정보 등 유출되었을 때 심각한 사생활 침해를 일으킬 수 있는 개인정보 항목을 선별한다.

○ 평가자의 작업 효율화를 위하여 직관적인 평가 항목을 아래의 6개로 선별한다.

① 혐오 표현	② 성적 표현	③ 욕설 표현	④ 차별적 표현
⑤ 기타 비도덕적 표현	⑥ 개인정보		

나. 말뭉치 내 비윤리적 표현 유형별 평가 진행 방법 지침

○ 국립국어원이 제공하는 평가용 말뭉치를 문서 단위로 평가자에게 제공한다.

○ 평가자의 작업 부담 및 평가의 균형을 고려하여 평가자별로 문서 유형, 문서 개수, 평가 발화(utterance) 개수를 균등하게 배정한다[2].

○ 평가용 말뭉치의 발화를 하나의 평가 단위로 하여 발화 내부에 포함된 평가 대상 항목을 평가자들이 선별한다.

○ 평가자들이 선별한 항목을 유형별 출현 빈도를 기준으로 정량 분석할 수 있는 평가 결과 파일을 산출물로 생성하여 분석 담당 연구진에게 제공한다.

2) 전체 평가자들이 균등한 양의 평가 작업을 할 수 있도록 평가 문서의 수량과 평가해야 할 문장의 수량을 평가자들에게 균등하게 배분하였다. 주관 기관을 통해 제공받은 웹 문서는 개별 문서 하위 문단의 개별 발화들이 구어 대화와 메신저 대화에 비해 길이가 긴 것들이 많았기 때문에 문장 부호를 단위로 분절한 후 분절된 개별 발화를 평가 문장 수량으로 환산하였다. 평가 작업의 편의를 위해 분절된 발화들은 이후 산출물 생성 단계에서 다시 JSON 원문의 형태와 비교를 통해 원문의 형태를 복원하였다.

2.3 사회적 인식 조사 평가 도구 자체 개발 및 적용

2.3.1 사회적 인식 조사 평가 도구 자체 개발

가. 평가 도구 자체 개발 과정

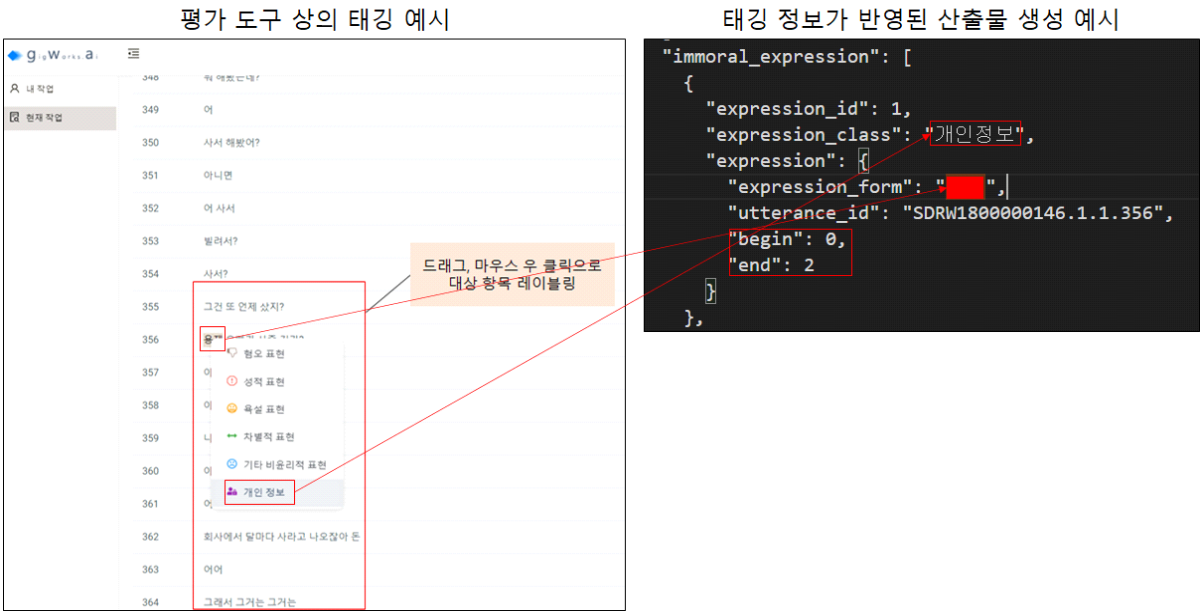
이번 연구 사업을 위해 공동 사업 수행 기관인 미디어 코퍼스에서 평가용 도구를 자체적으로 개발하였다.

주관 기관에서 평가를 위해 제공한 JSON 문서는 구조적으로 “데이터 이름: 값”의 형식에 개별 발화의 id, 화자 id, 주석, 발화 일시 등 발화에 대한 메타 데이터가 부착되어 있고, 분절된 발화 단위로 데이터가 그룹화되어 있다. 발화의 전후 맥락을 고려하여 내용의 적절성 여부를 판별해야 하는 작업의 특성을 고려할 때, 발화 단위로 분절되어 있을 뿐만 아니라 발화 자체 내용과 관련이 없는 메타 데이터까지 포함된 JSON 형식은 가독성 측면에서 평가자에게 그대로 배포하여 평가에 활용하기가 어렵다.

아울러 JSON 파일을 문서 형태로 평가자에게 제공하여 부적절한 표현과 내용을 찾고, 문서에 직접 표시하게 할 경우, 태깅 정보가 반영된 최종 산출물 생성 단계에서도 발화 안의 태그 위치 색인 정보(begin/end index)는 수기로 입력할 수밖에 없다. 이를 수기로 입력할 경우 작업 시간이 과다하게 걸리는 문제뿐만 아니라, 이 과정에서 발생하는 작업자의 실수가 데이터 품질에도 영향을 미칠 우려가 있다.

따라서 아래 [그림 4]와 같이 태깅 단계부터 산출물 생성 단계까지 작업의 효율과 데이터 품질을 고려하여 평가자들이 할당된 문서에 대하여 빠르게 읽고 평가가 가능하도록 하고 평가자 정보, 문서 정보, 태깅 정보를 일괄 DB에 저장하여 산출물 생성이 용이하도록 하는 것에 중점을 두고 평가 도구를 개발하였다. 이를 위해 JSON 형식으로 제공된 말뭉치 원문은 파싱 단계를 거친 후 평가자들이 읽어야 하는 발화만을 추출하여 가독성을 고려한 문서 보기 방식을 적용하고 태깅이 이루어진 부적절 표현의 위치 색인 정보는 마우스 드래그를 통한 입력으로 DB에 저장하여 아래 [그림 5]와 같이 산출물에 즉각 반영이 가능한 방식을 적용한 웹 기반으로 제작하였다.

[그림 4] 주관 기관이 제공한 JSON 원문을 파싱한 이후 평가 도구에 웹 문서로 변환한 화면



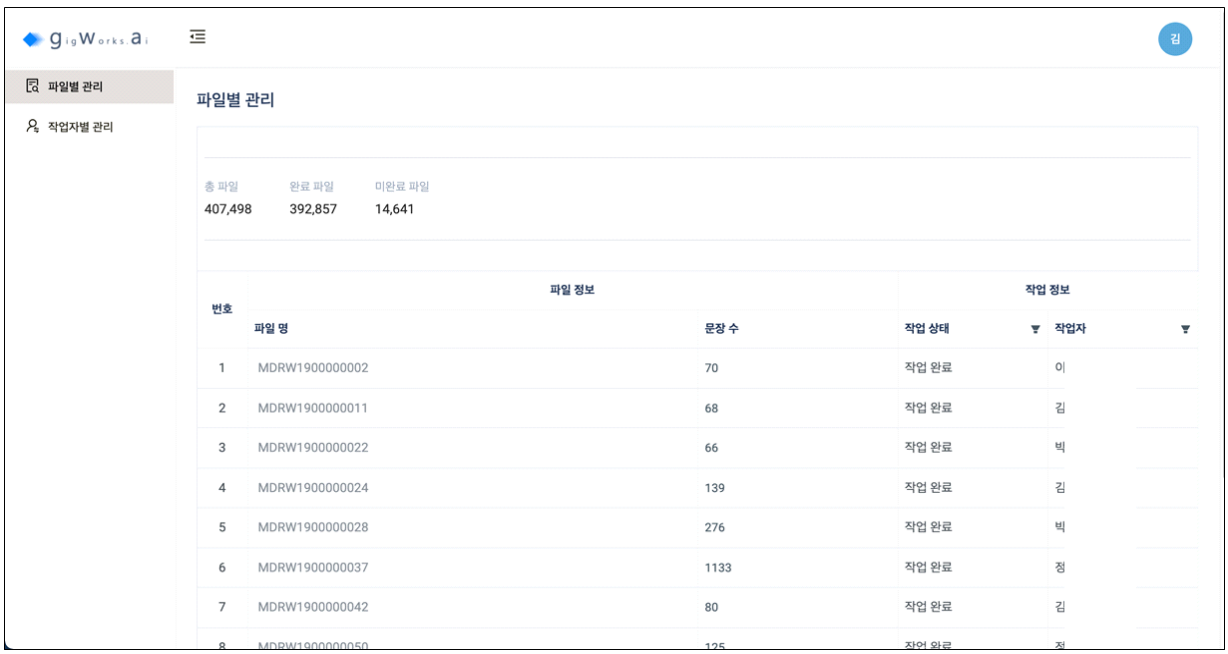
[그림 5] 평가 도구에서 이루어진 태깅 정보를 반영하여 생성한 산출물 예시

평가자들이 부적절한 표현을 찾아서 표기하는 태깅 작업뿐만 아니라 평가 작업 현황과 작업 품질 관리를 위해 관리자 기능을 동시 개발하여 관리자가 평가자별/문서별 진척 현황 관리 및 레이블링 현황 점검, 수정이 가능하도록 하였다.

또한 아래의 [그림 6]과 같이 관리자 기능을 별도로 개발하여 평가자들의 정보와 작업 진척 상황을 정확하게 파악하고 효율적으로 관리할 수 있도록 하였다. 관리자 기능을 통해서 [그림 7]과 같이 평가 문서별 작업 상태를 확인할 수 있

으며 현재 작업 진행 상태와 해당 문서의 평가자가 누구인지를 확인할 수 있도록 하였다. 또 [그림 8]과 같이 평가자의 구체적인 작업 내용을 개별적으로 검토할 수 있도록 하였다.

[그림 6] 관리자 기능의 평가자별 진척 상황 관리 화면



[그림 7] 관리자 기능의 평가 문서별 작업 상태 관리 화면

[그림 8] 관리자 기능의 평가자 작업 내용 확인 및 수정 화면

나. 평가 도구의 실제 적용과 관리

자체 개발된 평가 도구는 다음과 같은 방법으로 실제 평가 작업에 적용하였다.

먼저 평가자 개인별로 부여된 ID와 비밀번호로 접속하면 홈(home) 화면으로 진입하여 [그림 9]와 같이 평가 작업과 관련된 공지 사항 확인과 사업의 개요 및 목적을 확인할 수 있도록 하였다.

또한 평가 도구를 운영하는 중에 평가자들로부터 사용성 개선에 대한 피드백을 받아 지속적으로 기능에 대한 개선을 진행하였고 [그림 9]와 같이 홈페이지 공지를 통해 상세 안내를 제공하였다.

[그림 9] 평가 도구의 공지 사항과 프로젝트 개요 설명 화면

전체 메뉴는 작업 현황을 확인할 수 있는 '내 작업'과 실제 텍스트 레이블링 실행 화면인 '현재 작업' 두 가지로 단순하게 구성하였는데, '내 작업' 메뉴에서는 [그림 10]과 같이 나에게 배정된 문서의 수량과 내가 수행한 작업 수량, 남은 작업 수량을 확인할 수 있도록 하였고, 문서별로 포함되어 있는 발화의 개수 확인, 작업 상태 확인이 가능하도록 하였다.

그리고 발화의 수량에 따른 문서 정렬과 정렬된 문서 순서에 따른 순차적 작업 진행이 가능하도록 하여 평가자들이 원하는 방식에 따라 작업의 순서가 조정 가능하도록 하였다.

[그림 10] 평가 도구의 평가자 작업 현황 화면

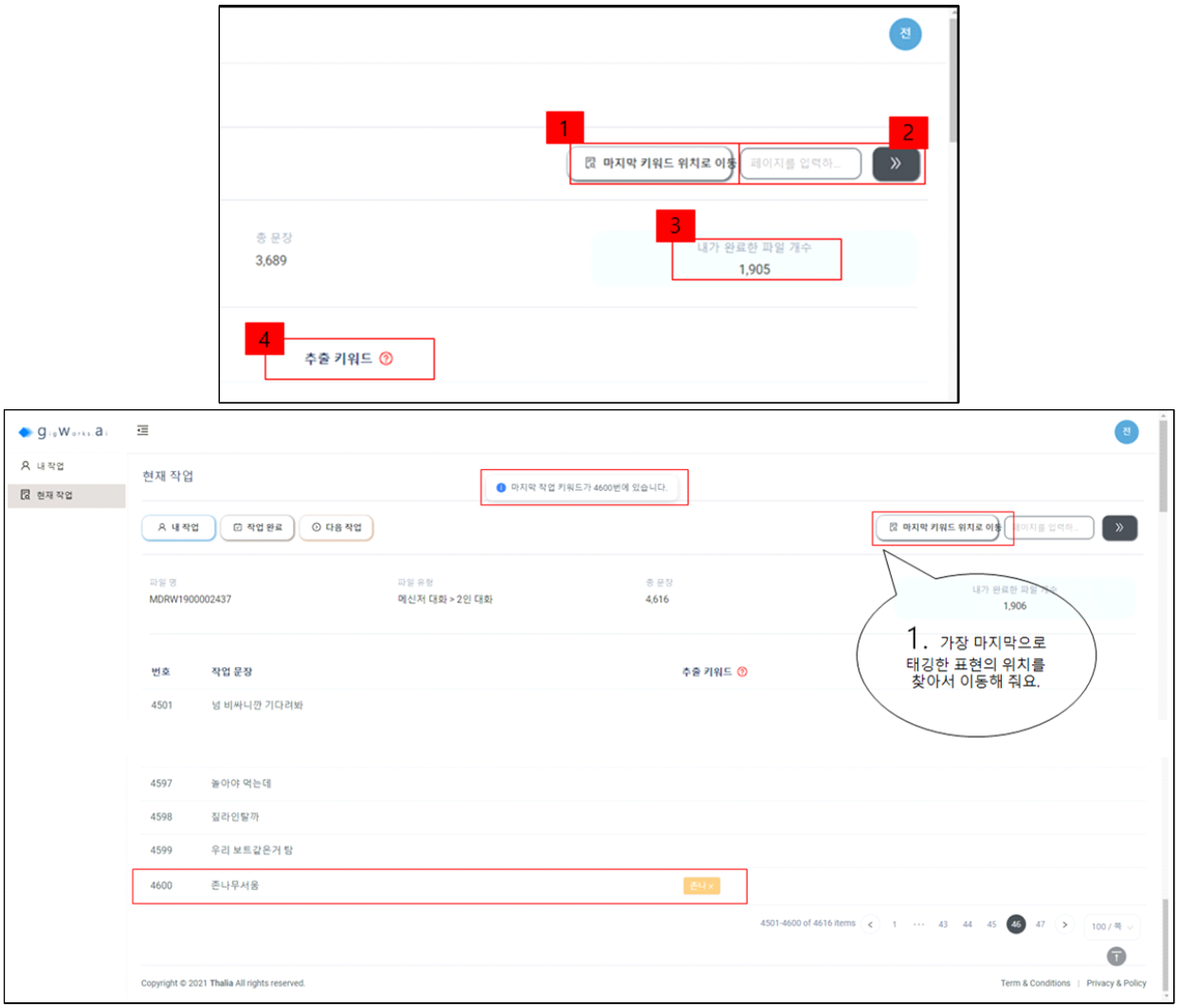
비윤리적 표현 및 개인정보 선별, 선별된 항목의 유형 분류는 별도의 키보드 사용 없이 마우스만으로도 진행이 가능하도록 하였다. 주관 기관이 제공하는 원시 말뭉치는 띄어쓰기가 되어 있지 않은 경우도 포함하고 있다는 점을 고려하여 [그림 11]과 같이 문제 표현을 찾으면 이를 드래그 하여 선택할 수 있도록 하였다. 그 후 마우스 우 클릭을 하여 유형 선택이 가능하도록 하였다.

[그림 11] 레이블링 진행 화면

평가자가 진행한 레이블링은 [그림 12]와 같이 유형에 따라 색깔별로 구분하여 작업 화면에 보이도록 하였고 평가자가 이를 확인하여 잘못이 있을 경우 ‘×’ 표시를 눌러 취소가 가능하도록 하였다.

[그림 12] 레이블링 선택 및 결과 반영 화면

또 아래 [그림 13]과 같이 마지막으로 태깅한 단어의 위치로 바로 이동할 수 있는 기능을 추가하여 작업을 잠시 중단하는 경우에도 빠르고 편하게 다시 시작할 수 있도록 하였다.



[그림 13] 평가 도구 기능 개선 및 개선 사항에 대한 설명 예시

평가자들의 실제 평가 기능과 개발한 관리자 기능을 통해 평가자별 작업 상황을 확인하고 이를 진도 관리와 평가자 경질과 결원에 대한 추가 선발에 활용하였다.

그리고 [그림 14]와 같이 평가자별 진도 관리를 위한 진행 현황 확인이 가능하도록 하였다. 특히 평가자별 진도 관리에서는 문서 개수에 따른 진행 현황뿐만 아니라, 작업자에게 배정된 발화의 수량에 따른 진척도 또한 확인이 가능하도록 하였다.

[그림 14] 관리자 기능의 평가자별 작업 진행률 확인 화면

관리자 기능의 진행 현황 확인 기능을 활용하여 확실한 작업 관리가 가능하도록 하기 위해 전체 조사 작업 분량을 1, 2차로 나누어 진행하였고 진행 현황 확인 기능을 활용하여 평가자별, 문서별 작업 진척 상황에 대한 관리를 실시간으로 지속하였다.

먼저 1차 조사 기간 동안 전체 작업량 50% 이상을 작업 완료하였으며 개인 사정 등의 사유로 8명의 중도 포기자가 발생하였다. 1차 조사 기간 중 7일 이상 작업 진척이 전혀 없는 불성실 평가자와 중도 포기자를 포함한 10인과 1차 조사 기간이 마감된 이후에도 35% 미만으로 작업한 3인을 경질하고 평가 데이터를 이관하여 평가자 추가 모집으로 결원을 보충하여 2차 조사 기간 완료까지 필요한 모든 작업을 완료하였다.

[그림 15] 관리자 기능의 평가자 평가 내용 확인 및 수정 화면

관리자는 개별 평가자들의 작업 내용에 대한 검수와 수정이 가능하도록 하였는데 [그림 15]와 같이 관리자 기능에서 평가자의 개별 평가 결과를 확인하고 수정이 가능하도록 하였다.

평가 도구를 통해 도출된 개별 데이터는 연구진들에게 제공되었고 이를 기초로 유형별 출현 빈도와 유형별 출현 비율을 종합적으로 분석하여 말뭉치 언어의 현황과 사회적 인식을 도출하는 데 활용하였다.

제 3 장

조사 진행 과정 및 내용

3. 조사 진행 과정 및 내용

3.1 사회적 인식 조사 평가 계획 및 일정 수립

말뭉치의 언어의 사회적 인식 조사의 다음 단계는 조사 평가 계획과 일정을 수립하는 것이다. 조사는 다음 절차에 따라 진행하였으며, 조사 기간 중 상시 모니터링을 진행하고 필요에 따라 평가자들에 대한 작업 내용 재교육이 이루어졌다. 과제 수행 단계를 요목화하면 다음과 같다.

① 사회적 인식 조사 평가 계획 수립
② 평가자 모집 기준 수립
③ 평가 진행 일정 수립
④ 평가자 모집
⑤ 평가자 교육 및 가이드라인 배포
⑥ 1차 조사 실시
⑦ 1차 조사 결과 수합
⑧ 1차 조사 결과 검수
⑨ 2차 조사 실시
⑩ 2차 조사 결과 수합
⑪ 2차 조사 결과 검수
⑫ 최종 조사 결과 분석
⑬ 데이터 가공 및 납품

3.1.1 사회적 인식 조사 평가자 선발

말뭉치 언어에 대한 사회적 인식 조사를 위해서는 한국어를 사용하는 일반 언어 사용자들을 최대한 편향 없이 모집할 필요가 있었다. 본 연구진이 언어의 사용과 인식에 영향을 줄 수 있는 요인으로 꼽은 것은 '성별'과 '연령대', 그리고 '거주 지역'이었다. 이러한 변수들에 따른 사회적 인식을 그대로 드러날 수 있도록 행정안전부가 2021년 5월 발표한 주민 등록 인구 통계를 반영하여 성별, 연

령별, 지역별 인구 비율을 반영하는 평가자 모집 기준을 수립하였다. 그 구체적인 기준은 다음과 같다.

○ 평가자의 성별, 연령별, 지역별 구성 비율, 인원수 등을 고려하여 조사 평가자를 모집한다.

○ 평가자 모집 일반 기준

구분	기준
모집단	· 행정안전부, 2021년 5월 주민등록 인구 통계
변수	· 성별 : 남자/여자 · 연령대 : 10대/20대/30대/40대 이상 · 지역 : 수도권/영남권/충청권/호남권/강원·제주권
할당	· 성별/연령대별/지역별 할당 표본 추출
기타 고려 사항	· 성별/연령별 현실적인 모집 가능성 고려

〈표 6〉 평가자 모집 일반 기준

○ 평가자 모집 비율 및 모집 현황

단위 (명)	남성				여성				합계
	10대	20대	30대	40대 이상	10대	20대	30대	40대 이상	
강원 제주권	-	1	-	-	-	1	-	-	2
수도권	4	7	7	4	5	11	9	6	53
영남권	2	3	3	2	3	5	6	3	27
충청권	-	1	1	1	-	2	2	1	8
호남권	1	1	1	1	1	2	2	1	10
합계	7	13	12	8	9	21	19	11	100

〈표 7〉 평가자 모집 비율 및 현황

3.1.2 사회적 인식 조사 평가 진행 일정 수립

○ 평가 진행 일정 계획표

공정	추진일정(월 단위)			
세부 일정	6월	7월	8월	9월
평가자 모집				
작업 지침 교육				
평가 데이터 수집 및 관리 (1차)				
조사 결과 정리 및 검토 보완 (1차)				
평가 데이터 수집 및 관리 (2차)				
조사 결과 정리 및 검토 보완 (2차)				
조사 결과 종합 및 분석				

〈표 8〉 평가 진행 일정 계획표

○ 평가 진행 세부 일정

- 피험자 모집 및 작업자 교육: 2021년 6월 20일 ~ 6월 30일
- 1차 조사: 2021년 7월 1일 ~ 7월 29일(4주간)
- 1차 조사 결과 정리 및 검토: 2021년 7월 30일 ~ 7월 31일
- 2차 조사: 2021년 8월 1일 ~ 8월 29일(4주간)
- 2차 조사 결과 정리 및 검토: 2021년 8월 30일 ~ 8월 31일
- 결과 종합 및 분석: 2021년 9월 1일 ~ 9월 30일

3.2 사회적 인식 조사 평가자 모집 및 평가 가이드라인 배포

3.2.1 평가자 모집

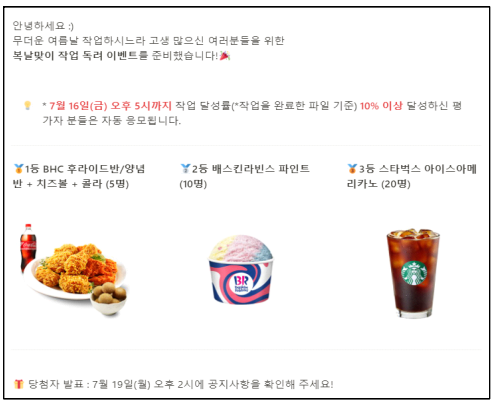
100명의 평가자를 모집하기 위해서는 정확한 작업 내용을 효과적으로 홍보해야할 필요가 있었다. 미디어 코퍼스가 자체적으로 보유한 홍보 채널을 활용하여 보다 많은 크라우드 워커들에게 조사의 목적 및 의의에 대해 홍보하였다.

[그림 16] 사업 홍보 홈페이지

[그림 17] 사업 소개

또한 정해진 사업 기간 내에 많은 양의 반복 작업이 이루어져야 했기 때문에 실제 조사 기간이 시작된 이후에도 작업자들의 효율적인 작업 진척을 위해 감독과 독려가 필요하였다. 작업량에 따른 급여 지급 이외에도 다양한 이벤트를 운영하여 평가자들의 작업을 독려하였다.

[그림 18] 사업 홍보 이벤트



[그림 19] 작업 독려 이벤트

연구에 대한 전반적인 설계와 완료된 후 2021년 6월 22일에서 6월 25일까지 3일간 평가자 모집을 진행하였고 해당 기간 동안 총 669명이 지원하였다. 신청 접수 마감 후 6월 27일까지 이틀간 성별, 연령, 지역의 변인 비율을 고려하여 2배수(200명)을 우선 선발하였고 1차 참여 확인 과정을 거쳤다. 참여 미확정 인원 및 조건 미달 지원자를 제외한 후 우선 선발된 2배수 내에서 2차 참여 절차 이후 총 100명의 평가자를 확정하였다.

확정된 참여자에게는 참여 안내 이메일을 통해 연구 참여에 대한 전반적인 안내를 제공함과 동시에 평가 자료와 결과에 대한 보안 각서, 연구 참여를 위한 개인정보 수집 동의서, 비용 지급을 위한 개인 서류 등을 요청하여 최종적으로

연구 참여에 대한 동의를 구하였다. 이와 더불어 사전 교육 참여 시간을 조사하여 3회에 걸쳐 실시간 사전 교육을 진행하였으며 불참 평가자에게는 동영상 교육 내용을 전달하여 평가 절차를 숙지하도록 하였다.

3.2.2 평가자 평가 가이드라인 배포

평가자들이 쉽고 빠르게 숙련될 수 있도록 비윤리적 표현 검토 및 평가를 위한 도구 사용 가이드라인과 이를 효율적으로 배포할 수 있도록 노션(notion) 사이트를 개설하였다. 그리고 이를 통해 가이드라인을 배포하는 한편 평가자들과 작업 진행과 관련된 사항을 공유하였다.

[그림 20] 작업 가이드 배포용 사이트

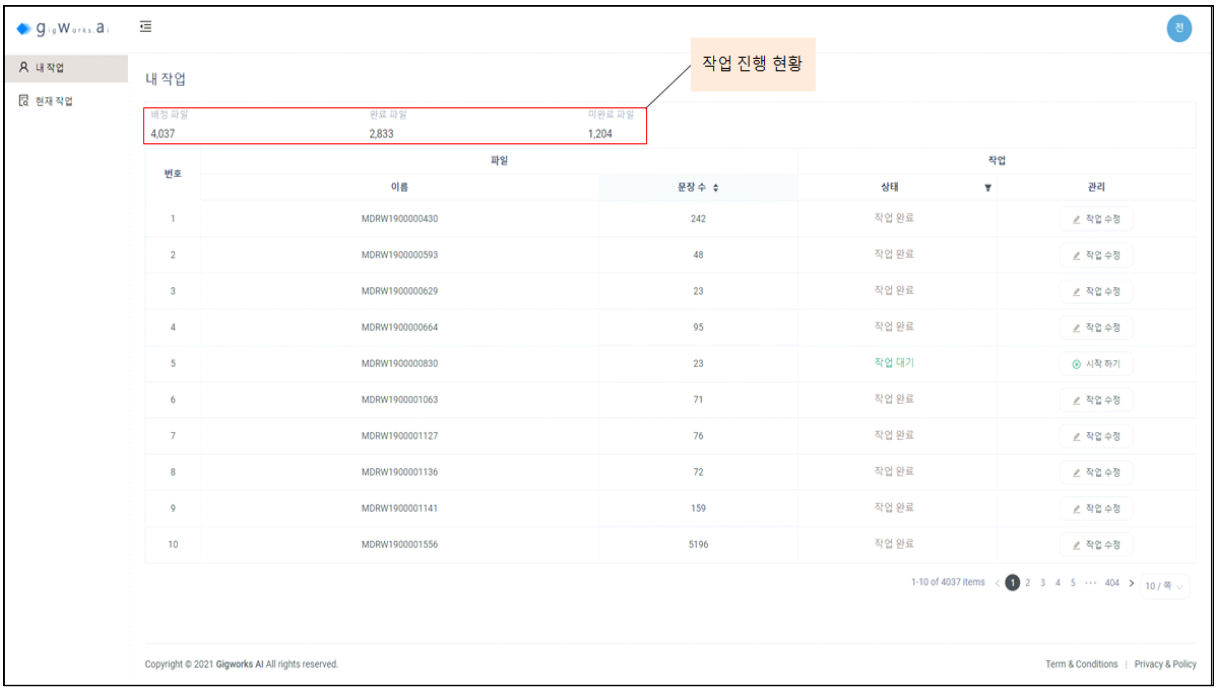
3.3 사회적 인식 조사 평가 실행 및 진행 교육

3.3.1 조사 평가 실시

미디어 코퍼스와 중앙대학교 인문콘텐츠연구소는 협업을 통해 웹 기반의 평가 도구를 자체 개발하였다. 웹 브라우저에서 로그인만 하면 바로 사용할 수 있는 형태로 특정 운영 체제나 작업 환경의 영향을 받지 않고 언제 어디에서나 효율적인 조사 작업이 가능하도록 하였다. 평가자의 평가와 작업 관리 모두 해당 도구를 활용하여 실시되었으며 아래는 그 예시이다.

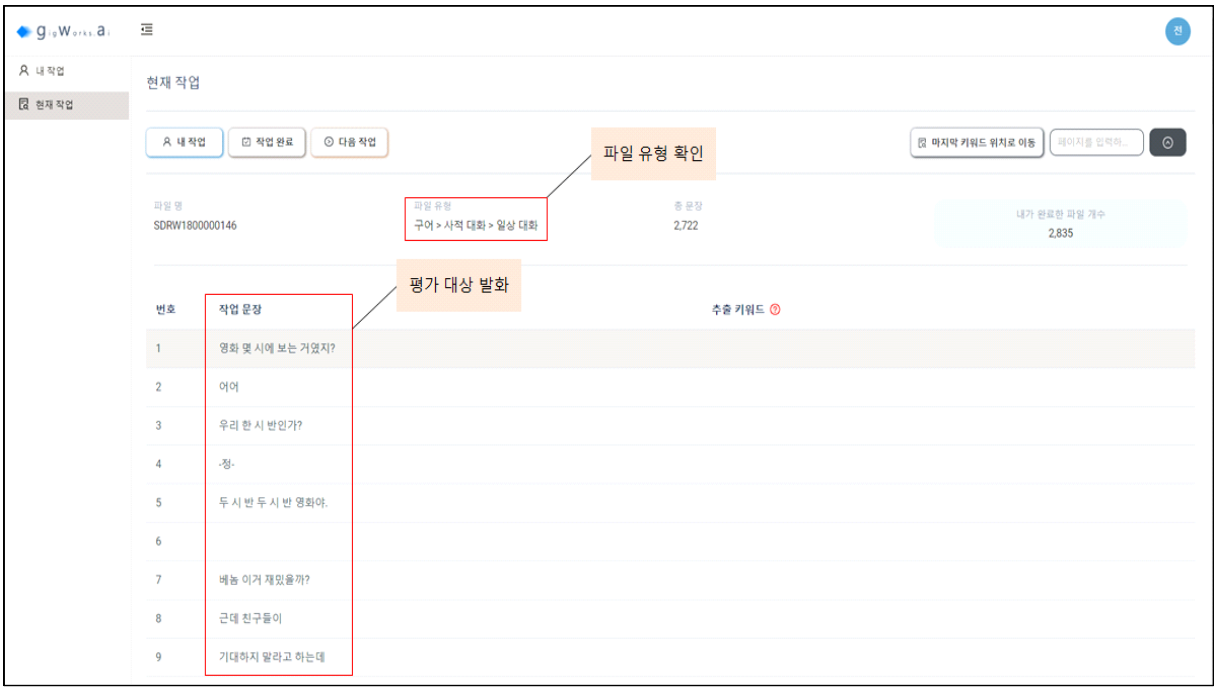
[그림 21]은 평가 도구에 접속하면 바로 볼 수 있는 메인 화면이며 작업 공지사항과 정보를 확인할 수 있도록 하였다. [그림 22]는 각 평가자별로 자신에게 할당된 작업의 현황을 확인할 수 있는 조회 화면이다. [그림 23]과 [그림 24]는 평가 대상 문장을 확인하고 비윤리적 표현을 태깅하는 실제 작업 과정을 담고 있다.

[그림 21] 평가 도구 메인화면



[그림 22] 평가 도구 작업 현황 조회 화면

[그림 23] 평가 도구 활용 화면 1



[그림 24] 평가 도구 활용 화면 2

3.3.2 조사 평가자 진행 교육 및 관리

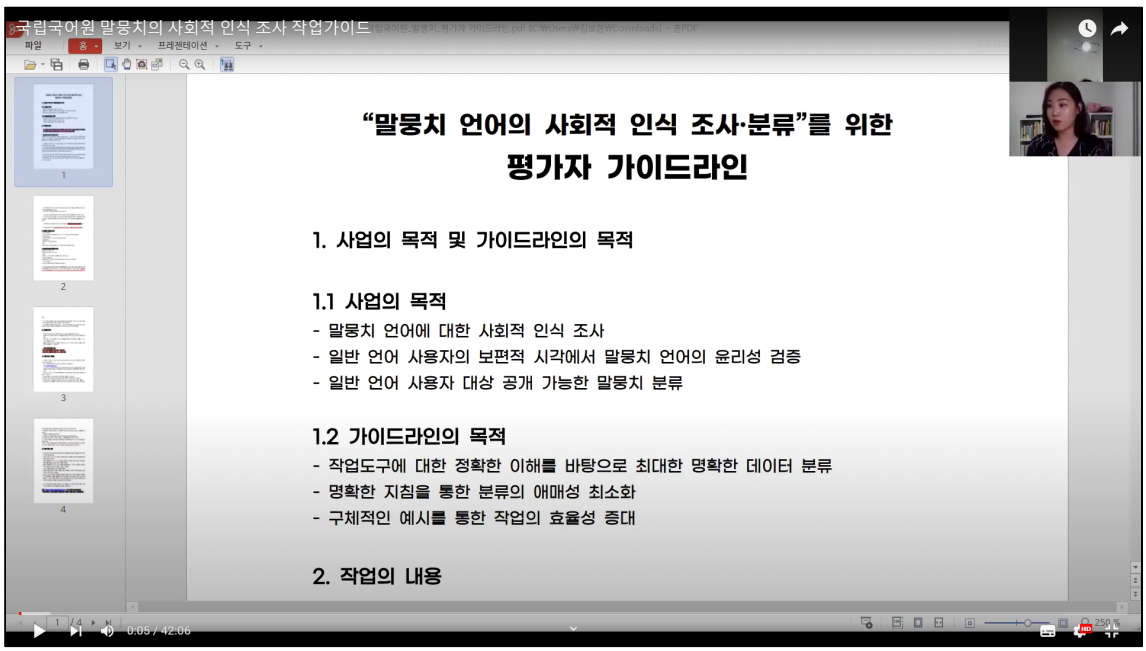
원활한 조사 진행을 위해서는 100명의 평가자 모두에게 빠른 시간 내에 정확하게 작업할 수 있도록 평가 절차를 숙지하도록 하는 것이 중요하다. 또 앞서 설정한 평가자 모집 기준에 따라 선발된 100명은 성별, 연령, 지역이 모두 고르게 분포되어 있어 생활양식이 모두 다르고 그에 따라 온라인 교육에 참여하기에 용이한 시간도 제각각이었다.

따라서 모든 평가자가 교육에 참가할 수 있도록 서로 다른 요일, 서로 다른 시간대에 총 3회에 걸쳐 실시간 온라인 교육을 실시하였다. 부득이하게 실시간 교육에 참여하기 어려운 평가자들을 위해서는 같은 교육 내용을 동영상으로 만들어 시청하도록 하고 가이드라인 파일을 별도로 제공하였다. 다음은 사전 교육의 세부적인 내용이다.

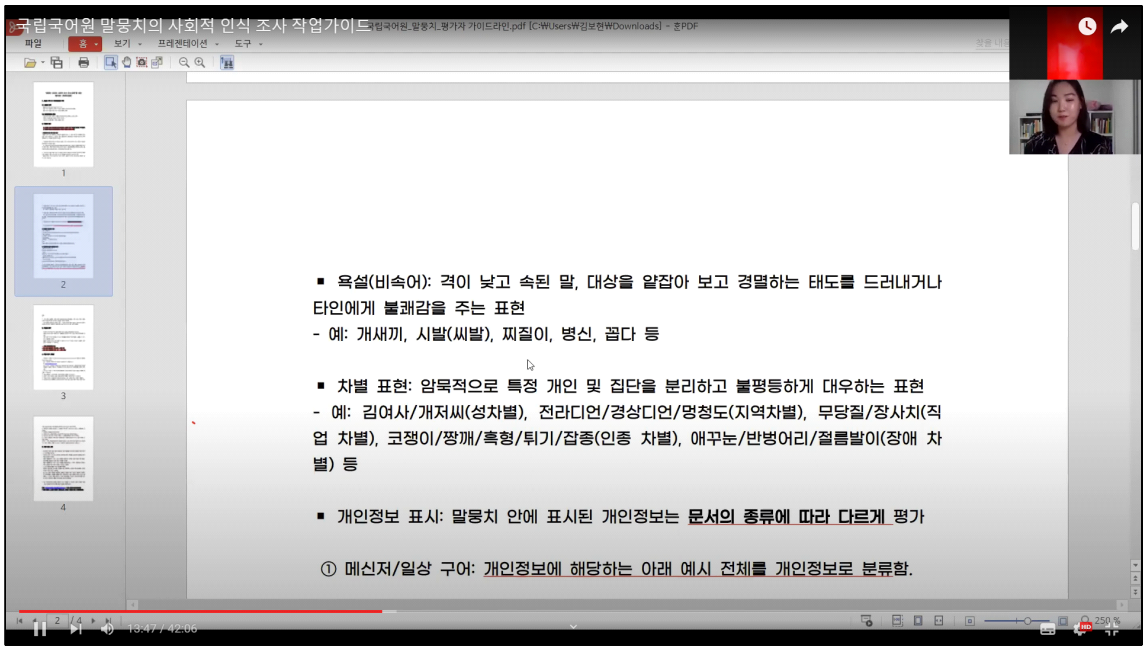
○ 100명의 평가자에게 3회에 걸쳐 작업 지침을 교육하였다(회당 30여 명). 교육은 온라인 화상 교육(Zoom 활용)을 원칙으로 하며 결석자에게 작업 지침 및 교육 동영상을 제공하였다. 사업 수행 기간 중 발생할 수 있는 문제들을 해결하고 평가자들과 소통할 수 있는 상시 대응팀을 운영하였다. 효율적 운

영을 위하여 평가자를 총 10팀으로 나누고 각 팀의 조장을 선발하여 원활한 작업 관리가 이루어질 수 있도록 하였다.

○ 정기적인 작업 진행 상황 점검으로 평가 품질을 유지하며 필요한 경우 평가자 재교육을 실시하였다. 불성실 평가자에 대해서는 1회 경고하고 이후에도 계속 작업 진척이 지연되면 평가자를 경질하고 진행 중인 작업을 인계하도록 하였다. 추가 모집은 초기 참여자 모집과 같은 방식으로 미디어 코퍼스의 프로젝트 홍보 홈페이지를 활용하여 8월 5일부터 8월 9일까지 5일간 진행되었으며 총 99명이 참여를 신청하였고 경질된 평가자 변인을 고려하여 최종 13인을 선발하였다. 추가 선발된 평가자에게도 참여 안내 및 제반 서류를 요청하여 최종적으로 연구 참여에 동의하는 절차를 거쳤으며, 사전 교육 영상을 배포하는 것으로 진행하였다.



[그림 25] 평가자 교육 영상 캡처 화면 1



[그림 26] 평가자 교육 영상 캡처 화면 2

3.4 조사 현황 관리

가. 1차 조사 진행 및 수합 현황

○ 1차 조사 결과 수합일: 2021년 8월 1일

○ 1차 조사 기간: 2021년 7월 5일 ~ 8월 1일

○ 1차 조사 목표 작업량 달성

- 전체 407,498개 파일 중 203,749개, 25,190,902 어절 중 1,500만 어절 이상 (전체 작업량의 50% 이상) 달성하였다.
- 중간 결과물을 국립국어원에 전달하였다.

나. 1차 조사 이후 평가자 결원 보충

- 조사 기간 중 작업 진행 상황을 수시로 모니터링하였다.
- 개인적 사정 등의 사유로 8명의 중도 포기자가 발생하였으며 이후 불성실 평가자 2명의 경질과 함께 추가 모집으로 결원을 보충하였다.
- 작업 과정 중 발생한 문의 사항들에 안내와 재교육을 진행하였다.
- 1차 조사 기간 중 7일 이상 작업 진척이 전혀 없거나 중도 포기한 10명

을 해임 후 재모집하였다.

- 1차 조사 기간이 마감된 이후에도 35% 이상 작업하지 못한 3명을 해임 후 재모집하였다.
- 재모집한 작업자들에게 잔여 작업량을 재배분하여 당초 계획대로 작업 진행하였다.

작업자	성별	나이	지역			미완료	완료	총	진행률	미완료	완료	총	진행률
			출신	성장	거주								
변 은	여자	20	서울	서울	서울	0	0	0	0.0%	0	0	0	0.0%
이 민	남자	32	경기	경기	경기	0	0	0	0.0%	0	0	0	0.0%
이 영	여자	51	전북	경기	경기	0	0	0	0.0%	0	0	0	0.0%
이 희	여자	29	강원	강원	강원	0	0	0	0.0%	0	0	0	0.0%
이 나	여자	36	경기	경기	경기	0	0	0	0.0%	0	0	0	0.0%
이 은	여자	26	경북	경북	경북	0	0	0	0.0%	0	0	0	0.0%
정 숙	여자	52	경기	경기	경기	0	0	0	0.0%	0	0	0	0.0%
최 실	여자	54	서울	경기	경기	0	0	0	0.0%	0	0	0	0.0%
윤 옥	여자	55	경기	경기	경기	0	0	0	0.0%	0	0	0	0.0%
김 우	여자	18	서울	서울	서울	0	0	0	0.0%	0	0	0	0.0%
강 호	남자	25	서울	서울	서울	2602	1412	4014	35.2%	58058	2855	60913	4.7%
김 영	남자	16	광주	광주	광주	4002	34	4036	0.8%	44648	16264	60912	26.7%
임 묵	남자	41	부산	부산	부산	3999	38	4037	0.9%	41179	19733	60912	32.4%

〈표 9〉 평가자 재모집 사례

다. 2차 조사 진행 및 수합 현황

○ 2차 조사 결과 수합일 : 2021년 09월 05일

○ 2차 조사 기간: 8월 3일 ~ 9월 5일

○ 조사 목표 작업량 달성

- 407,498개 문서, 25,190,902 어절에 대한 조사를 완료하였다.

라. 2차 조사 결과 보완

○ 2차 조사 결과 검토

- 데이터의 정확성 검토를 위해 언어 및 철학 전공자를 검토자를 지정하여 이들을 대상으로 검토자용 가이드라인을 배포하고 교육을 실시하였다.

○ 2차 조사 결과 자문/검토에 따른 작업 주요 내용

- 하나의 문서 전체에 아무런 태그도 표시되지 않은 경우 실제로 문제가 되는 표현이 없는지 검토 후 정제 수준을 결정하도록 한다.
- 태그된 비윤리적 표현 내에 불필요한 공백이 삽입된 경우 이를 수정하도록 한다.
- 이미 필터링된 개인정보가 중복 태그된 경우 (예를 들어, &name) 이를 수정하도록 한다.
- 비윤리적 표현으로 분류가 애매한 표현에 태그된 경우 검토 의견을 기술한다.

3.5 최종 조사 결과 분류 및 데이터 납품

3.5.1 최종 조사 결과 분류

총 조사 대상 데이터는 407,498개 문서, 약 25,190,902 어절로 메신저 말뭉치 6,712,968 어절, 웹 말뭉치 9,766,838 어절, 구어 말뭉치 8,711,096 어절로 구성되어 있었으며 이에 대한 분류 항목별 검토와 문서별 검토를 진행하였다.

정제 대상 내용 및 표현으로는 혐오 표현, 성적 수치심을 일으킬 수 있는 표현, 욕설, 차별 표현, 기타 문제가 될 소지가 있는 표현, 개인정보 등 6가지 범주를 설정하여 제시하였다. 그리고 거주 지역・성별・연령 변수가 고르게 분포된 100명의 일반 언어 사용자를 평가자로 하여 조사 대상 데이터에 정제 대상 내용 및 표현이 포함 여부를 조사 완료하였다.

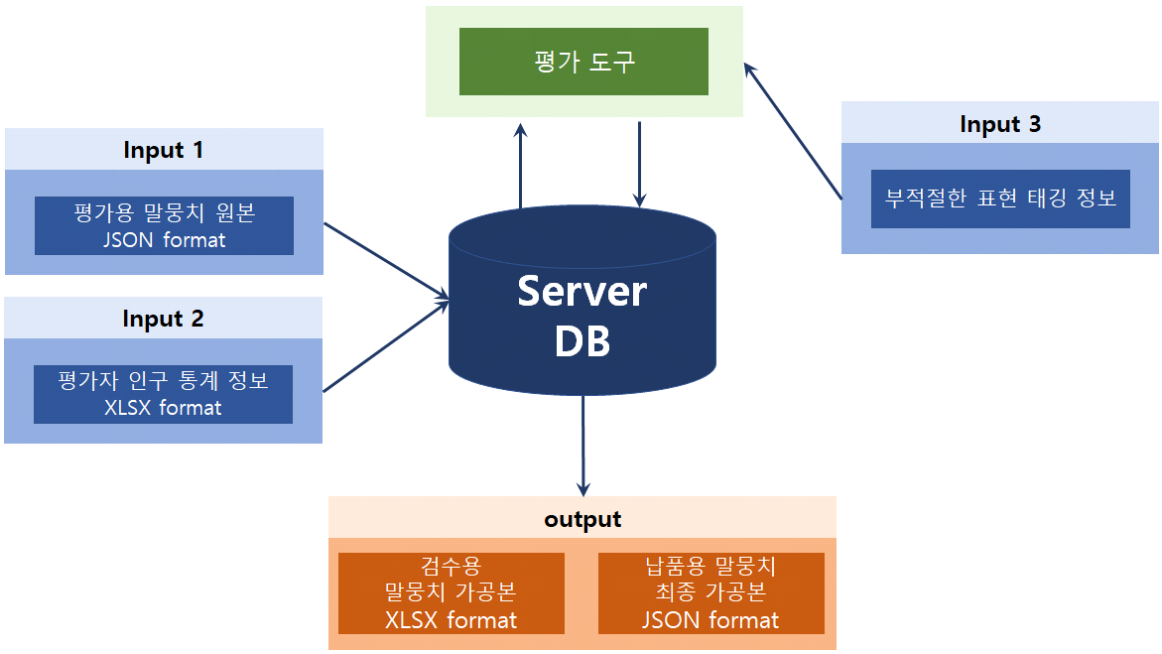
최종 검수 과정을 거친 후 JSON 및 엑셀 파일 형식의 조사 결과 데이터를 완성하였다.

3.5.2 최종 조사 결과 데이터 가공 및 납품

국립국어원에서 평가용으로 제공한 자료는 평가 결과를 반영하여 국립국어원에서 제시한 기준 형식에 맞추어 가공한 후 납품하였다.

최초 자료 제공 단계에서부터 최종 산출물 생성까지의 단계를 도식화하면 [그

림 27]과 같다.



[그림 27] 데이터 입력과 산출물 생성

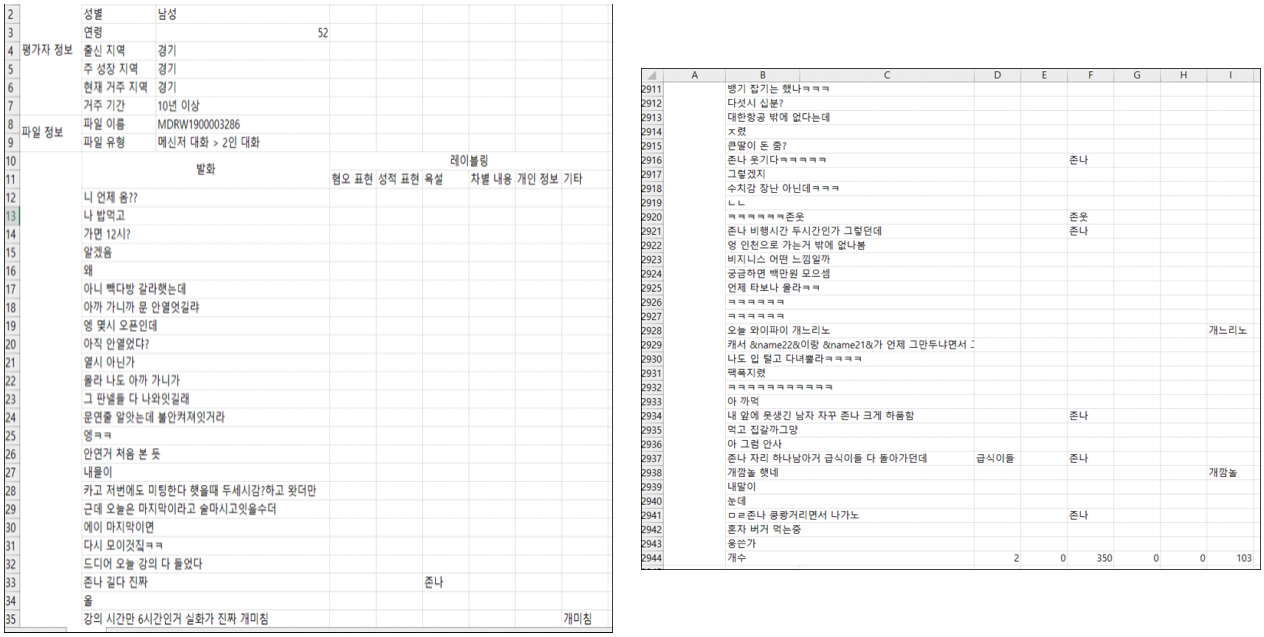
먼저 국립국어원으로부터 평가용 자료 원본을 제공받아 이를 평가 도구 서버 DB에 입력하였다. DB 입력과 파싱 단계 이후, 평가자들이 문서를 읽고 부적절 표현을 찾아서 태깅하는 것과는 무관한 JSON의 속성자(property)와 메타 정보는 평가자들이 보게 될 문서에는 반영되지 않도록 처리하였다.

그리고 평가 대상 3종의 말뭉치를 분석한 결과, 웹 말뭉치의 경우에는 구어 대화, 메신저 대화와 동일한 단위임에도 불구하고 긴 어절로 이루어진 경우가 다수를 차지하였다. 이를 별도로 분할하지 않고 평가자들에게 그대로 제공할 경우 평가자별로 할당받게 될 어절 수량의 불균형이 발생할 가능성이 있고, 평가자가 읽어야 할 단위가 길어짐에 따라 독해 부담량이 커질 수도 있다는 점을 고려하여 문장 부호를 기준으로 분할된 단위로 문서의 행을 구성하도록 하였다.

평가자 모집을 진행하면서 평가자의 성별, 연령, 출생, 주 성장지, 거주 지역 정보를 수집하여 평가 도구 서버 DB에 입력하였고, 해당 정보는 평가 도구의 관리자 기능에서 평가자별 평가 진척 상황을 확인할 때 확인이 가능하도록 하였다.

이후 평가자들이 평가 도구를 통해 문서별 이루어진 태깅 정보는 태깅 항목, 항목별 유형, 태그 위치 색인 정보(begin/end index)로 세부 분류하여 서버 DB에 입력하였다.

서버 DB에 저장된 평가용 말뭉치 원본, 평가자 인구 통계 정보, 태깅 정보는 산출물 생성 단계에서 중간 점검을 위한 XLSX 형식과 최종 납품을 위한 JSON 형식으로 가공이 이루어졌다.



2		성별	남성						
3		연령		52					
4	평가자 정보	출신 지역	경기						
5		주 성장 지역	경기						
6		현재 거주 지역	경기						
7		거주 기간	10년 이상						
8	파일 정보	파일 이름	MDRW1900003286						
9		파일 유형	메신저 대화 > 2인 대화						
10		발화		레이블링					
11				혐오 표현	성적 표현	욕설	차별 내용	개인 정보	기타
12		니 언제 옴??							
13		나 밥먹고							
14		가면 12시?							
15		알겠음							
16		왜							
17		아니 백다방 갈라했는데							
18		아까 가니까 문 안열엇길랴							
19		엥 몇시 오픈인데							
20		아직 안열었다?							
21		엩시 아닌가							
22		몰라 나도 아까 가니가							
23		그 판넬들 다 나와잇길래							
24		문연줄 알앗는데 불안켜져잇거라							
25		엥ㅋㅋ							
26		안연거 처음 본 듯							
27		내몰이							
28		카고 저번에도 미팅한다 했을때 두세시감?하고 왓더만							
29		근데 오늘은 마지막이라고 술마시고잇을수더							
30		에이 마지막이면							
31		다시 모이것짘ㅋㅋ							
32		드디어 오늘 강의 다 들었다							
33		존나 길다 진짜				존나			
34		올							
35		강의 시간만 6시간인거 실화가 진짜 개미침							개미침

	A	B	C	D	E	F	G	H	I
2911		뱅기 잡기는 했나ㅋㅋㅋ							
2912		다섯시 십분?							
2913		대한항공 밖에 없다는데							
2914		ㅈ했							
2915		큰딸이 돈 줌?							
2916		존나 웃기다ㅋㅋㅋㅋㅋ				존나			
2917		그럴겠지							
2918		수치감 장난 아닌데ㅋㅋㅋ							
2919		ㄴㄴ							
2920		ㅋㅋㅋㅋㅋㅋ존웃				존웃			
2921		존나 비행시간 두시간인가 그럴던데				존나			
2922		엉 인천으로 가는거 밖에 없나봄							
2923		비지니스 어떤 느낌일까							
2924		궁금하면 백만원 모으셈							
2925		언제 타보나 몰라ㅋㅋ							
2926		ㅋㅋㅋㅋㅋㅋ							
2927		ㅋㅋㅋㅋㅋㅋ							
2928		오늘 와이파이 개느리노							개느리노
2929		캐서 &name22&이랑 &name21&가 언제 그만두냐면서							
2930		나도 입 털고 다녀볼라ㅋㅋㅋㅋ							
2931		땍쪽지했							
2932		ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ							
2933		아 까먹							
2934		내 앞에 뭇생긴 남자 자꾸 존나 크게 하품함				존나			
2935		먹고 집갈까그랑							
2936		아 그럼 안사							
2937		존나 자리 하나님아거 급식이들 다 돌아가던데		급식이들		존나			
2938		개깜놀 햇네							개깜놀
2939		내말이							
2940		눋데							
2941		ㅁㄹ존나 뭉뭉거리면서 나가노				존나			
2942		혼자 버거 먹는중							
2943		옹쓴가							
2944		개수		2	0	350	0	0	103

[그림 28] XLSX 형식 산출물 예시

XLSX 형식의 산출물은 [그림 28]과 같이 평가자의 인구 통계 정보와 평가 문서 내용, 태깅 항목과 유형별 빈도 등을 확인할 수 있도록 구성하여 평가자들이 수행한 태깅 내용을 점검하고 유형별 출현 빈도를 파악하기 위한 용도로 생성하였다.

최종 산출물은 국립국어원의 요청을 반영하여 모두의 말뭉치 형식을 토대로 JSON 형식으로 생성하였다. 서버 DB에 저장된 평가 대상 문서 정보, 평가자 정보, 평가 원문 정보를 포함하여 태깅이 이루어진 항목의 유형 정보와 태그 위치 색인 정보를 포함하도록 하였다.

JSON 형식의 최종 산출물 생성 단계에서 평가자들에게 작업을 균등하게 할당하기 위해서 문장 부호를 단위로 분할했던 웹 유형의 평가 대상 자료는 국립국어원이 최초에 제공한 웹 말뭉치 원문과 1:1 매칭하여 분할 이전의 결합 형태로 다시 복원이 이루어졌다.

평가 문서 메타 정보
id/metadag

문서 및 평가자 메타 정보
document/investigator

문장(발화) 및
부적절 항목 정보

[그림 29] 최종 산출물 JSON 형식 예시

제 4 장

조사 결과 분석

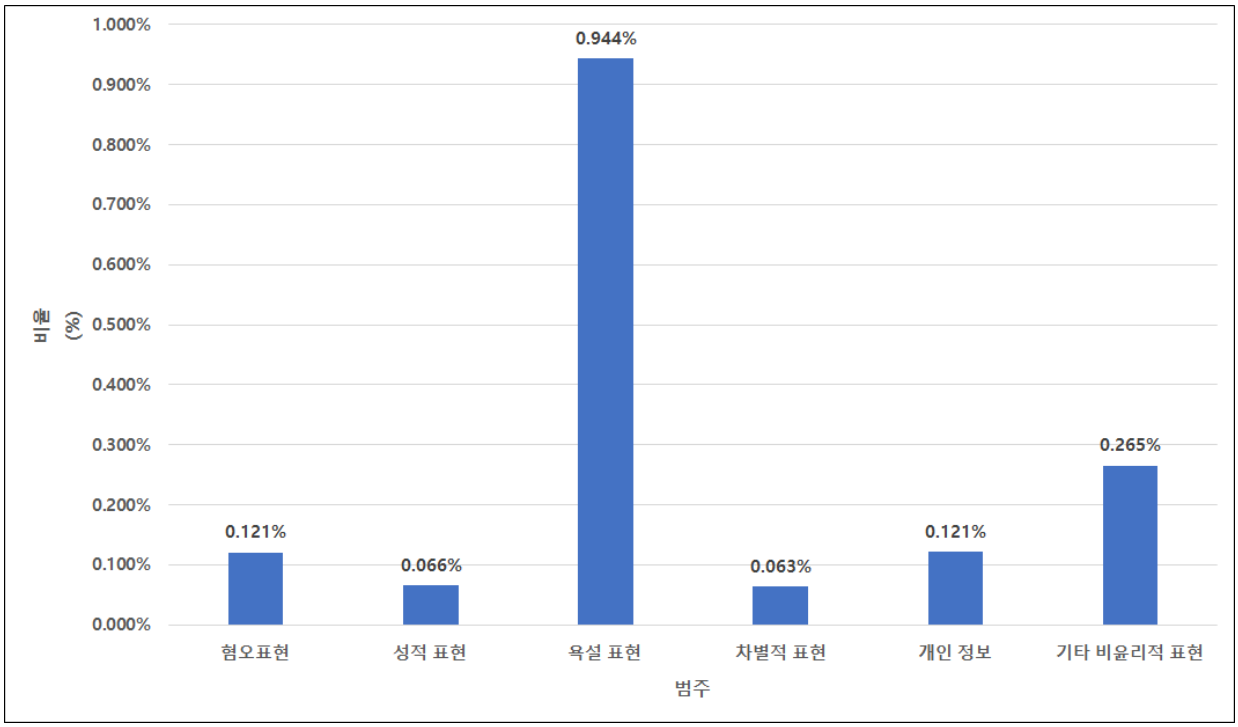
4. 조사 결과 분석

4.1 평가자의 변인별 비윤리적 표현 유형 빈도 및 비율 분석

성별, 연령, 지역, 직업 등 평가자 변인별로 비윤리적 표현 유형 태깅 빈도표와 이를 시각화한 그래프는 아래와 같다.

구분		발화 (개)	혐오 표현		성적표현		욕설 표현		차별적 표현		개인 정보		기타 비윤리적 표현	
			건	비율 (%)	건	비율 (%)	건	비율 (%)	건	비율 (%)	건	비율 (%)	건	비율 (%)
성별	남성	2,429,401	2,671	0.11	1,418	0.058	18,665	0.768	1,582	0.065	2,777	0.114	7,393	0.304
	여성	3,722,734	4,771	0.128	2,641	0.071	39,407	1.059	2,320	0.062	4,684	0.126	8,925	0.24
연령	10대	1,051,771	920	0.087	644	0.061	10,285	0.978	560	0.053	977	0.093	2,718	0.258
	20대	2,134,785	2,316	0.108	1,564	0.073	18,558	0.869	1,369	0.064	2,054	0.096	2,279	0.107
	30대	1,964,164	2,780	0.142	1,255	0.064	19,948	1.016	1,136	0.058	2,442	0.124	7,754	0.395
	40대 이상	1,001,415	1,426	0.281	596	0.118	9,281	1.851	837	0.16	1,988	0.404	3,567	0.722
지역	수도권	3,375,085	2,946	0.259	1,793	0.147	30,122	2.735	1,244	0.094	3,852	0.378	6,985	0.462
	영남권	1,603,453	2,490	0.582	1,223	0.296	18,022	4.472	1,007	0.235	1,948	0.493	5,492	1.301
	호남권	564,474	970	0.565	431	0.246	4,889	2.615	922	0.525	1,231	0.624	1,523	0.774
	충청권	487,299	903	0.674	521	0.384	4,102	2.737	685	0.524	408	0.309	2,023	0.933
	강원/ 제주권	121,824	133	0.109	91	0.075	937	0.769	44	0.036	22	0.018	295	0.242
직업	가정주부	365,474	326	0.089	194	0.053	3,055	0.836	117	0.032	373	0.102	1,018	0.279
	경영/ 관리직	182,737	294	0.161	136	0.074	2,025	1.108	44	0.024	1,104	0.604	750	0.41
	교육 종사자	365,472	499	0.137	228	0.062	2,494	0.682	252	0.069	100	0.027	1,356	0.371
	기술직 종사자	339,447	454	0.134	257	0.076	2,990	0.881	176	0.052	343	0.101	2,579	0.76
	무직	791,861	950	0.12	507	0.064	7,591	0.959	546	0.069	902	0.114	2,385	0.301
	사무 종사자	1,177,066	1,511	0.128	693	0.059	11,634	0.988	898	0.076	2,099	0.178	3,594	0.305
	서비스 종사자	121,825	47	0.039	33	0.027	486	0.399	82	0.067	30	0.025	44	0.036
	전문가	416,415	253	0.061	124	0.03	4,206	1.01	51	0.012	172	0.041	642	0.154
	판매 종사자	60,912	17	0.028	21	0.034	146	0.24	22	0.036	1	0.002	19	0.031
	학생	2,026,366	1,722	0.085	1,319	0.065	18,464	0.911	1,039	0.051	1,968	0.097	3,708	0.183
	기타	304,560	1,369	0.45	547	0.18	4,981	1.635	675	0.222	369	0.121	223	0.073

〈표 10〉 평가자 변인별 비윤리적 표현 유형 태깅 빈도



[그림 30] 전체 조사 대상 발화 중 비윤리적 표현 유형별 비율

평가자에게 할당된 총 6,152,135 발화의 1.58%인 97,254건을 대상으로 조사 분석을 시행하였다. 이 조사 대상 중에서 우리가 비윤리적 표현으로 구분한 6가지 범주(혐오 표현, 성적 표현, 욕설 표현, 차별적 표현, 개인정보, 기타 비윤리적 표현)에 따라 조사 결과로 도출된 건수와 비율은 다음과 같다. 혐오 표현은 7,442건으로 전체 조사 대상 대비 0.121%로 나타났다. 성적 표현은 4,059건으로 0.066%, 욕설은 58,072건으로 0.944%, 차별적 표현은 3,902건으로 0.063%, 개인정보는 7,461건으로 0.121%, 기타 비윤리적 표현은 16,318건으로 0.265%로 나타났다. 6개 모든 범주에서 조사 대상 발화 수 대비 검출 건수가 1%를 넘지는 않았으며, 6개 범주의 비율을 모두 합하면 1.581%에 해당한다.

국립국어원의 모두의 말뭉치에서 배포하고 있는 데이터 중 본 연구의 조사 대상 말뭉치는 그 방대한 수에 비해 비윤리적 표현에 해당하는 비율은 낮다. 그럼에도 불구하고 비윤리적 표현의 정제의 중요성은 여전히 요구된다. 이를테면 사회적으로 어떤 누구라도 '한 마디' 말실수를 저지를 경우, 그것이 특히 '사회적으로 용인되지 못하는' 비윤리적 표현이라면 그 사람이 아무리 평소에 좋은 말, 바른 말을 사용해 왔다고 하더라도 그 사람의 언행에 대해서 질타를 하고 잘못된 언행에 대한 책임은 추궁되기 마련이다. 우리가 비윤리적 표현으로 구분한 혐오,

성적, 욕설, 차별 등은 '단 한 건'의 언행이라도 실제로 어떤 사람이 행한다면 사회적으로 그 사람은 도덕적 비난의 대상이 될 수 있다. 위의 조사 대상에서 특히 '욕설 표현'이 다른 범주의 표현들보다 상대적으로 가장 많이 조사된 것을 눈여겨볼 필요가 있다. 욕설 표현은 58,072건으로 0.944%로 다른 모든 범주의 비윤리적 표현들보다 압도적으로 많았다. 이는 곧 모두의 말뭉치에 '욕설 표현'이 많이 포함되어 다는 추론을 가능하게 한다. 대중에게 공개되는 말뭉치 언어에서의 욕설 표현은 사회적으로 논란의 여지없이 즉각적인 비난의 대상이 되기 때문에 말뭉치 정제 과정에 걸러져야 한다.

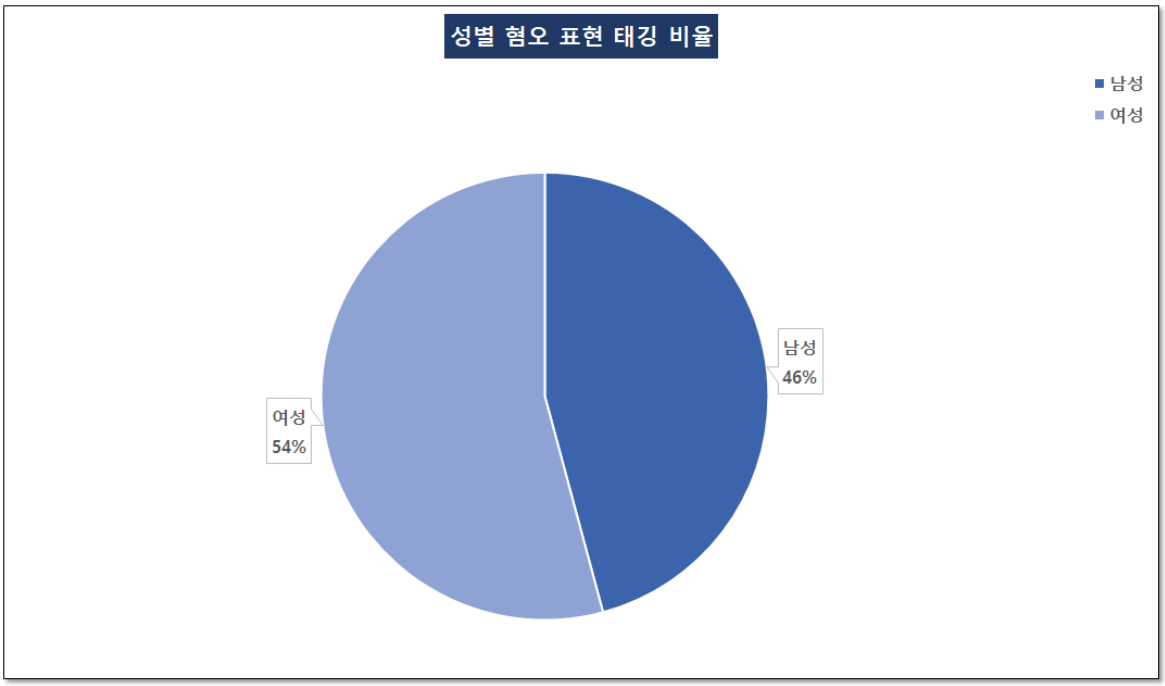
한편, 욕설 표현 다음으로 많은 건수를 기록한 경우는 '기타'인데, 기타 비윤리적 표현으로 분류된 것들에는 품위가 없거나 거친 말에 해당하는 비속어들이 대부분이었다. 특히 발화자가 특정 의미 단어 앞에 '개-', '존-' 등을 붙여 자신의 과장된 정서나 인식 대상의 왜곡된 상태를 부정적으로 표현하거나 특정 문맥과 의미 연관 없이 습관적으로 사용하는 특징이 있다. 이러한 비속어들은 일견 욕설 표현과 유사해 보이지만 의미 내용과 발화 맥락에 따라서 욕설 표현과는 차이가 있다고 할 수 있다.

다음, 혐오, 성적, 차별적 표현들은 욕설 표현에 비해 그 건수와 비율이 상대적으로 낮게 나타났는데 그중에서도 혐오 표현이 성적, 차별적 표현들보다 상대적으로 높은 건수와 비율(7,442건, 0.121%)로 나타났음을 알 수 있다. 혐오 표현의 경우 최근 사회적으로 큰 이슈화 되면서 아무리 사소한 경우라고 하더라도 사회적으로 용인되기 힘든 표현으로 주목되고 있으며, 특히 특정 집단에 대한 혐오로 드러날 경우 윤리적 논란이 되기 일쑤이다. 따라서 '모두의 말뭉치'에서도 혐오 표현은 정제되도록 향후 데이터를 개선해 나갈 필요가 있다.

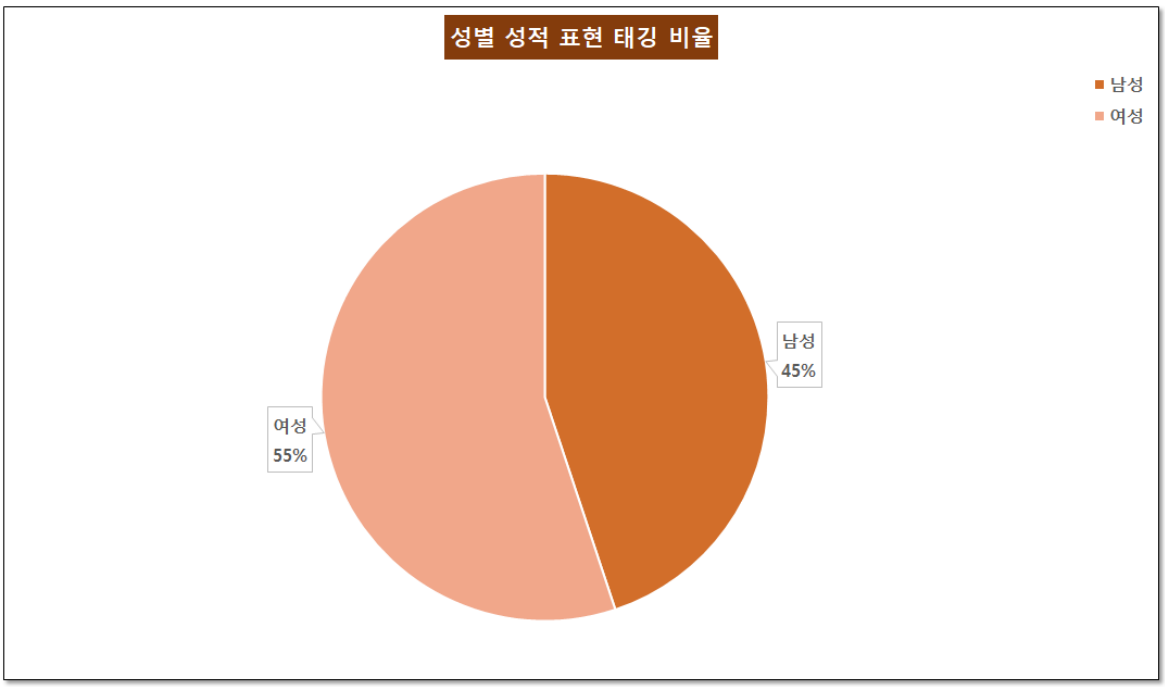
성적 표현과 차별적 표현들도 그 건수와 비율이 비록 상대적으로 작지만 최근 사회적으로 용인되기 힘들 만큼 민감한 표현들이 공론화되었을 경우 도덕적 비난의 대상이 되기 때문에 모두의 말뭉치에서 정제되고 개선되어야 할 필요가 있다.

한편, 우리는 국립국어원에서 제공받은 전체 말뭉치에 대한 비윤리적 표현 유형 빈도를 평가자의 변인별로 쉽게 비교할 수 있도록 그래프를 제시하고 조사 과정이나 결과에서 유의할 점에 대해 간략한 서술이 필요하다고 판단하였다. 다음은 100명의 평가자의 '성별', '연령', '지역', '직업' 변인에 따른 그래프와 이에 대한 설명이다.

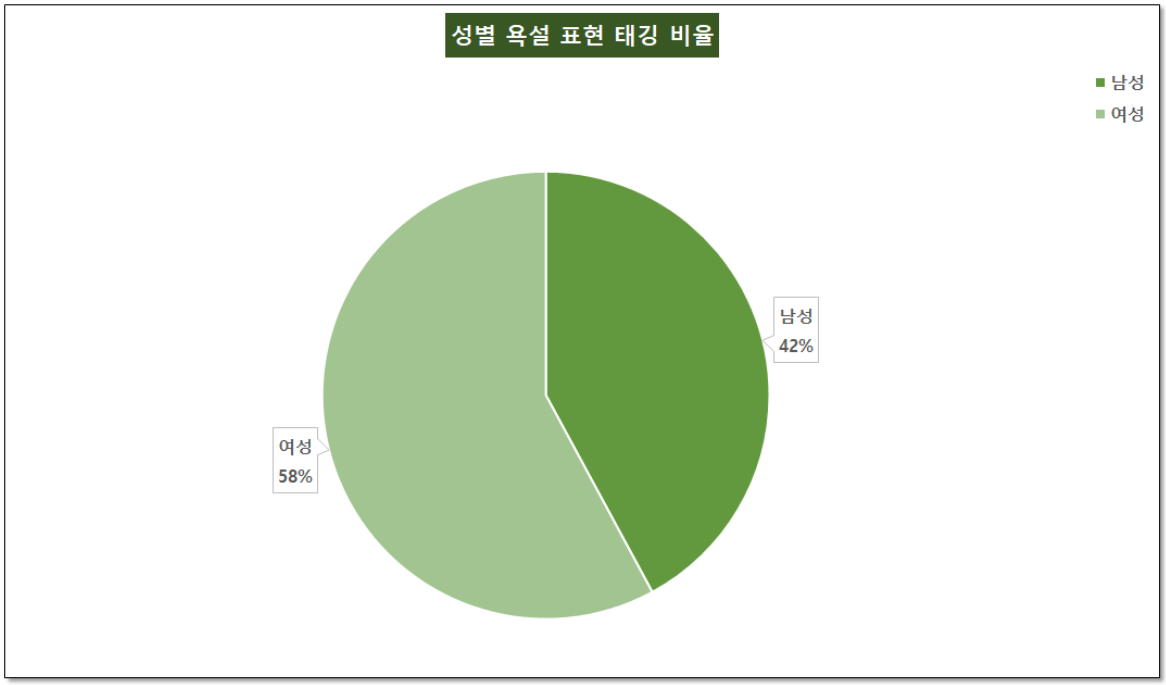
○ 평가자의 [성별] 변인에 따른 그래프



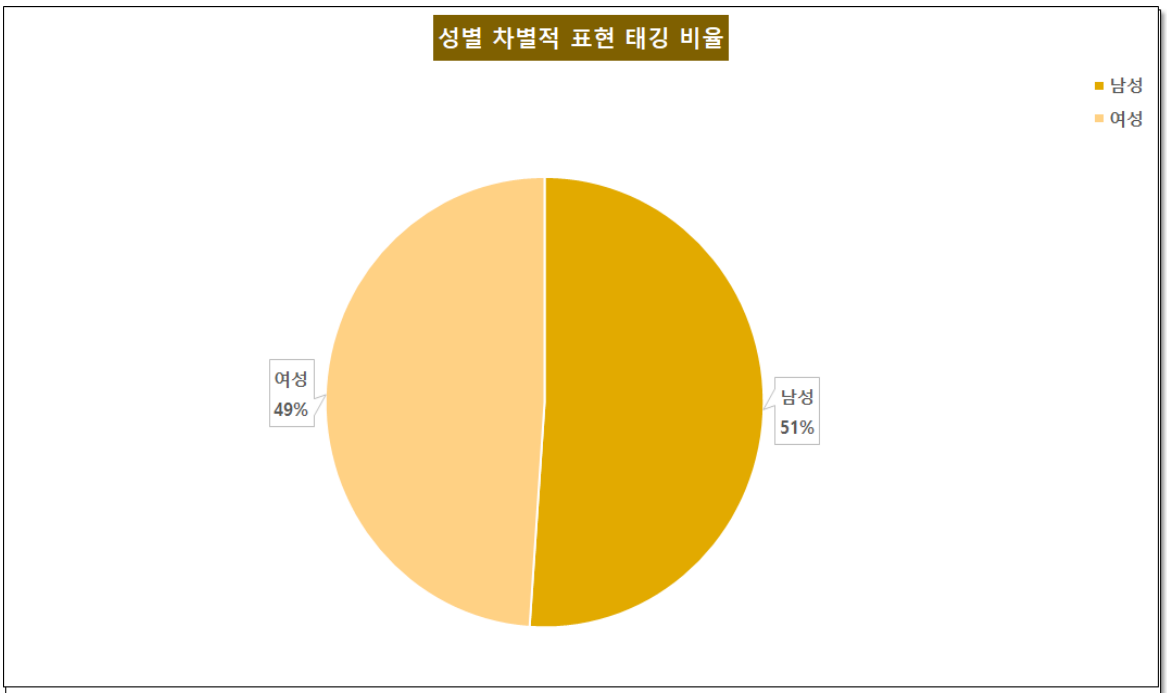
[그림 31] 성별에 따른 혐오 표현 태깅 비율



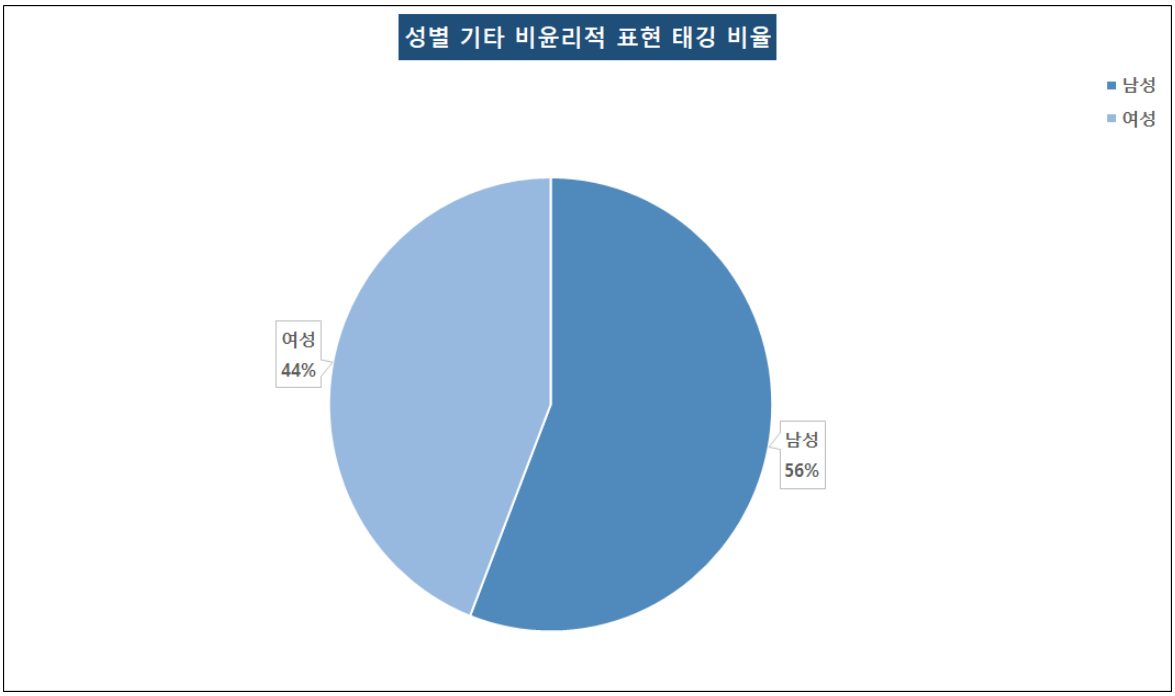
[그림 32] 성별에 따른 성적 표현 태깅 비율



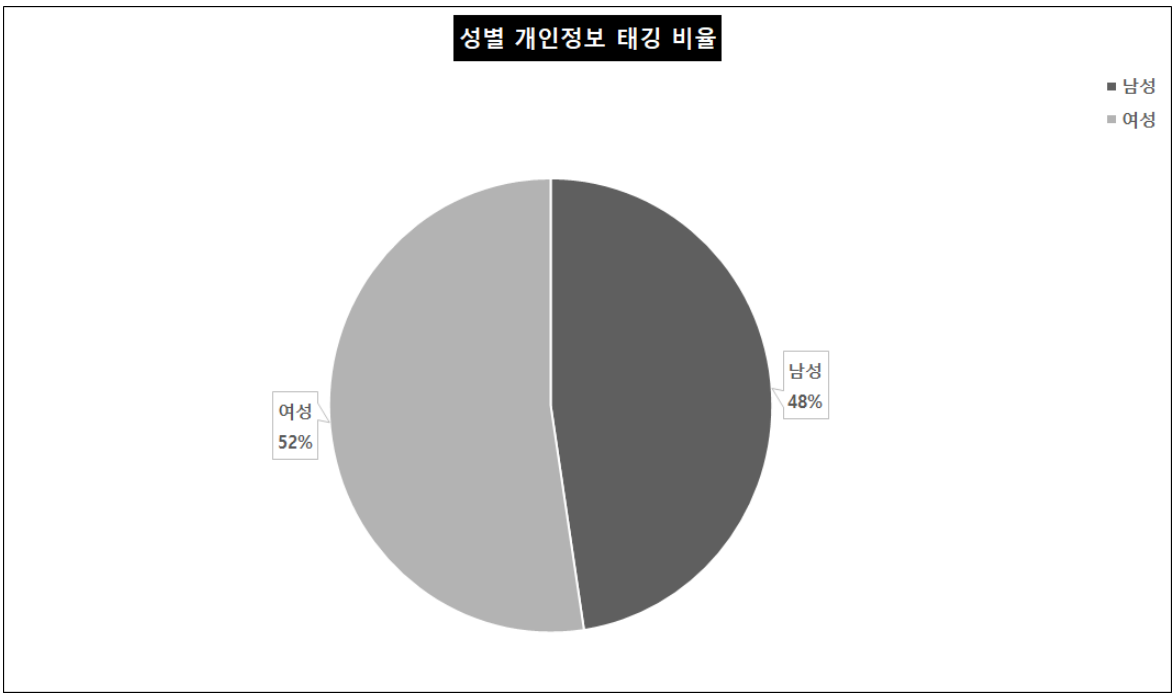
[그림 33] 성별에 따른 욕설 표현 태깅 비율



[그림 34] 성별에 따른 차별 표현 태깅 비율



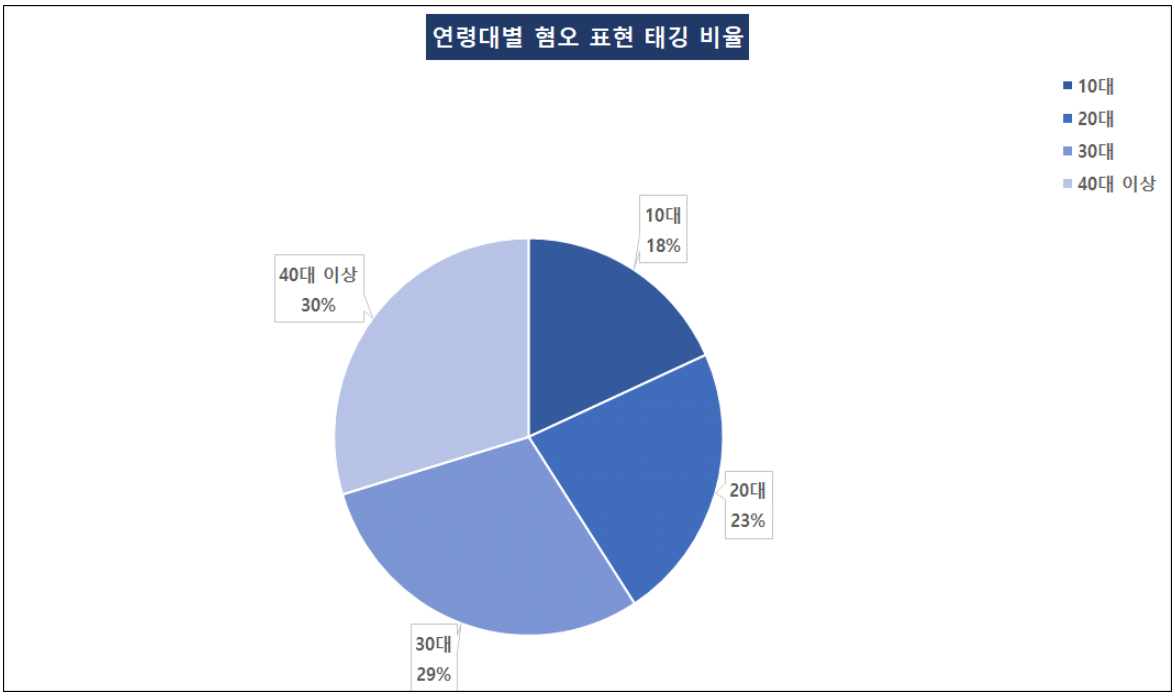
[그림 35] 성별에 따른 기타 비윤리적 표현 태깅 비율



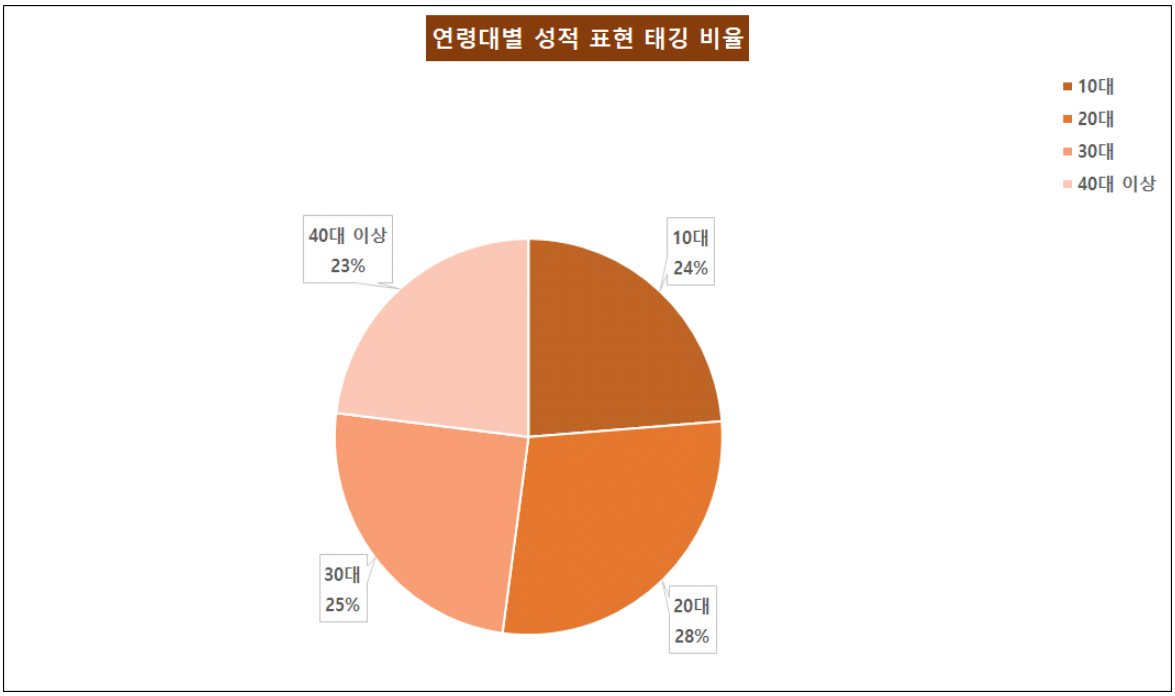
[그림 36] 성별에 따른 개인정보 노출 태깅 비율

남성 평가자는 할당된 2,429,401 발화의 1.42%인 34,506건을, 여성 평가자는 할당된 3,722,734 발화의 1.69%인 62,748건을 비윤리적 표현으로 식별하였다.

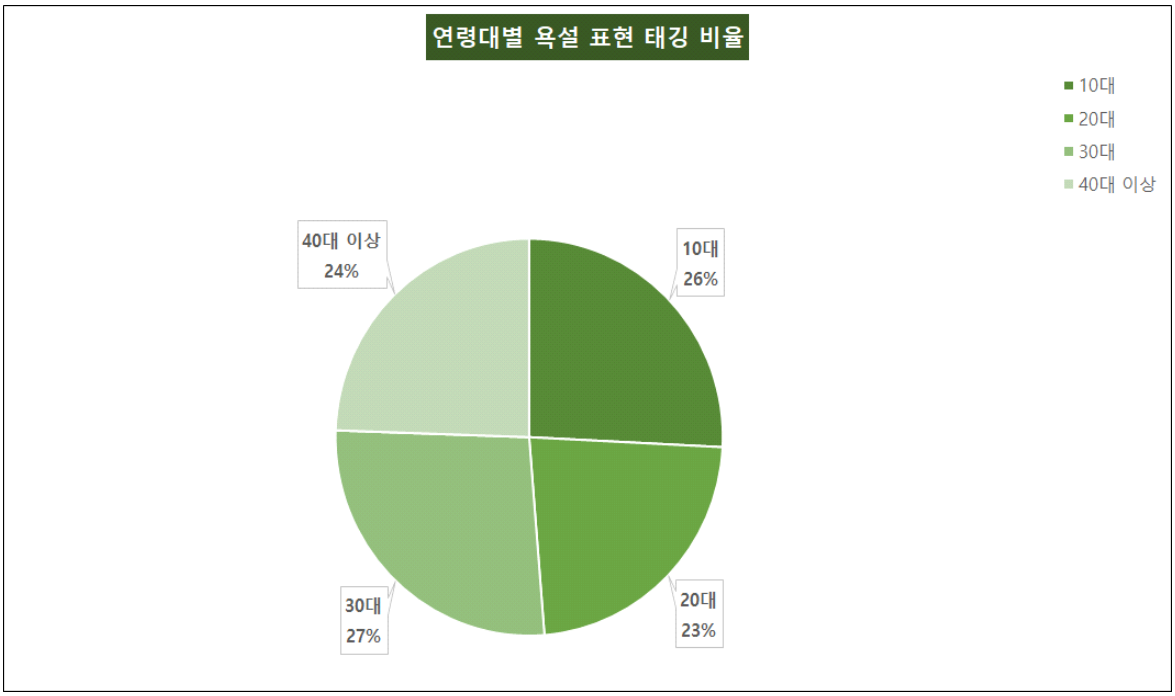
○ 평가자의 [연령] 변인에 따른 그래프



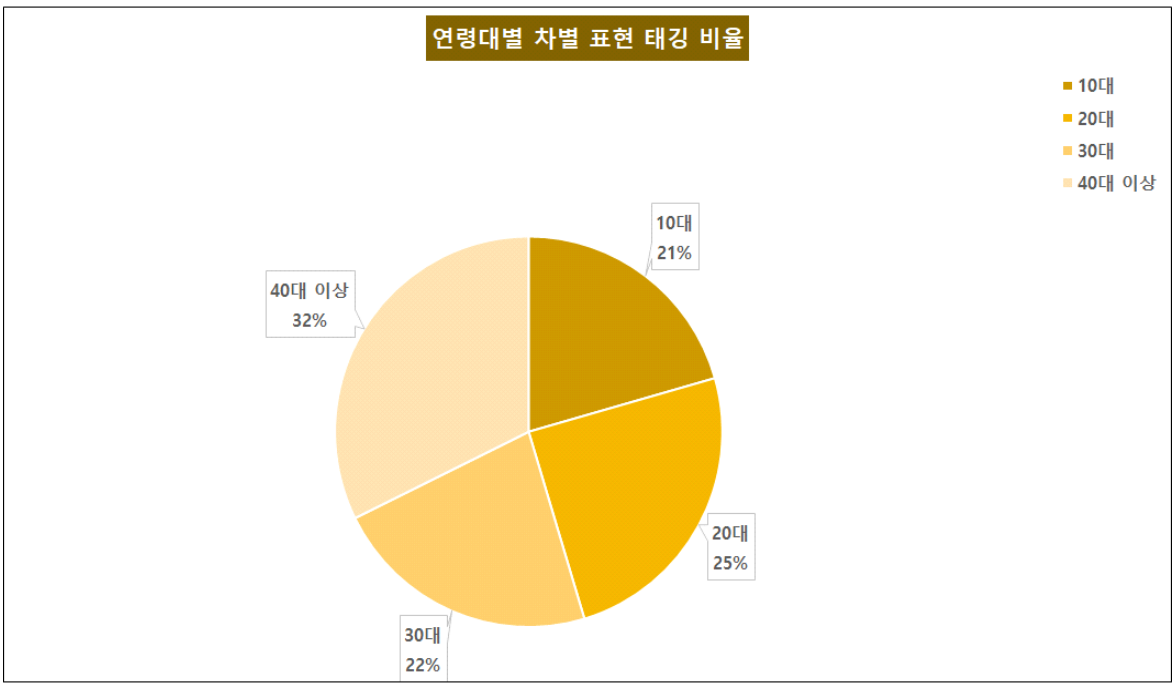
[그림 37] 연령대에 따른 혐오 표현 태깅 비율



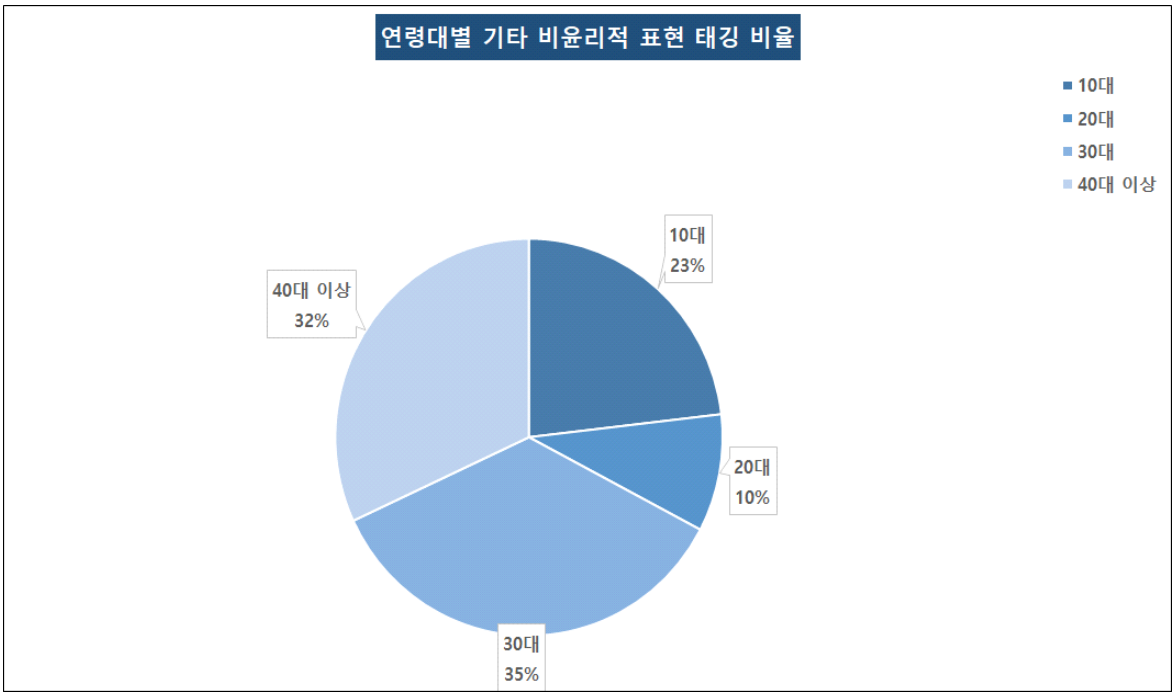
[그림 38] 연령대에 따른 성적 표현 태깅 비율



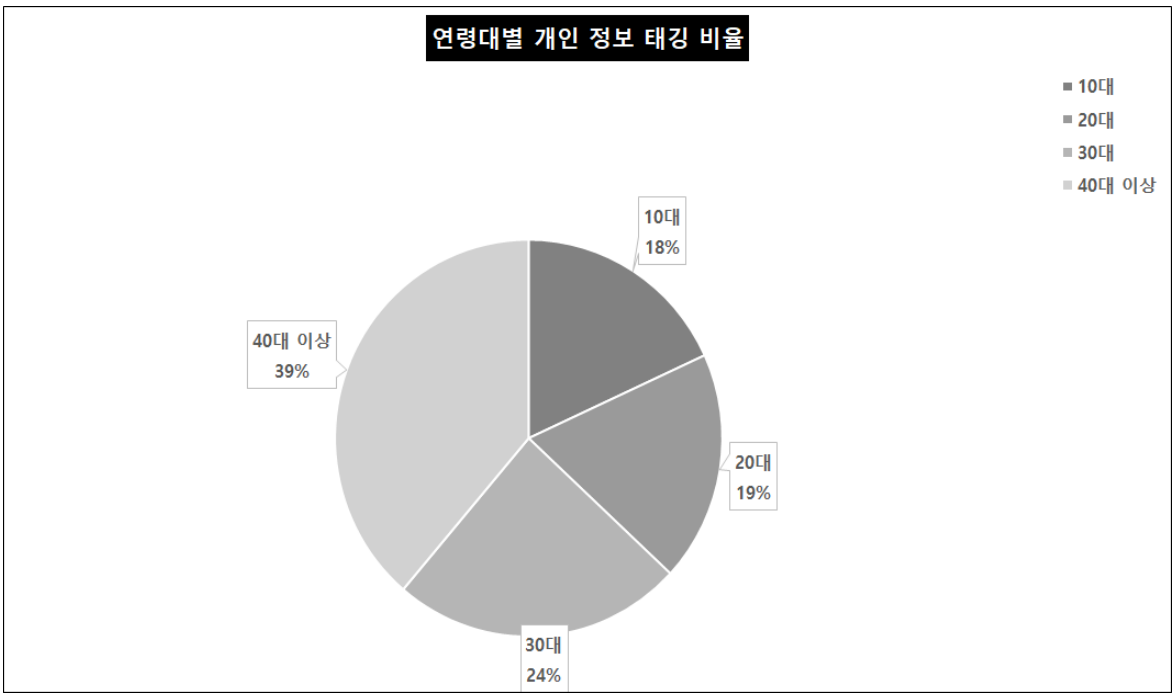
[그림 39] 연령대에 따른 욕설 표현 태깅 비율



[그림 40] 연령대에 따른 차별 표현 태깅 비율



[그림 41] 연령대에 따른 기타 비윤리적 표현 태깅 비율



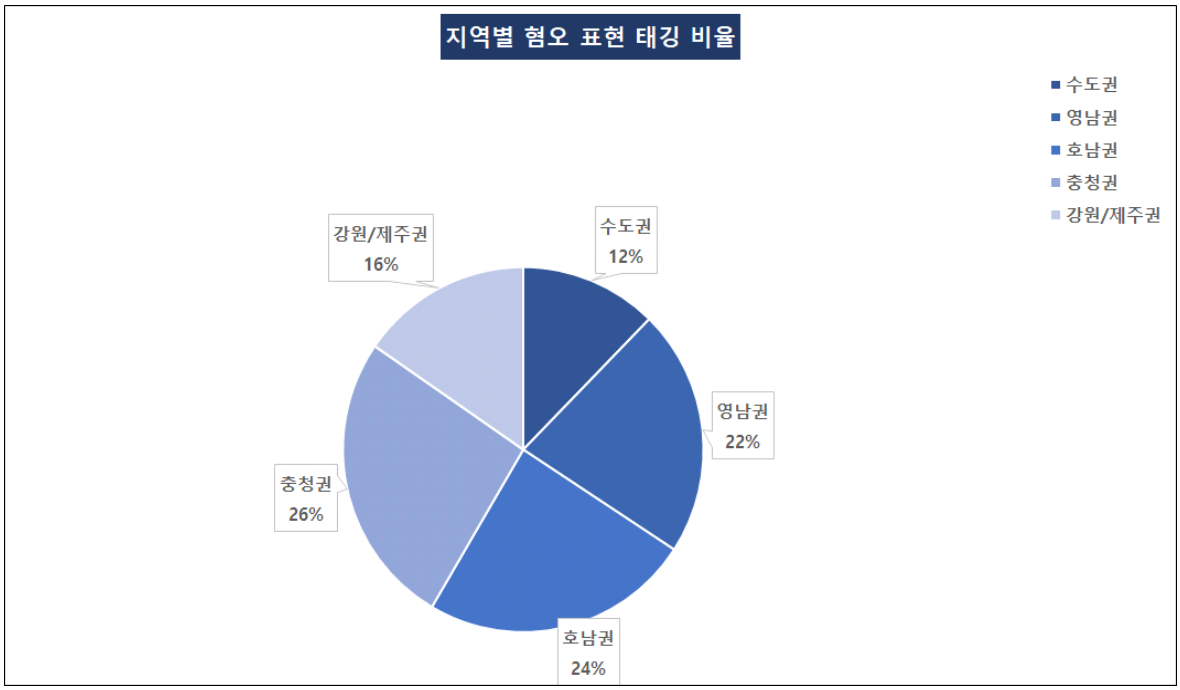
[그림 42] 연령대에 따른 개인정보 노출 태깅 비율

10대 평가자는 할당된 1,051,771 발화의 1.53%인 16,104건을, 20대 평가자는 할당된 2,134,785 발화의 1.32%인 28,140건을, 30대 평가자는 할당된 1,964,164 발화의 1.8%인 35,315건을, 40대 이상 평가자는 할당된 1,001,415 발화의 1.77%인

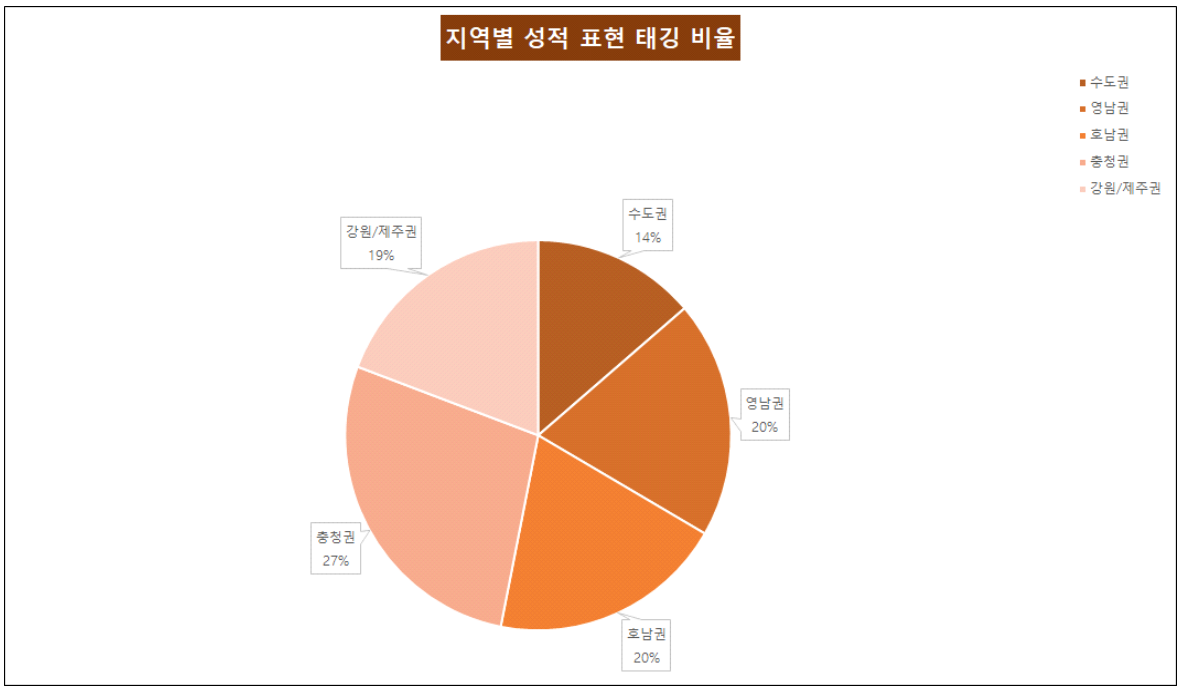
17,695건을 비윤리적 표현으로 식별하였다.

연령별 배당 비율은 20대와 30대에 치우쳤으며 상대적으로 40대 이상의 평가자에게 다소 낮게 배당되었다. 10대와 40대에 비해 20대와 30대에게 상대적으로 많은 발화가 배당되었다는 점이 평면적이고 정량적 기준에서 보면 문제의 소지가 있어 보인다. 하지만 본고의 연구 대상인 온라인 및 메신저 대화 활동량을 고려한다면 이와 같이 배당하는 것이 일상의 언어 사용 상황을 적절하게 고려한 것이라고 판단된다.

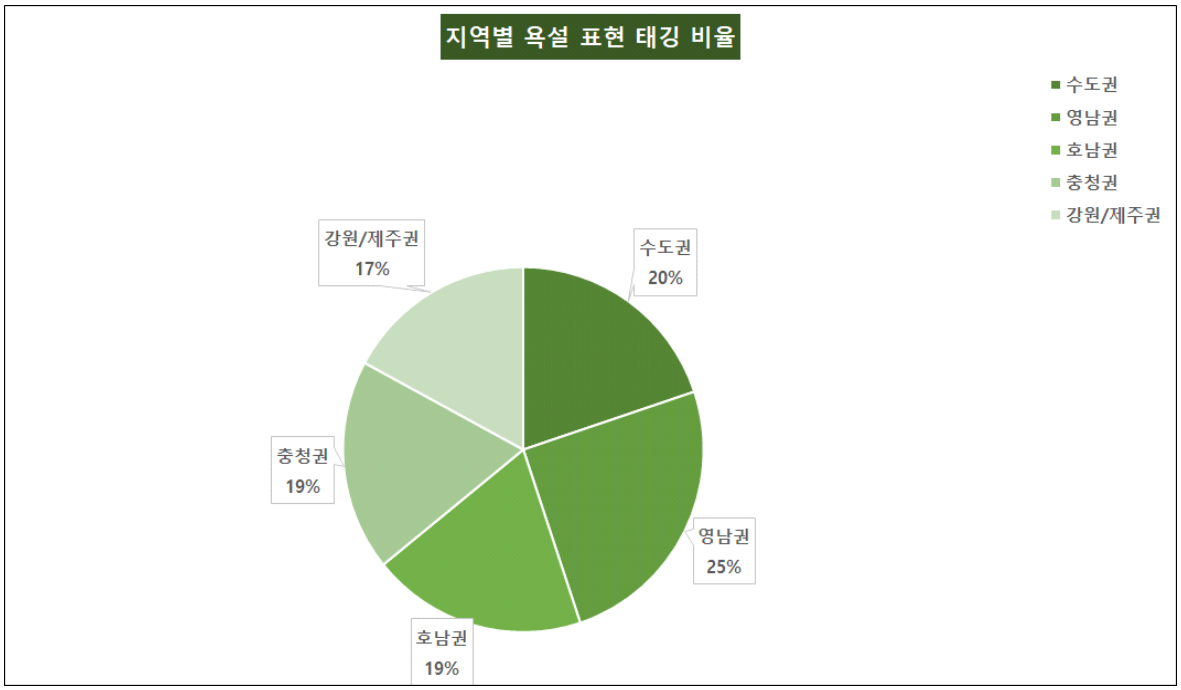
○ 평가자의 [지역] 변인에 따른 그래프



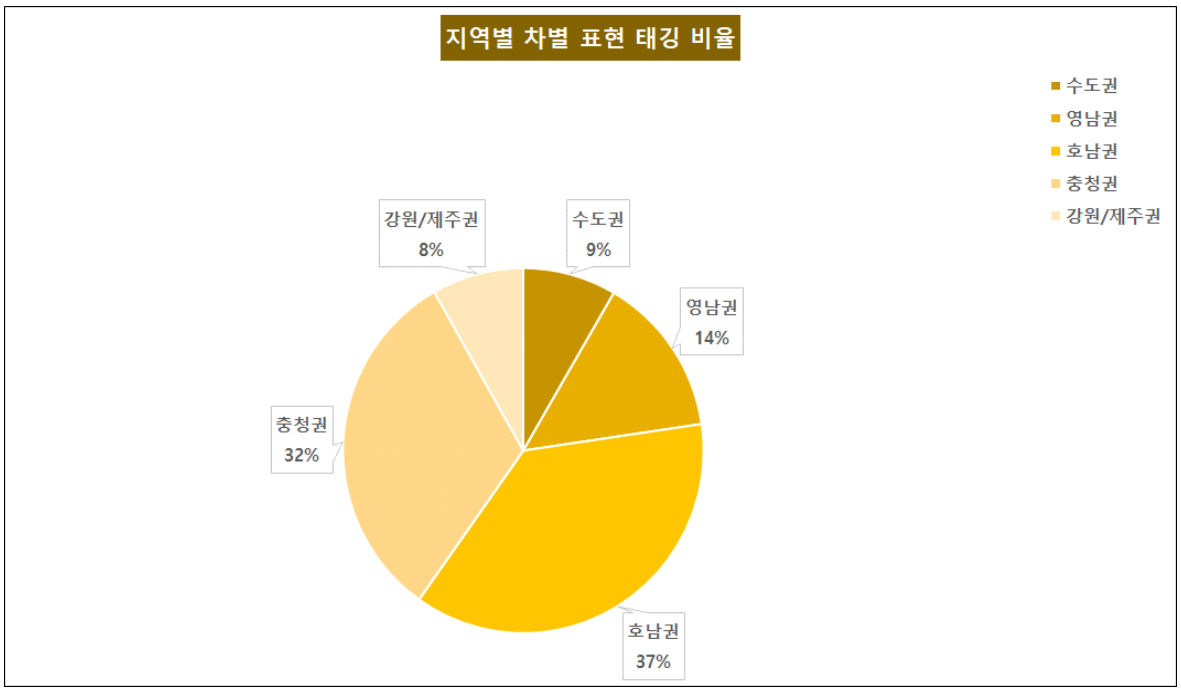
[그림 43] 지역에 따른 혐오 표현 태깅 비율



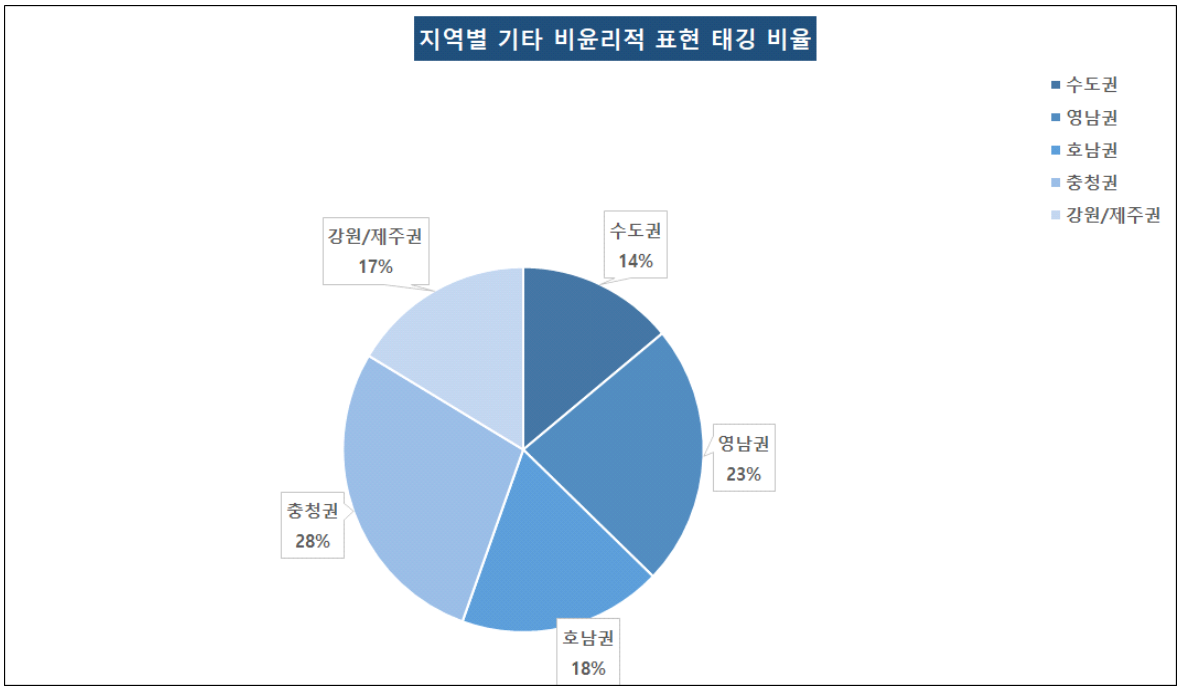
[그림 44] 지역에 따른 성적 표현 태깅 비율



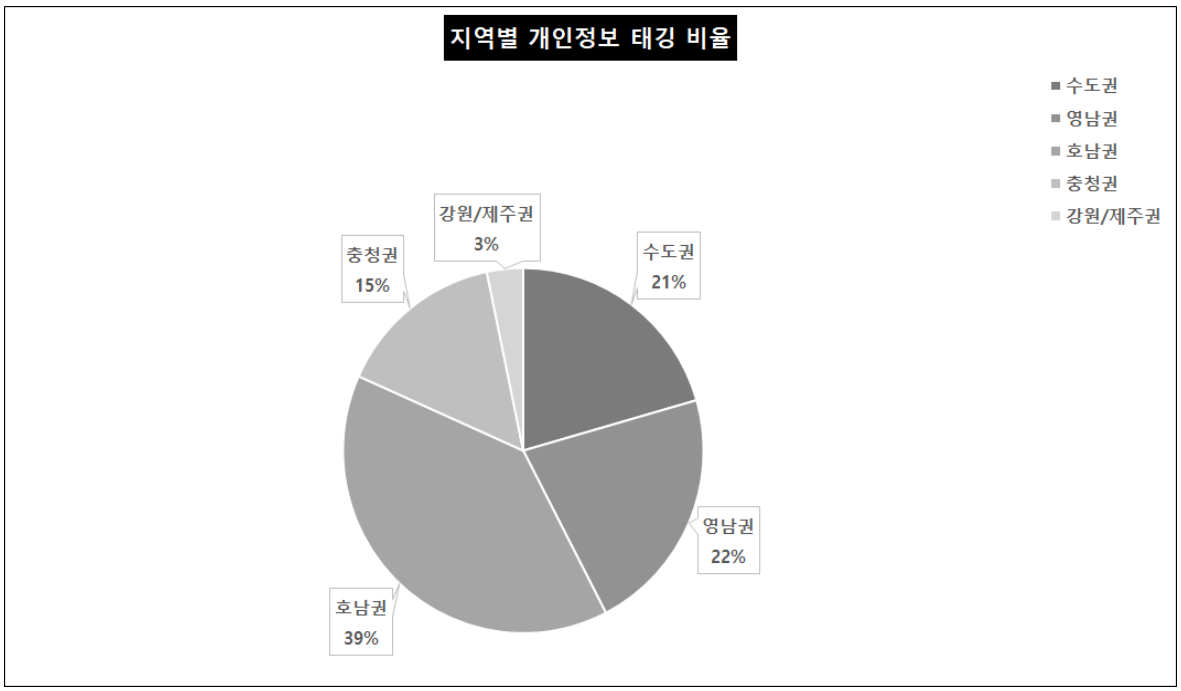
[그림 45] 지역에 따른 욕설 표현 태깅 비율



[그림 46] 지역에 따른 차별 표현 태깅 비율



[그림 47] 지역에 따른 기타 비윤리적 표현 태깅 비율



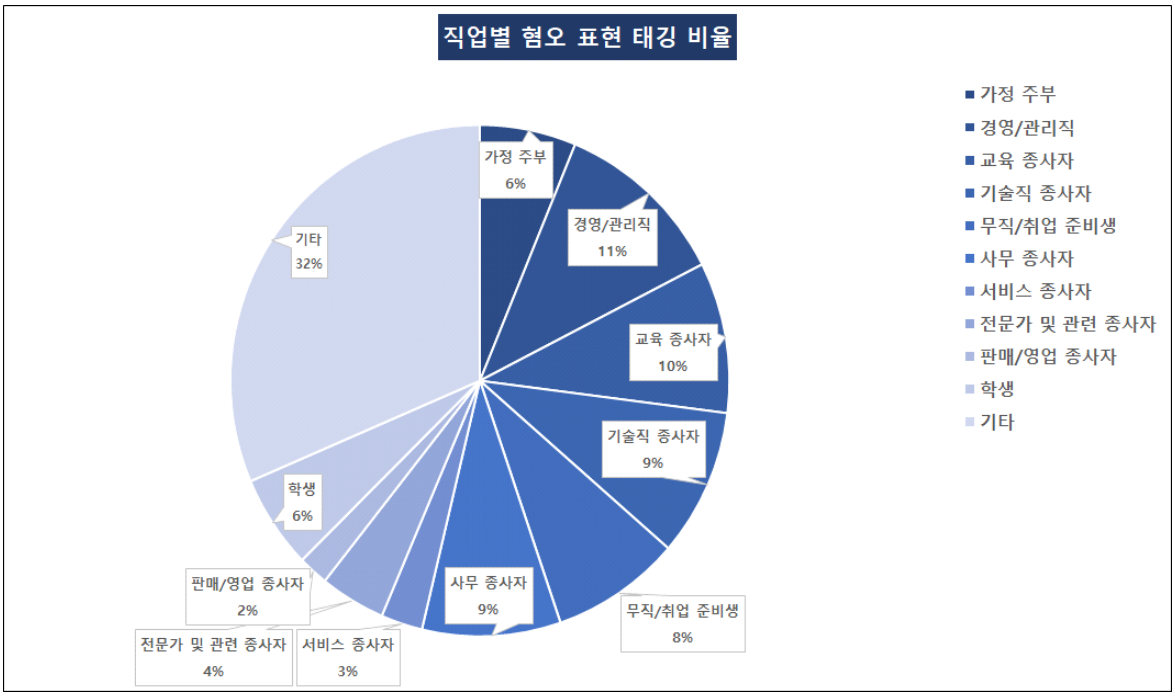
[그림 48] 지역에 따른 개인정보 태깅 비율

‘지역’ 변인은 총 5개 권역(수도권, 영남권, 호남권, 충청권, 강원/제주권)으로 구분하였다. 수도권 평가자는 할당된 3,375,085 발화의 1.39%인 46,942건을, 영남권 평가자는 할당된 1,603,453 발화의 1.88%인 30,182건을, 호남권 평가자는 할당된 564,474 발화의 1.77%인 9,966건을, 충청권 평가자는 할당된 487,299 발화의

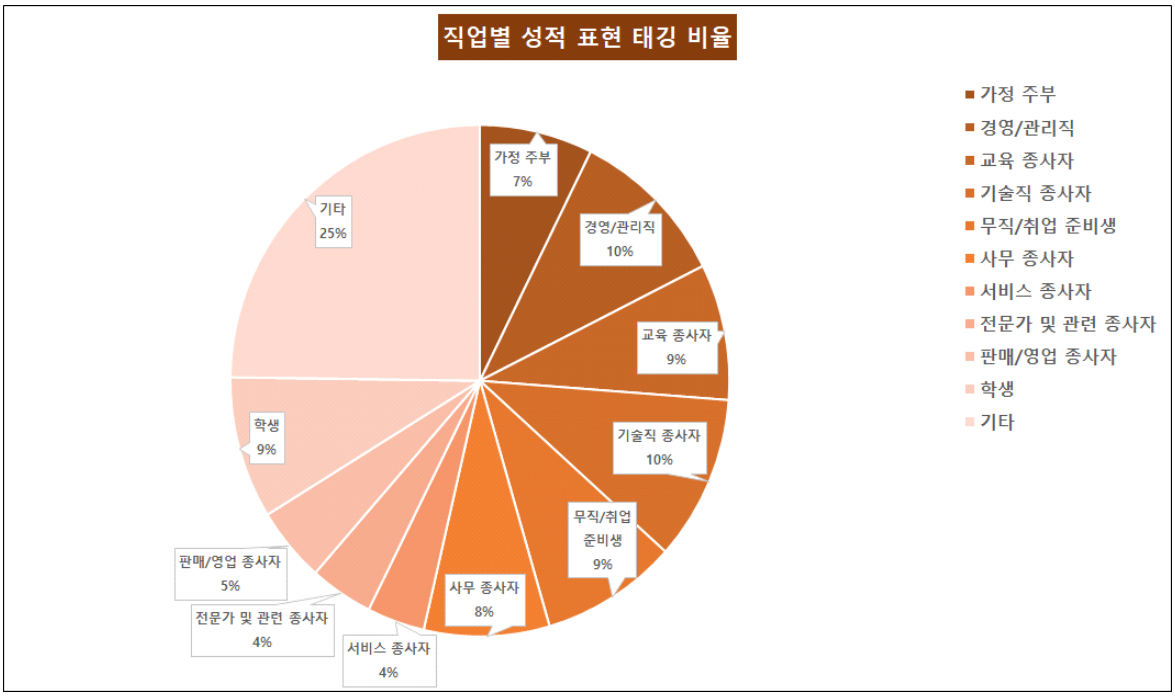
1.77%인 8,642건, 강원/제주권 평가자는 할당된 121,824 발화의 1.25%인 1,522건을 비윤리적 표현으로 식별하였다.

지역별 거주 인구의 수를 고려해 볼 때, 수도권에 그 외 지역보다 더 많은 평가자를 할당한 것은 실제 인구통계학적 특성을 반영한다는 점에서 적절하다고 판단된다.

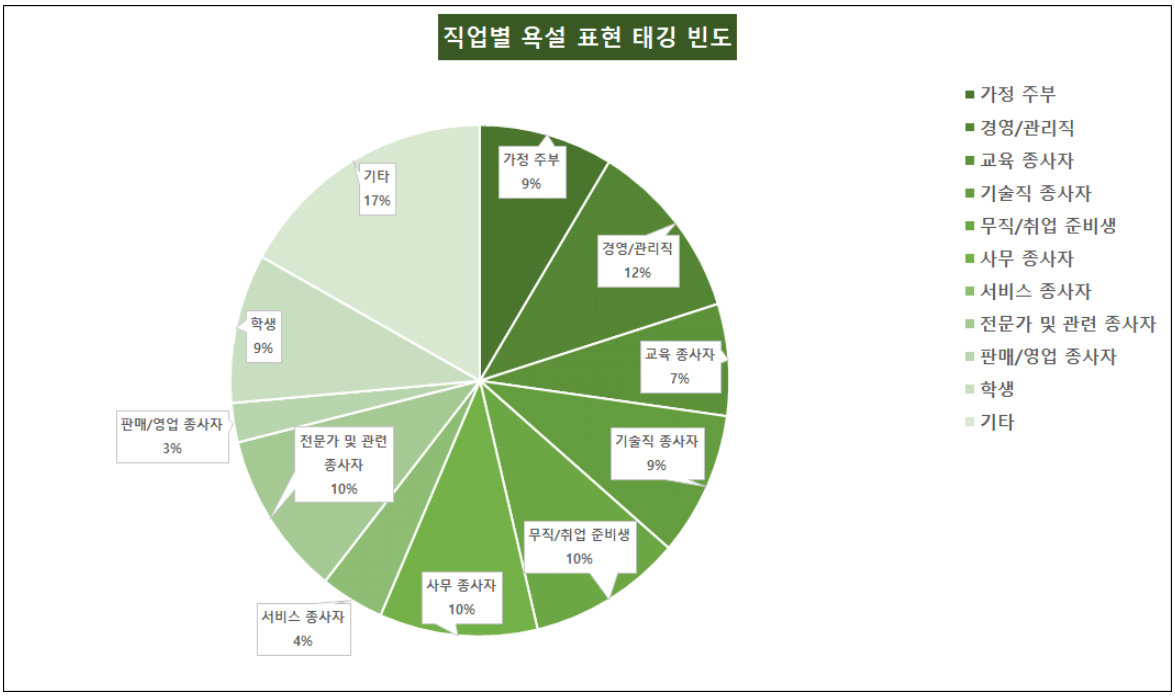
○ 평가자의 [직업] 변인에 따른 그래프



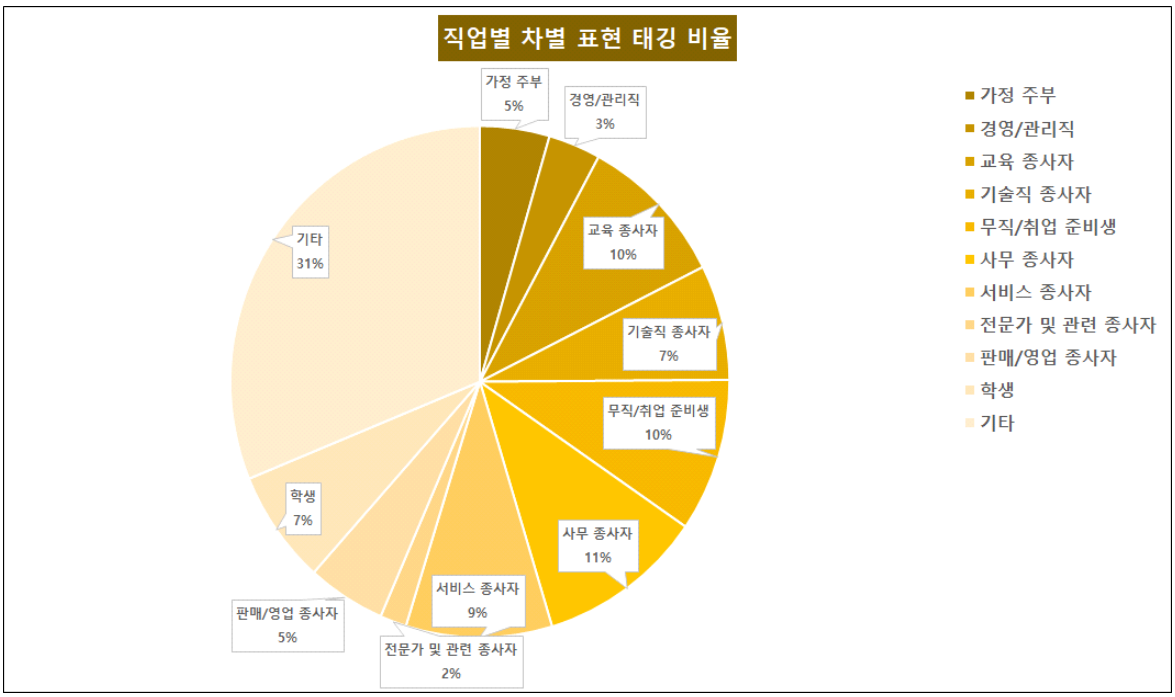
[그림 49] 직업에 따른 혐오 표현 태깅 비율



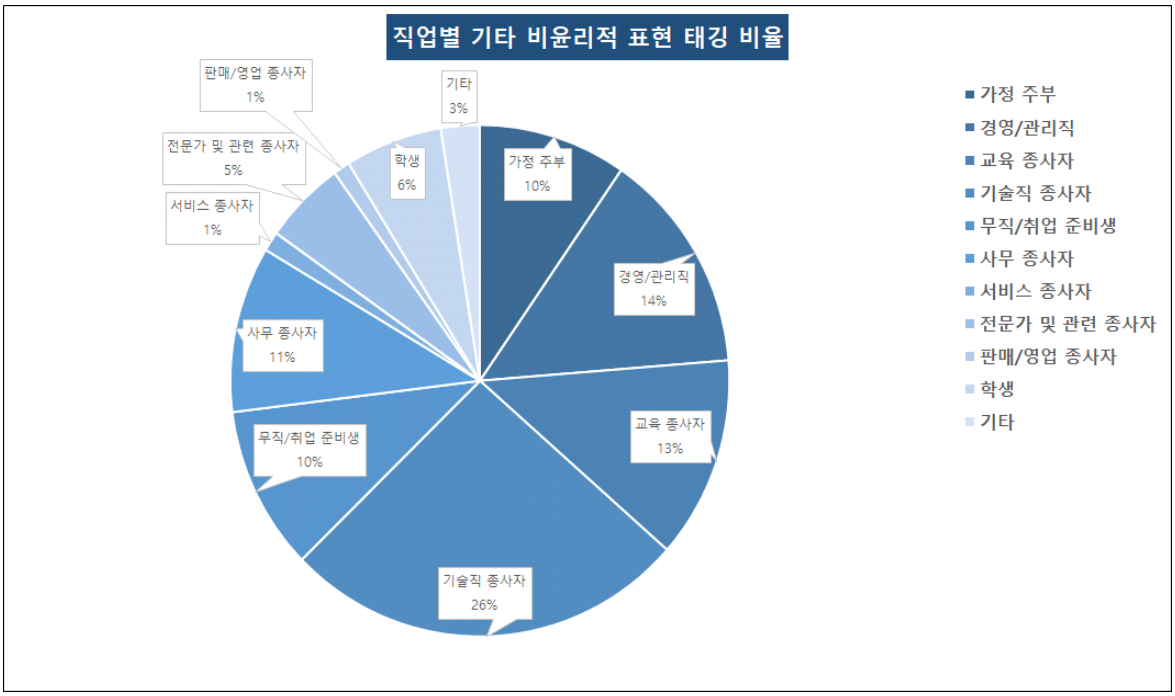
[그림 50] 직업에 따른 성적 표현 태깅 비율



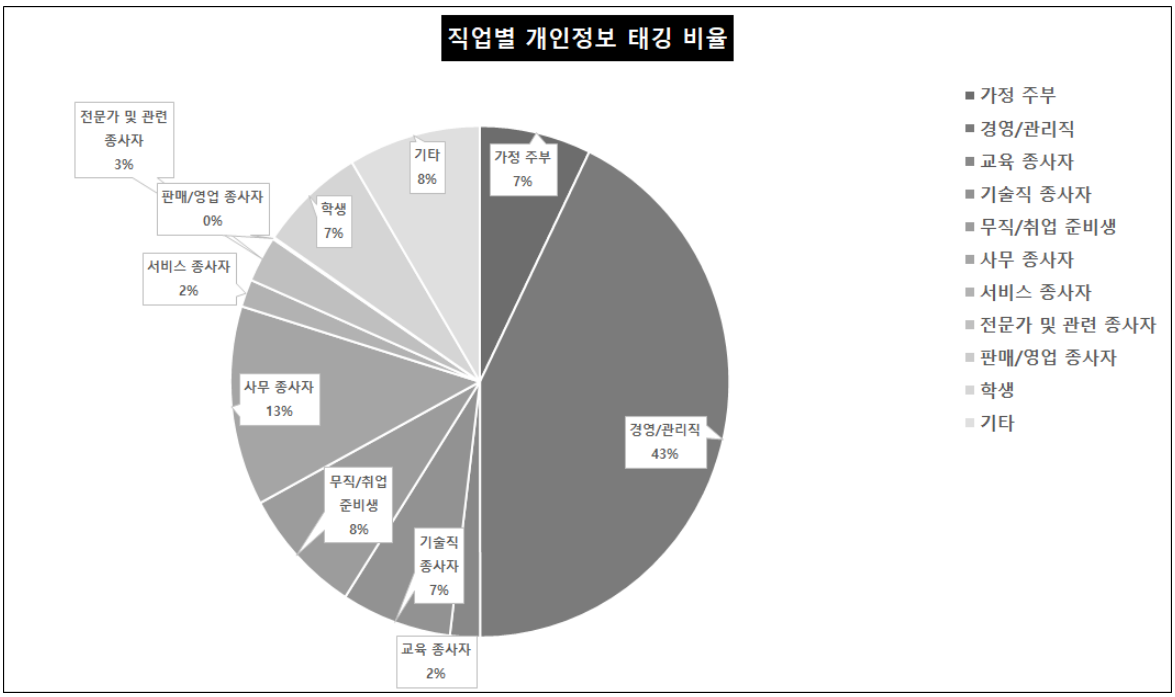
[그림 51] 직업에 따른 욕설 표현 태깅 빈도



[그림 52] 직업에 따른 차별 표현 태깅 비율



[그림 53] 직업에 따른 기타 비윤리적 표현 태깅 비율



[그림 54] 직업에 따른 개인정보 노출 태깅 비율

'직업' 변인은 평가에 참여 확정된 100명의 직업을 분석하여 가능한 범위에서 묶어서 범주를 구성하였다. 가정주부 평가자는 할당된 365,474 발화의 1.39%인 5,083건을, 경영/관리직 평가자는 할당된 182,737 발화의 2.38%인 4,353건을, 교

육 종사자 평가자는 할당된 365,472 발화의 1.35%인 4,929건을, 기술직 종사자 평가자는 할당된 339,447 발화의 2%인 6,799건, 사무 종사자 평가자는 할당된 1,177,066 발화의 1.74%인 20,429건을, 서비스 종사자 평가자는 할당된 121,825 발화의 0.59%인 722건을, 전문가/관련 종사자 평가자는 할당된 416,415 발화의 1.31%인 5,448건을, 판매/영업 종사자 평가자는 할당된 60,912 발화의 0.37%인 226건을, 학생 평가자는 할당된 2,026,366 발화의 1.39%인 28,220건을, 무직/취업 준비생 평가자는 할당된 791,861 발화의 1.63%인 12,881건을, 기타 직업 평가자는 할당된 304,560 발화의 2.68%인 8,164건을 비윤리적 표현으로 식별하였다.

'직업' 변인의 배당 분포 중에서 발화 수가 가장 많이 배당된 경우는 학생(2,026,336건)이며, 전체 발화 수 대비 34% 정도에 해당한다. 그 다음으로 많이 배당된 경우는 사무 종사자(1,177,066건)로서 20% 정도에 해당한다. 이 두 직업의 비율을 합치면 50% 이상으로 절반 이상의 발화 수가 이 두 직업군에 배당되었음을 알 수 있다. 나머지의 다양한 직업군들의 경우 무직/취업 준비생(791,861건, 13% 정도)을 제외하고 모두 30만 건 대 이하의 발화가 배정되었다. 따라서 직업 변인에서는 학생, 사무 종사자, 무직/취업 준비생에 해당하는 평가자의 의견이 더 많은 비율로 반영되었다는 점을 고려할 필요가 있다.

4.2 말뭉치 문서 종류별 비윤리적 표현 유형 비율 분석 및 정제 수준

우리가 분석 대상으로 삼은 자료는 모두 말뭉치 중 웹 문서, 메신저 대화, 일상 구어 대화 이렇게 세 종류이다. 우리는 온라인상에서 불특정 다수를 대상으로 하는 웹 말뭉치, 대면하지 않은 상황에서 PC나 모바일을 통해 상대방과 대화를 나누는 메신저 대화 말뭉치, 대면하여 말로 이야기한 내용이 담긴 일상 구어 대화 말뭉치는 구축 특성에 따라 비윤리적 표현 포함 유형이 다를 것이라는 전제하에 분석을 수행하였다.

우리의 전제를 구체적으로 말하자면 다음과 같다. 화자는 자신의 생각을 표현할 때 대화의 상황과 전달 방식에 따라 비윤리적 표현의 강도를 달리한다. 대화 당사자 등의 사회적 지위가 불균형한 상태에서 이루어지는 대화의 상황에서는 높은 강도의 비윤리적 표현 출현의 빈도가 높다. 또한, 대화 과정에 있지 않은

제3자에 대해 두 명의 화자가 대화를 나눌 때에도 비윤리적 표현의 강도가 높은 경향이 있다. 아울러, 익명이 보장된 상태의 발화 내용에도 높은 강도의 비윤리적 표현이 자주 등장한다. 이러한 점을 고려하여 말뭉치의 종류별로 비윤리적 표현의 빈도를 분석하는 것이 필요하다.

문서 종류별 비윤리적 표현 유형 태깅 빈도를 분석한 결과는 다음과 같다.

구분		발화 (개)	혐오 표현		성적 표현		욕설 표현		차별적 표현		개인 정보		기타 비윤리적 표현	
			건	비율 (%)	건	비율 (%)	건	비율 (%)	건	비율 (%)	건	비율 (%)	건	비율 (%)
말뭉치 문서종류	웹 (누리소통망)	1,370,659	4,158	0.303	2,668	0.195	22,127	1.614	1,942	0.142	2,302	0.168	7,811	0.57
	웹 (리뷰)	162,789	11	0.007	15	0.009	53	0.033	31	0.019	239	0.147	79	0.049
	웹 (게시판)	7,774	10	0.129	12	0.154	41	0.527	2	0.026	3	0.039	12	0.154
	웹 (블로그)	114,748	19	0.017	16	0.014	43	0.037	26	0.023	160	0.139	30	0.026
	메신저 대화 (2인)	2,777,154	2,875	0.104	1,197	0,043	33,425	1.204	1,461	0.053	2,780	0.1	7,676	0.276
	메신저 대화 (다자)	198,643	74	0.037	47	0.024	1,831	0.922	23	0.012	142	0.071	383	0.193
	일상구어 대화	1,520,368	295	0.019	104	0.007	552	0.036	417	0.027	1,835	0.121	327	0.022
합계		6,152,135	7,442	0.121	4,059	0.066	58,072	0.944	3,902	0.063	7,461	0.121	16,318	0.265

〈표 11〉 말뭉치 문서종류별 비윤리적 표현 유형 태깅 빈도

위의 도표에서 제시된 각 문서 종류별 발화 개수와 비윤리적 유형의 태깅 건수 및 비율을 서술하면 다음과 같다. 웹(누리소통망)의 총 발화 개수는 1,370,659개이며, 이 중 혐오 표현으로 태깅된 건수는 4,158건으로 발화 개수 대비 태깅 건수의 비율은 0.303%이다. 성적 표현은 2,668건으로 비율은 0.195%이며, 욕설 표현은 22,127건으로 비율은 1.614%이다. 차별적 표현은 1,942건으로 0.142%이다. 개인정보 노출은 2,302건으로 비율은 0.168%이다. 기타 비윤리적 표현은 7,811건으로 비율은 0.57%이다.

웹(리뷰)의 총 발화 개수는 162,789개이며, 이 중 혐오 표현으로 태깅된 건수는

11건으로 발화 개수 대비 태깅 건수의 비율은 0.007%이다. 성적 표현은 15건으로 비율은 0.009%이다. 욕설 표현은 53건으로 0.033%이다. 차별적 표현은 31건으로 0.019%이다. 개인정보 노출은 239건으로 0.147%이다. 기타 비윤리적 표현은 79건으로 0.049%이다.

웹(게시판)의 총 발화 개수는 7,774개이며, 이 중 혐오 표현으로 태깅된 건수는 10건으로 발화 개수 대비 태깅 건수의 비율은 0.129%이다. 성적 표현은 12건으로 0.154%이다. 욕설 표현은 41건으로 0.527%이다. 차별적 표현은 2건으로 0.026%이다. 개인정보 노출은 3건으로 0.039%이다. 기타 비윤리적 표현은 12건으로 0.154%이다.

웹(블로그)의 총 발화 개수는 114,748개이며, 이 중 혐오 표현으로 태깅된 건수는 19건으로 발화 개수 대비 태깅 건수의 비율은 0.017%이다. 성적 표현은 16건으로 비율은 0.014%이다. 욕설 표현은 43건으로 비율은 0.037%이다. 차별적 표현은 26건으로 비율은 0.023%이다. 개인정보 노출은 160건으로 비율은 0.139%이다. 기타 비윤리적 표현은 30건으로 비율은 0.026%이다.

메신저 대화(2인)의 총 발화 개수는 2,777,154개이며, 이 중 혐오 표현으로 태깅된 건수는 2,875건으로 발화 개수 대비 태깅 건수의 비율은 0.104%이다. 성적표현은 1,197건으로 비율은 0.043%이다. 욕설 표현은 33,425건으로 비율은 1.204%이다. 차별적 표현은 1,461건으로 비율은 0.053%이다. 개인정보 노출은 2,780건으로 비율은 0.1%이다. 기타 비윤리적 표현은 7,676건으로 비율은 0.276%이다.

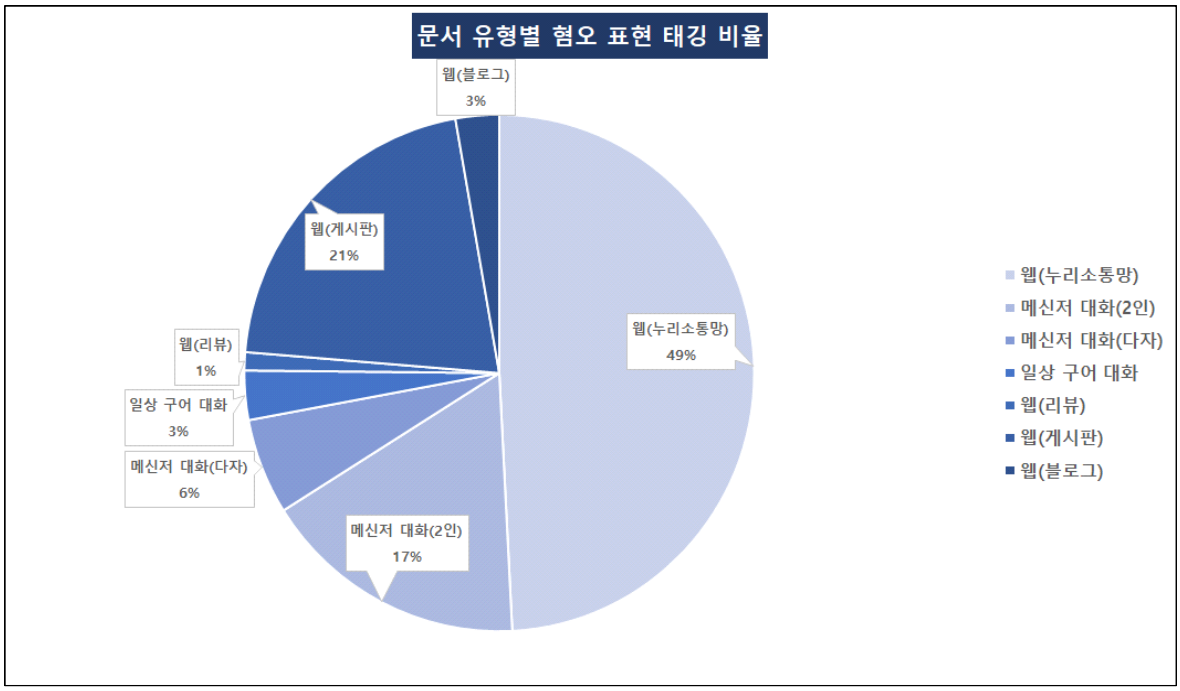
메신저 대화(다자)의 총 발화 개수는 198,643개이며, 이 중 혐오 표현으로 태깅된 건수는 74건으로 발화 개수 대비 태깅 건수의 비율은 0.037%이다. 성적 표현은 47건으로 비율은 0.024%이다. 욕설 표현은 1,831건으로 0.922%이다. 차별적 표현은 23건으로 0.012%이다. 개인정보 노출은 142건으로 0.071%이다. 기타 비윤리적 표현은 383건으로 0.193%이다.

일상 구어 대화의 총 발화 개수는 1,520,368개이며, 이 중 혐오 표현으로 태깅된 건수는 295건으로 발화 개수 대비 태깅 건수의 비율은 0.019%이다. 성적 표현은 104건으로 비율은 0.007%이다. 욕설 표현은 552건으로 0.036%이다. 차별적 표현은 417건으로 0.027%이다. 개인정보 노출은 1,835건으로 0.121%이다. 기타 비윤리적 표현은 327건으로 0.022%이다.

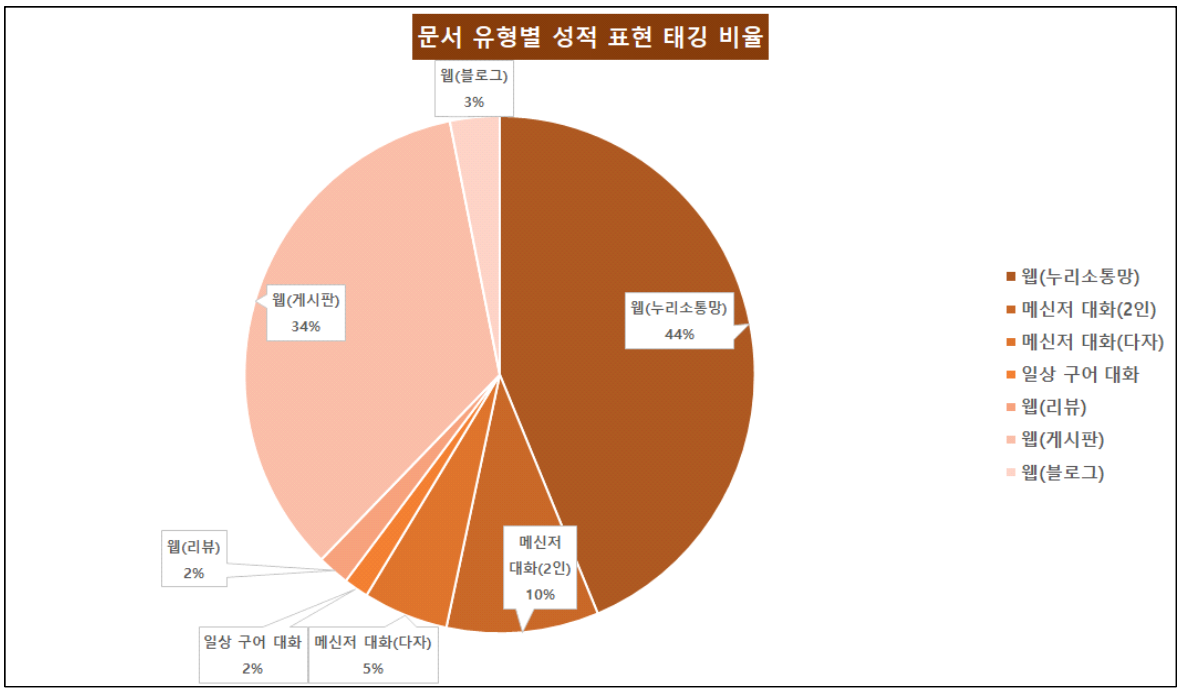
위의 도표의 수치를 분석한 말뭉치 종류별 비윤리적 표현의 빈도를 볼 때, 몇

가지 점에서 말뭉치 문서 종류별로 비윤리적 표현의 빈도의 차이가 있음을 우리는 확인할 수 있다. 우선, 모든 세부 종류 중에서 웹(누리소통망)의 경우는 다른 종류들보다 훨씬 더 많은 비율로 비윤리적 표현이 확인되었다. 웹(누리소통망)의 말뭉치는 혐오, 성적, 욕설, 차별, 개인정보, 기타 비윤리적 표현 모든 유형에서 가장 높은 비율을 보여주고 있다.

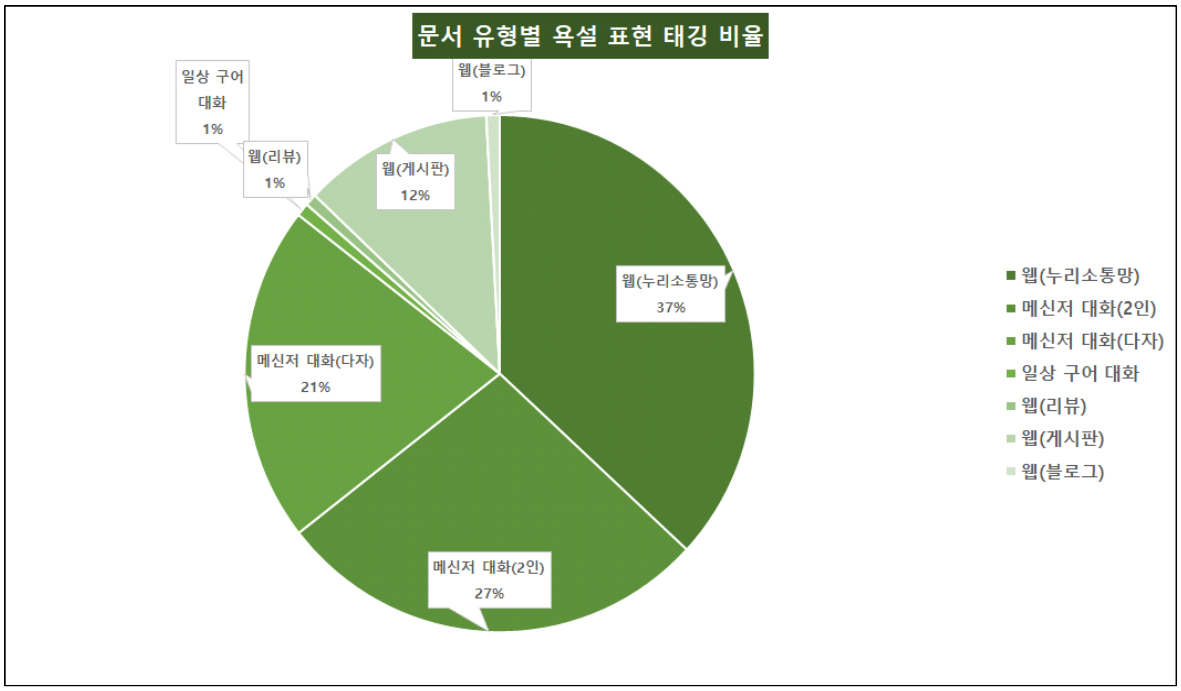
위에서 분석한 내용을 그래프로 제시하면 다음와 같다.



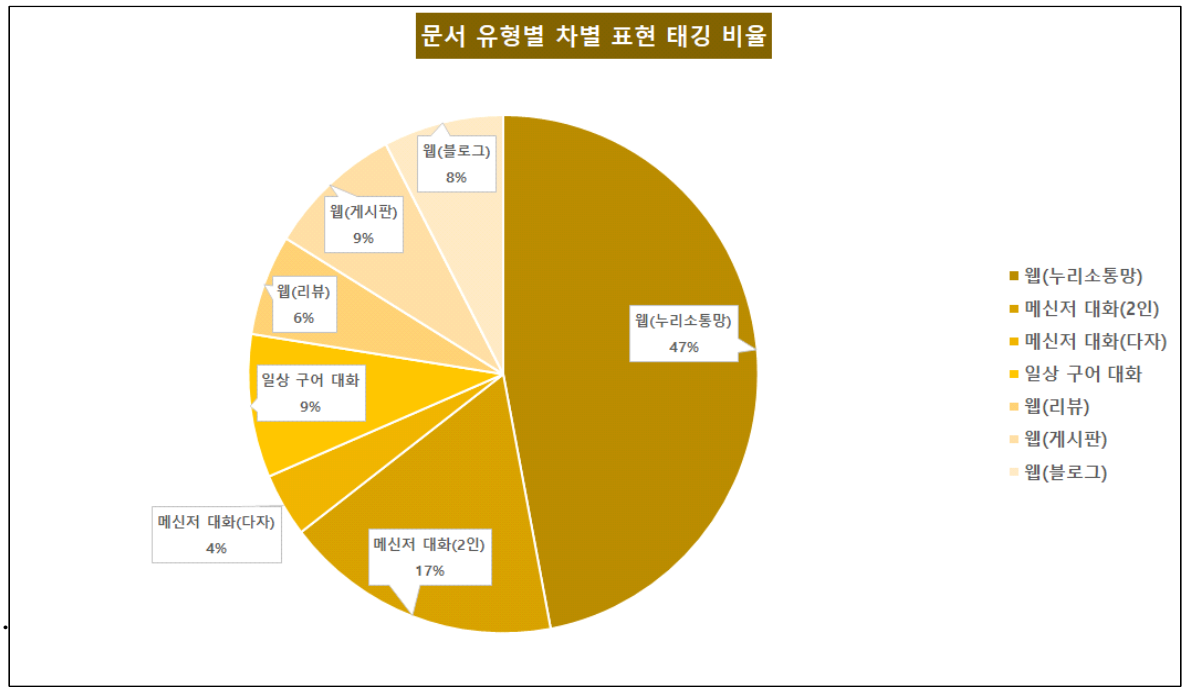
[그림 55] 문서 종류별 혐오 표현 태깅 분포 비율



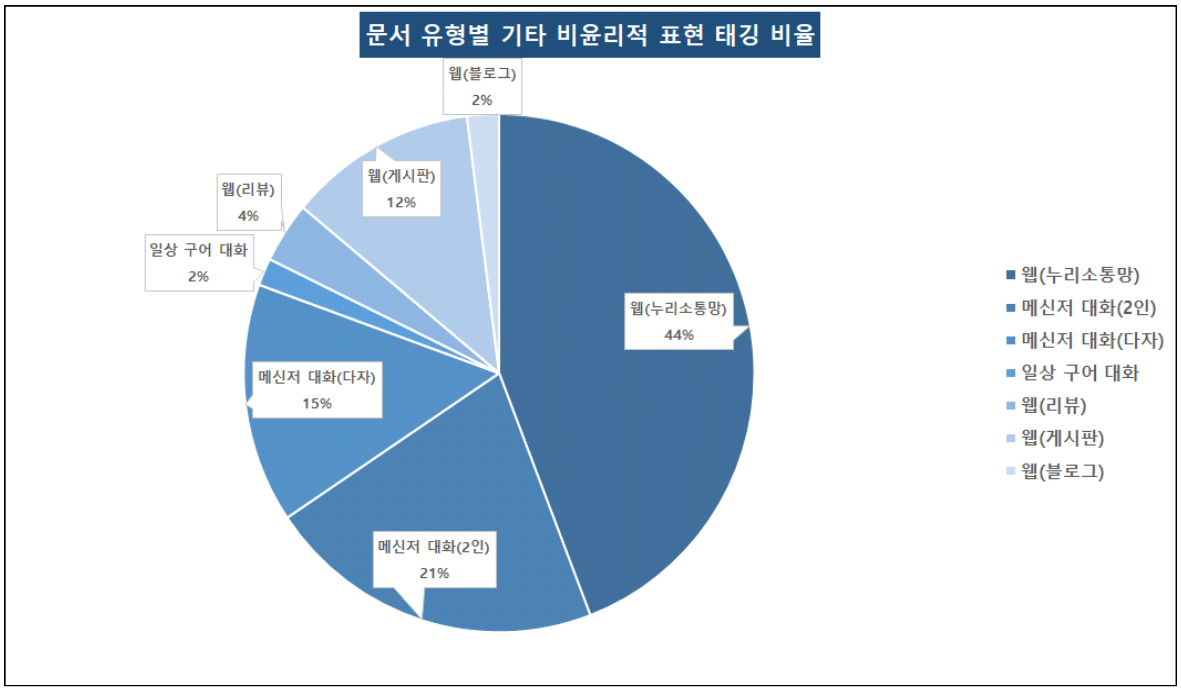
[그림 56] 문서 종류별 성적 표현 태깅 분포 비율



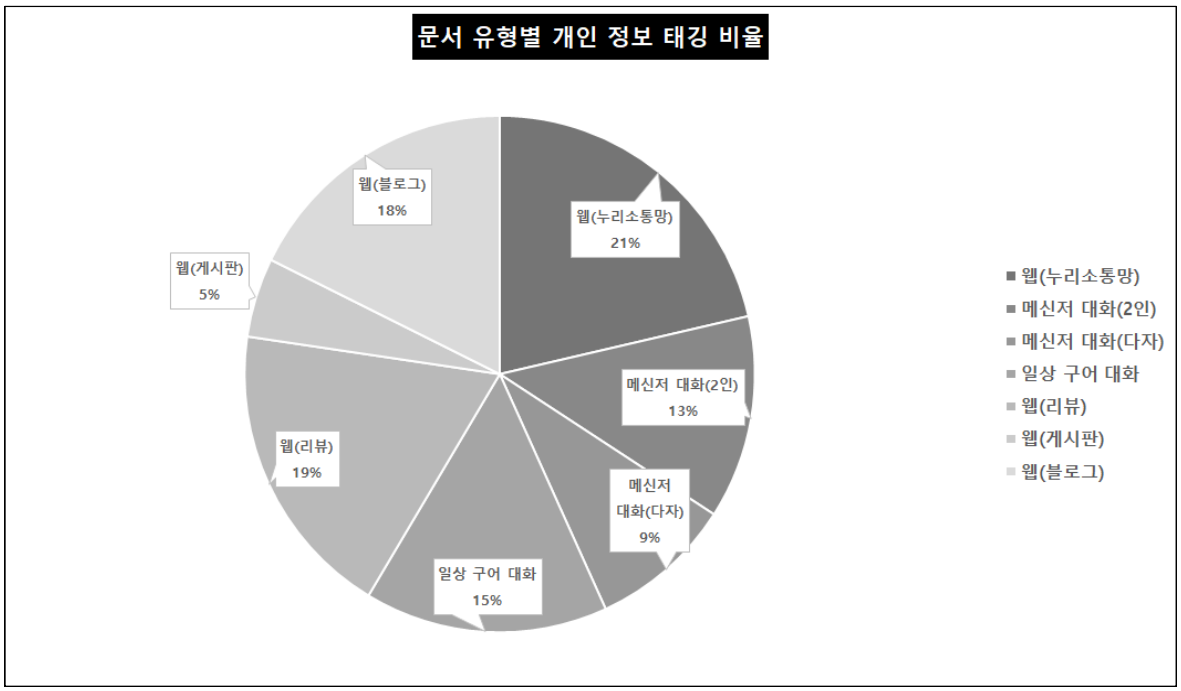
[그림 57] 문서 종류별 욕설 표현 태깅 분포 비율



[그림 58] 문서 종류별 차별 표현 태깅 분포 비율



[그림 59] 문서 종류별 기타 비윤리적 표현 태깅 분포 비율



[그림 60] 문서 종류별 개인정보 노출 태깅 분포 비율

위의 그래프들에서 보이는 바와 같이, 분석 대상을 웹 말뭉치로 한정한다면, 누리소통망 문서 다음으로 게시판 문서에서 개인정보 노출을 제외한 모든 유형의 태깅 비율이 높게 나타나고 있음을 알 수 있다. 반면 리뷰 문서나 블로그 문서의 경우에는 다른 웹 문서 종류보다 태깅 비율이 낮게 나타났다.

메신저 대화 말뭉치의 경우를 살펴보면, ‘2인 대화’와 ‘다자 대화’에서 비윤리적 표현 태깅 빈도가 확연한 차이를 보인다. 6가지 비윤리적 표현 유형의 모든 항목에서 ‘2인 대화’의 경우가 ‘다자 대화’의 경우보다 태깅 비율이 높게 나타났다. 유형별로 살펴보면 다음과 같은 편차를 보인다. 혐오 표현(0.104 > 0.037), 성적 표현(0.043 > 0.024), 욕설 표현(1.204 > 0.922), 차별적 표현(0.053 > 0.012), 개인정보 노출(0.1 > 0.071), 기타 비윤리적 표현(0.276 > 0.193). 이러한 점으로 볼 때, 향후 말뭉치 언어 수집 과정과 활용 등에서 ‘2인’의 메신저 대화의 경우 비윤리적 표현의 빈도가 높게 나타나게 됨을 인지하고 이를 충분히 고려할 필요가 있다.

일상 구어 대화는 다른 종류의 말뭉치들보다 비윤리적 표현의 빈도가 상대적으로 낮게 나타났다. 그 이유로서 직접 대면하여 대화를 할 때, 발화자는 비윤리적 표현을 삼가게 된다는 점을 추론할 수 있다. 일반적으로 사람들은 대화 상대자가 지금 현재 바로 눈앞에 대면해 있을 경우 상대방의 표정과 태도를 읽으며 감정을 느낄 수 있고, 상호 공감이나 역지사지의 감정이입을 통해서 발화자는 자신의 표현이 상대방에게 미칠 정서적 영향을 고려하여 비윤리적 표현을 삼가는 경향이 있다.

이상에서 분석한 내용을 종합적으로 고려하여 도출한 향후 말뭉치 수집 및 활용을 위한 몇 가지 시사점은 다음과 같다. 일상 구어 대화 말뭉치에서 등장하는 비윤리적 표현의 빈도가 다른 종류의 말뭉치보다 전반적으로 적게 나타나므로 이 종류의 말뭉치를 인공지능 챗봇 등에 활용한다면 상대적으로 ‘안정적’이라고 할 수 있다. 한편, 웹 말뭉치 중에서 누리소통망 문서나 게시판 문서의 경우는 다른 종류의 말뭉치보다 전반적으로 비윤리적 표현의 빈도가 매우 크게 나타났으므로 이 종류의 말뭉치를 활용할 때에는 보다 신중한 판단이 필요하다고 할 수 있다. 따라서 웹 문서를 수집하거나 활용할 때에는 가급적 리뷰나 블로그 문서를 보다 많이 수집 및 활용하는 것이 좋다고 할 수 있다. 또한, 메신저 대화의 경우, ‘2인 대화’로 수집된 말뭉치에 전반적으로 비윤리적 표현이 매우 많이 등장하므로, 이러한 사실을 말뭉치 수집이나 활용 시에 고려할 필요가 있다. 메신저 말뭉치는 다른 말뭉치에 비해 욕설 표현의 빈도가 매우 높게 나타났는데, 2인 대화뿐만 아니라 다자간 대화에서도 욕설 표현의 비율이 상당히 높다는 것을 확인할 수 있었다.

다음으로 우리는 문서 종류별로 말뭉치 내 비윤리적 표현의 정제 수준을 검토하였다. 문서의 비윤리성 판별에 대하여 일차적으로 '비윤리적인 문서'와 '비윤리적이지 않은 문서'로 구분하였고, 이차적으로 비윤리적 문서를 정제 수준에 따라 '상'과 '중'으로 분류하였다. '비윤리적 문서' 중 '상'은 비윤리적 표현의 등장 빈도가 해당 문서가 포함하고 있는 발화에 대하여 100% 이상[3]일 경우, '중'은 1%~99%일 경우에 해당한다. '비윤리적이지 않은 문서', 즉 '하'로 표시된 문서는 비윤리적 표현이 전혀 등장하지 않는 문서(0%)이다. 다음의 표는 문서 종류별로 정제 수준을 상중하로 분류한 표이다. 문서 종류별로 비윤리적 표현의 빈도를 '동일 종류 내 비율'과 '전체 문서 내 비율'도 함께 아래와 같이 도표화하였다.

3) 정제 수준은 문서에 포함된 전체 발화 수에 대한 비윤리적 표현의 비율(비윤리적 표현 수/전체 발화 수 *100, 단위 %)을 토대로 산출하였다. 다만, 비윤리적 표현에 대하여 개별 표현별 어절 단위로 태깅이 가능하도록 하였기 때문에 하나의 발화(또는 문장)에 두 개 이상의 비윤리적 표현이 등장하는 경우 문서 내의 전체 발화 수보다 비윤리적 표현이 많아질 수 있는데, 이러한 경우 비윤리적 표현의 비율이 100% 이상으로 계산될 수 있다.

문서 종류	정제 수준	문서 수 (건)	동일 종류 내 비율 (%)	전체 문서 내 비율 (%)
웹(블로그)	상	-	-	-
	중	168	8.195	0.041
	하	1,882	91.805	0.462
소계		2,050	100	0.503
웹(게시판)	상	9	1.741	0.002
	중	32	6.19	0.008
	하	476	92.07	0.117
소계		517	100	0.127
웹(리뷰)	상	-	-	-
	중	200	8.22	0.049
	하	2,233	91.78	0.548
소계		2,433	100	0.597
웹(누리소통망)	상	18,757	4.74	4.603
	중	14,858	3.755	3.646
	하	362,065	91.504	88.851
소계		395,680	100	97.1
일상 구어 대화	상	-	-	-
	중	574	25.809	0.141
	하	1,650	74.191	0.405
소계		2,224	100	0.546
메신저 대화(2인)	상	-	-	-
	중	1,255	28.108	0.308
	하	3,210	71.892	0.788
소계		4,465	100	1.096
메신저 대화(다자)	상	-	-	-
	중	60	46.512	0.015
	하	69	53.488	0.017
소계		129	100	0.032
전체	상	18,766		4.605
	중	17,147		4.208
	하	371,585		91.187
합계		**407,498**		**100**

<표 12> 비윤리적 표현 유형 빈도에 따른 말뭉치 문서 종류별 정제 수준 분류

주지할 것은 '하'로 분류한 '비윤리적이지 않음'의 개념이 '윤리적임'을 의미하는 것은 아니라는 것이다. 본 연구의 목적은 비윤리적인 표현이 등장하는 문서를 걸러 내어 공공 언어로서 공개 여부를 제고하는 것에 있기 때문에, 우리가 '비윤

리적이지 않음’이라고 분류하였다고 해서 그 말뭉치 문서 종류가 ‘윤리적임’이라는 뜻을 지닌 도덕적 판단이나 정당화를 부여하지는 않는다. 이는 우리의 정제 수준 분류 결과를 근거로 해당 말뭉치를 ‘윤리적 말뭉치’라고 이름 붙일 수 없다는 뜻이며, 그러한 도덕적 판단이나 정당화를 부여하는 것은 본 연구의 과업과는 그 차원을 달리한다는 것을 의미한다. 그렇기 때문에 우리가 연구 결과로서 어떤 종류의 말뭉치가 ‘비윤리적이지 않음’이라는 평가를 제시한다고 해서, 그것이 ‘도덕적으로 검증을 받았다고 하거나 혹은 도덕적으로 인정 혹은 허용되는 말뭉치’라는 뜻으로 어떤 인공지능 기술이나 제반 산업 분야에 적용되거나 활용되어서는 안 된다.

우리는 비윤리적 표현이 한 번이라도 등장하면 해당 문서는 비윤리적 문서로 규정하였다. 대면 대화와 비대면 대화 혹은 문서 등 말뭉치 종류의 다양성을 고려하여 비윤리적 표현의 범주에 ‘표현 자체가 함축하는 비윤리성’외에도 ‘담화 맥락에서 규정되는 비윤리성’도 포함하였다. 그리고 조사 대상자들에게 이 두 가지 범주의 비윤리적 표현을 모두 검토하도록 하였다. 이 두 요소를 중심으로 평가자는 말뭉치에 포함된 언어의 비윤리성을 판단하며 그 기준은 평가자의 도덕적 직관에 의존한다. 실제 생활 세계에 대한 직접적인 윤리 의식이 이 사회에서 통용되는 말뭉치의 사회적 인식 조사에 가장 근접한 판별 요소이기 때문이다. 이러한 사실에 의존할 때 문서의 비윤리성 정도는 평가자의 직관에 따른 비윤리적 표현의 등장 횟수에 의해 그 적절성이 확보될 수 있다.

우리는 조사 결과를 토대로 ‘비윤리적이지 않은 문서’는 일차적으로 공개하지 않고 우선 검토를 거쳐 공개 여부를 판정할 것을 권한다. 검토 및 정제 후 공개 여부를 국립국어원에서 결정할 수 있다. 또한 문서의 분할 공개 여부, 문건을 열람할 수 있는 권한자의 분류 여부도 고려해야 할 필요도 있을 것으로 보인다.

조사 결과 비윤리적 문서는 전체 조사 대상 말뭉치 중 8.8%에 불과하였지만 앞서 언급하였듯이 공공언어의 특성상 단 한 건의 비윤리적 표현이라도 사회적으로 용인되기 힘들기에 비윤리적 문서로 분류된 말뭉치 언어에 대한 철저한 검토와 정제가 필요하다고 판단된다.

제 5 장

보고서 활용 방안 및 정책 제언

5. 보고서 활용 방안 및 정책 제언

이 연구는 기존 국립국어원에서 구축한 말뭉치 언어 자료에서 비윤리적 표현이 있는지를 유형별로 조사하고 데이터를 분석하여 배포 가능한 정제 수준을 제시하는 것을 목표로 진행되었다. 우리는 국립국어원에서 기존에 구축하였던 모두의 말뭉치 중 일부 말뭉치에 나타난 언어를 대상으로 '사회적으로 용인되지 않는' 비윤리적 표현으로 분류할 수 있는 범주를 6가지로 설정하고(혐오 표현, 성적 표현, 욕설 표현, 차별적 표현, 개인정보, 기타) 범주별 빈도를 산출하기 위해 성별, 연령, 지역, 직업 등 다양한 변인으로 분류되는 일반 언어 사용자 100명을 조사 평가자로 활용하여 그들의 직관에 근거한 태깅 수치와 비율을 비교 분석하였다.

본 보고서는 연구의 목적 수립부터 데이터 분류 및 구축, 납품 상세 내역에 이르기까지 우리가 수행한 연구 진행 과정 전체를 과정별로 정리한 것으로서 향후 유관한 연구를 발주하고 이를 진행함에 있어 선행 사례로 활용될 수 있을 것이다. 특히, 전무하다시피 한 비윤리적 언어 검출 유형을 다소 거칠지만 기존 연구를 통하여 비윤리적 표현의 유형을 6개로 분류하였다는 것은 후속 연구에 도움이 될 것으로 사료된다. 후속 연구를 통해 이 6개의 분류 유형 기준을 상세화 및 정교화할 수 있다면 비윤리적 언어 검출 시스템 개발에 도움이 될 것이다.

아울러, 본 보고서의 제4장(조사 결과 분석)에서 두 가지(평가자 변인별 분석, 말뭉치 문서 종류별 분석) 차원에서 분석한 결과는 향후 말뭉치 언어의 수집 및 활용 과정에서 정책의 방향을 설정하거나 개선의 과제를 수립하는 데 몇 가지 주요한 시사점을 줄 것으로 기대한다.

말뭉치 언어에 대한 사회 및 산업계의 관심이 높은 점을 고려할 때, 향후 국립국어원에서 구축하는 말뭉치 언어가 지속 가능한 수준으로 유지 및 개선되기 위해서는 이번 연구의 후속으로 보다 심도 깊은 추가 연구가 가능하도록 다양한 사업 발주를 통한 연구 지원이 필요할 것으로 사료된다. 예컨대 비윤리적 언어 표현에 대한 보다 폭넓은 윤리학적 배경에 대한 연구를 토대로 여기서 우리가 제시한 비윤리적 표현 유형 분류 체계를 더 상세화하는 한편, 전문가 전문과 같은 정성적인 평가 검증 과정을 추가하여 신뢰할 수 있고 활용도가 높은 연구 결

과를 제시할 수 있는 후속 연구가 필요할 것으로 생각된다. 이 보고서는 이러한 후속 사업을 위한 토대로 활용될 수 있을 것이다.

<Abstract>

A Study on Social Recognition and Classification of Corpora Languages

The purpose of this study is to investigate the inappropriate expression and content of general language users, categorize the types of unethical expressions, and suggest the level of immorality to decide if the corpus need to be refined for the corpora built by National Institute of Korean Language.

To that end, we set 5 categories of unethical expressions, including “hatred”, “discrimination”, “sexual expression”, and “cursing”, and “other unethical expressions” that do not belong to them. In addition, “personal information” that cannot be disclosed to the public was also targeted for identification. Accordingly, 100 evaluators were recruited considering gender, age, region and conducted a survey on the unethical expressions and personal information for 25,190,902 claus, 407,498 document.

In the course of the investigation, evaluation guidelines and evaluator education were provided for all evaluators. Also the results of the investigation were divided into two parts, and the accuracy of the data was improved through expert consultation. To do this efficiently, evaluators used evaluation tools produced by the participating organization Media Corpus. The whole results were categorized in the form of JSON files and Excel files respectively, and the contents of the statistical analysis and visualized graphs were presented in this paper. Finally, we summarized the utilization methods and related policy recommendations of this study.

Keywords: Corpus language, Unethical expression, Unethical expression type, Unethical expression sensitivity, Unethical expression classification guidelines.

Project Director: Lee Chankyu(Chung-ang University)

참고문헌

김진웅(2021), 자연언어처리에서 윤리적 문제와 해결 방안: 연령 및 지역 편향성 극복의 출발점으로서 방언자료 수집, *연구방법논총, 6*(1), 157-180.

김형주・변순용・김민수(2018), 한국 초등 도덕과 교육과 관련된 독일 도덕 교과서 분석 연구 - 가치영역별 도덕 역량 지향 모델을 중심으로, *초등도덕교육, 59,* 165-196.

변순용(2020), 데이터 윤리에서 인공지능 편향성 문제에 대한 연구, *윤리연구, 1*(128), 143-158.

변순용(2020), AI 시민성 교육에 대한 시론. *초등도덕교육, 67,* 427-445.

서상규・김형정(2005), 구어 말뭉치 설계의 몇 가지 조건. *언어 사실과 관점, 16,* 5-29.

윤상오(2018), 인공지능 기반 공공서비스의 주요 쟁점에 관한 연구: 챗봇(ChatBot) 서비스를 중심으로, *한국공공관리학보, 32*(2), 83-104.

이청호・김봉제・김형주・변순용・이찬규, (2021). 윤리적 인공지능을 위한 비도덕 문장 판별 온톨로지 구축에 대한 연구, *인공지능인문학연구, 7,* 149-170.

조태린(2021), 언어의 품격과 공공언어의 품격(성) 문제에 대한 비판적 고찰, *문법 교육, 41,* 97-124.

조태린・김신각・신유리・공나형・신아영(2018), 비윤리적 언어 표현의 의미 자질에 따른 하위 유형 연구 - 대화형 인공지능의 윤리적 언어 표현을 위한 기초 연구(2) -, *한민족문화연구, 63*(63), 147-184.

최현철・변순용(2019), 인공적 도덕 행위자에 대한 융합접근의 철학적 기획, *윤리연구, 1*(124), 1-16.

Armstrong, H.(2015), *Machines that learn in the wild: Machine learning capabilities, limitations and implications.* Technical report, Nesta, London, England

Hosseini, H, Kannan, S, Zhang, B, & Poovendran, R.(2017), *Deceiving Google's perspective API built for detecting toxic comments,* arXiv 1702:08138.

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M.(2018). Benchmarking aggression identification in social media. In Proceedings of the First Workshop on Trolling, *Aggression and Cyberbullying* (TRAC-2018): 1-11.

Vakkuri, V., Kemell, K. -., Jantunen, M., Halme, E., & Abrahamsson, P.(2021). ECCOLA — A method for implementing ethically aligned AI systems. *Journal of Systems and Software, 182* doi:10.1016/j.jss.2021.111067

방송통신심의위원회(2020), *방송심의에 관한 규정,* 방송통신심의위원회 규칙 제150호.

방송통신심의위원회(2019), *방송언어 가이드라인.*

<기획·연구>

국립국어원 이승재 언어정보과장

국립국어원 유희정 학예연구사

국립국어원 한송이 연구원

<사업 참여자>

사업 책임자 이찬규(중앙대학교 인문콘텐츠연구소 소장)

사업 참여자 김민수(동서울대학교)

박일섭(미디어 코퍼스)

김보현(중앙대학교 인문콘텐츠연구소)

이은재(중앙대학교 인문콘텐츠연구소)

조재원(미디어 코퍼스)

안윤(미디어 코퍼스)

현재홍(미디어 코퍼스)

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금낭화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2022년 3월 15일

발행일: 2022년 3월 17일

인 쇄: 도서출판 이삭

※ “이 책은 국립국어원의 용역비로 수행한 ‘말뭉치 언어의 사회적 인식 조사·분류’ 사업의 결과물을 발간한 것입니다.”