

대규모 악취 민원 정보 핵심 키워드 추출 LLM 모델 개발

이호정, 황병일, 김동주, 서영주, 김다현*

포항공과대학교 인공지능연구원

ghwjd0112@postech.ac.kr, bihwang@postech.ac.kr, kbb0320@postech.ac.kr,
yjsuh@postech.ac.kr, *kdhyun8011@postech.ac.kr

Development of an LLM Model for Extracting Key Keywords from Large-Scale Odor Complaint Data

Lee Ho Jung, Hwang Byeong Il, Kim Dong Ju, Suh Young Joo, Kim Da Hyun*

POSTECH Institute of Artificial Intelligence

요약

본 연구는 산업 도시에서 급증하는 악취 민원 처리를 효율화하기 위해 소형 대형언어모델(Small Large Language Model, sLLM) 기반의 악취 민원 자동 키워드 추출 시스템을 개발하였다. 기존 텍스트 마이닝 기반 접근은 비정형 표현 처리와 실시간 대응에 한계가 있어, 본 연구에서는 대형 언어 모델(Large Language Model, LLM)을 활용한 도메인 특화 모델을 제안하였다. 환경 공무원 인터뷰를 통해 핵심 키워드(발생 위치, 악취 종류, 원인 추정 지역, 강도 변화, 지속 시간)를 정의하고, 포항시 실제 민원 904건과 GPT-4 기반 데이터 증강을 통해 학습 데이터를 구축하였다. 한국어 특화 sLLM을 full fine-tuning으로 학습한 결과, 모든 모델의 성능이 향상되었으며 DeepSeek-R1-Distill-Qwen-1.5B가 **재현율 94.9%, F1-score 95.0%**로 최고 성능을 보였다. 제안 시스템은 악취 민원 데이터베이스 구축을 자동화하여 수기 처리의 비효율을 해소하고, 향후 음성 민원(음성-텍스트 변환 연동) 및 악취 발생 예측 시스템으로 확장 가능성을 지닌다.

I. 서론

최근 산업화와 도시화로 악취 민원이 급증하며, 포항시와 같은 산업 도시에서 공장, 폐기물 처리 시설 등으로 인한 악취가 주민 생활을 저해하는 사회적 문제로 대두되고 있다. 현재 악취 민원은 주로 전화나 인터넷을 통해 접수되며, 공무원이나 관련 기관에서 이를 수기 방식으로 데이터베이스(Database, DB)에 기록하고 대응하고 있다. 이 과정에서 민원 내용의 핵심 정보를 추출하고 분류하는 작업이 필수적이거나, 수기 처리의 한계가 뚜렷하게 드러난다. 특히 주거지역 인근에서 악취가 발생할 경우, 동시에 다발적으로 많은 민원이 한꺼번에 들어오는 현상이 빈번하다. 예를 들어, 특정 공장에서 화학 물질 누출이 발생하면 수십에서 수백 건의 민원이 단 시간에 집중될 수 있으며, 이를 수기로 정리하고 DB에 입력하는 데는 상당한 시간과 인력이 소요된다. 이로 인해 민원 처리 지연, 정보 오류 발생, 그리고 대응 전략 수립의 비효율성이 초래된다. 또한, 민원 텍스트의 자연어 특성상 비정형적 표현, 모호한 서술 또는 감정적 언사가 포함되어 정확한 키워드 추출이 어렵다.

이에 본 연구는 인공지능(Artificial Intelligence, AI), 구체적으로 대형 언어 모델(Large Language Model, LLM)을 활용한 악취 민원 자동화 시스템을 제안한다. AI를 통해 민원 텍스트에서 핵심 키워드를 자동 추출함으로써, 각 키워드(발생 위치, 악취 종류, 원인 추정 지역, 악취 강도 변화, 지속 시간)에 따른 민원 DB를 체계적으로 구축할 수 있다. 이렇게 구축된 DB는 데이터 분석을 통해 악취 발생 패턴을 파악하고, 예방적 해결 방안을 모색하는 데 활용될 수 있다. 궁극적으로 이는 악취 저감 정책 수립에 기여하며, 주민들의 삶의 질 향상과 환경 관리 효율성을 높이는 데 목적이 있다. 본 연구는 소형 LLM(Small Large Language Model, sLLM)을 기

반으로 한 모델 개발을 중점으로 하며, 실제 포항시 민원 데이터를 바탕으로 실증적 접근을 취한다.

II. 관련 연구

1. 민원 텍스트 분석 관련 연구

국내 연구에서는 딥러닝과 텍스트 마이닝을 활용한 민원 분석이 주목받고 있다. 한 연구에서는 개체명 인식모델을 미세 조정하여 120 다산콜 민원 텍스트에서 공간 정보(자치구, 도로, 건물 번호)를 추출하고, 지오코딩과 지리정보시스템으로 민원 맵핑을 구현하였다[1]. 다른 연구에서는 공공 민원 데이터를 텍스트 마이닝과 토픽 모델링으로 분석하여 강동구 사례를 통해 공간복지 향상을 위한 모델을 제시하였다[2]. 또 다른 연구에서는 유해화학물질 민원에 잠재 의미 분석과 잠재 디리클레 할당을 적용하여 주요 토픽(배관 관리, 사고 예방 등)을 도출하고, 수요자 중심 정책을 제안하였다[3]. 대부분 연구가 전통적 머신러닝이나 간단한 딥러닝에 의존해 비정형적 표현 및 모호한 서술 처리에 한계가 있으며, 소규모 데이터셋으로 일반화가 어렵고, 복잡한 문맥 반영과 실시간 시스템 적용이 부족하여 새로운 방법이 필요하다.

2. 악취 민원 특화 연구

악취 민원에 특화된 연구는 상대적으로 적으나, 머신러닝과 AI를 활용한 사례가 증가하고 있다. 한 연구에서는 제강 사업장을 대상으로 머신러닝을 활용해 악취 비산을 예측하고, 악취 물질 농도와 민원 발생을 연계 분석하여 민원 최소화 전략을 도출하였다[4]. 다른 연구에서는 주거-산업 복합지역에서 AI를 이용해 악취 농도를 예측하고, 민원 증가 원인을 분석하

였다[5]. 또 다른 연구에서는 텍스트 마이닝을 통해 환경 민원에서 키워드를 추출하고, 머신러닝으로 우선순위를 부여하였다[6]. 그러나 악취 민원 특화 연구가 부족해 도메인 특화 키워드 추출이 미흡하고, 센서 데이터나 단순 텍스트 마이닝에 의존하며 LLM 활용이 드물다. 소규모 데이터셋으로 다양한 패턴을 포괄하지 못하고, 실시간 시스템 적용이 제한적이기 때문에 관련 연구가 필요한 실정이다.

III. 실험

1. 민원 키워드-텍스트 쌍 데이터셋 구축

악취 민원 전문가(환경 공무원, 악취 관리 전문가)들의 인터뷰를 통해 핵심 키워드를 설정하였다. 인터뷰 결과, 악취 민원에서 가장 중요한 정보로 발생 위치(예: 아파트, 사거리), 악취 냄새 종류(예: 타는 냄새, 썩는 냄새), 악취 원인 추정 지역(예: 인근 공업단지, 매립장), 악취 냄새 강도의 변화(예: 점점 강해짐, 갑자기 약해짐), 악취 지속 시간(예: 1시간 이상, 간헐적) 등 5가지 키워드를 선정하였다. 이는 민원 대응 시 필수적인 요소로, 발생 위치, 냄새 종류, 원인 추정 지역은 악취 원인을 신속히 파악하고 해당 기업이나 시설에 제재를 가할 근거를 마련하는 데 도움이 되며, 악취 지속 시간과 냄새 강도의 변화는 기상 데이터와 조합하여 공무원들이 악취 원인 지역을 직접 추정하고 예방 조치를 취하는 데 핵심적으로 활용된다. 초기 데이터는 포항시의 실제 민원 대장에서 904건을 기반으로 수집되었으며, 각 민원 내용에 대해 선정한 5개의 키워드를 매핑하여 구조화된 데이터로 변환하였다. 그러나 개인정보 보호 등의 이유로 실제 민원 텍스트 데이터를 대규모로 확보하는 것이 불가능하여, 생성형 인공지능(GPT-4 기반 모델)을 활용해 민원 텍스트를 증강하였다. 증강 과정에서는 협조적, 화가 난, 불확실한, 긴급 상황 등 다양한 시나리오를 반영하여 정보 불확실성, 감정 표현을 포함한 텍스트를 생성, 실제 민원 상황의 다양성을 모방하여 모델의 일반화 능력을 강화하였다. 최종적으로 총 904건의 데이터를 생성하였으며, 이 중 859개를 훈련 데이터로, 45개를 테스트 데이터로 분할하였다.

2. 모델 학습 및 결과 분석

비교를 위해 한국어 처리에 강한 sLLM 모델을 선정하였다: Qwen3-0.6B, Llama-3.2-1B, DeepSeek-R1-Distill-Qwen-1.5B, Ko Alpaca-Polyglot-5.8B. 이러한 모델들은 파라미터 규모가 작아 실시간 추론에 적합하며, Hugging Face에서 제공되는 미세 조정 가능한 버전이다. 훈련은 full fine-tuning 방식으로 진행하였다. 전체 파라미터를 학습함으로써 도메인 적응을 강화하였으며, 에폭 5로 설정하여 과적합을 방지하였다. 입력 형식은 민원 텍스트를 입력받고, 출력으로 5가지 키워드를 구조화된 형제(예: "발생 위치: 공장, 악취 종류: 화학 냄새")로 생성하도록 설계하였다. 비교를 위해 fine-tuning 전(기본 모델)과 후의 성능을 평가하였다.

성능 평가는 단어 겹침을 기반으로 한 재현율과 F1-score를 사용하였다. 이 함수를 사용한 이유는 키워드 추출의 특성 때문이다. 악취 민원 키워드는 짧고 구체적인 단어/구로 구성되며, 정확한 일치보다는 의미적 겹침이 중요하다. 기존 지표(예: BLEU)는 문장 수준에서 적합하나, 키워드 추출에서는 단어 집합 기반의 Recall(실제 키워드를 얼마나 포괄적으로 추출했는가)과 F1-score(정밀도와 재현율의 균형)가 더 적합하다. 특수문자/번호 제거를 통해 노이즈를 최소화하고, 한국어의 조사/어미 변형을 고려하여 견고한 평가를 가능하게 한다. 테스트 데이터 45건에 대해 평균 재현율과 F1-score를 계산하였다.

결과는 [표 1]과 같다. Fine-tuning 후 모든 모델에서 성능이 크게 향상

되었으며, DeepSeek-R1-Distill-Qwen-1.5B가 재현율 94.9%, F1-score 95.0%로 최고 성능을 보였다. KoAlpaca-Polyglot-5.8B는 fine-tuning 전 재현율 3.2%, F1-score 1.2%로 매우 낮았으나, fine-tuning 후 재현율 92.8%, F1-score 93.1%로 큰 개선을 보였다. 이는 도메인 적응의 효과와 한국어 특화 사전 학습의 이점을 입증한다.

[표 1] sLLM 모델 Fine-Tuning 전후 모델별 성능 비교

Model	Base		Fine-tuning	
	Recall	F1	Recall	F1
Qwen3-0.6B	0.741	0.753	0.944	0.943
KoAlpaca-Polyglot-5.8B	0.032	0.012	0.928	0.931
Llama-3.2-1B	0.374	0.168	0.947	0.948
DeepSeek-R1-Distill-Qwen-1.5B	0.288	0.065	0.949	0.950

IV. 결론

본 연구는 악취 민원 텍스트에서 핵심 키워드를 자동 추출하는 LLM 기반 시스템을 개발하여 수기 DB 구축의 비효율성을 해결하고자 하였다. 관련 연구 검토를 통해 기존 텍스트 마이닝과 머신러닝의 한계를 확인하고, 본 연구의 LLM 활용이 이를 보완할 수 있음을 입증하였다.

현재 시스템은 인터넷 텍스트 민원만 처리 가능하며, 전화 민원은 음성-텍스트 변환(Speech-to-Text, STT) 기술 제약으로 지원하지 못하나, 향후 전화망과 STT 모델 연동을 통해 확장할 예정이다. 이 시스템은 포항시 등 지방자치단체에서 실시간 민원 처리와 데이터 기반 악취 저감 정책 수립에 활용될 수 있으며, 장기적으로 악취 발생 핫스팟 매핑과 예방 정책 제안을 통해 환경 문제 해결에 기여할 것으로 기대된다.

ACKNOWLEDGMENT

이 연구는 2025년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업(RS-2022-NR070870) 지원을 받아 수행되었으며, 과학기술정보통신부-경찰청이 공동 지원한 '폴리스랩 2.0 사업'(RS-2023-00281072)과 정부(과학기술정보통신부)의 재원으로 한국지능정보사회진흥원(NIA)의 지원을 받아 수행된 「2024년 지역특화형 초거대 AI 클라우드 팜 실증 및 확산 환경 조성」 사업(2025-지능기술인프라-위25)의 지원을 받아 수행된 연구입니다.

참 고 문 헌

- [1] 박진홍, 강민규. (2024). 개체명 인식 모델링을 통한 120 다산콜 민원 텍스트의 공간정보 추출 및 기초분석.
- [2] 신동윤. (2023). 거주민 공간복지 향상을 위한 공공 개방 민원 데이터 분석 모델.
- [3] 김강현, 홍정열. (2023). 텍스트마이닝을 활용한 유해화학물질 관련 민원의 주요 토픽 분석.
- [4] 김동한, 오용환. (2024). 머신러닝 기반 악취 비산 영향 예측 및 민원 발생 최소화를 위한 통합 모델 구축.
- [5] 이송은, 정아현. (2024). 주거-산업 단지 내 악취물질 농도 예측을 위한 AI 기반 시계열 예측 모델의 성능 평가.
- [6] Fabiana Manservigi, Michele Banzi. (2023). Environmental complaint insights through text mining based on the complaints.