대규모 언어모델(LLM) 기반 AI 댓글 탐지: 제21대 한국 대통령 선거 사례 연구

(A Case Study on AI Comment Detection Based on Large Language Models: The 21st Korean Presidential Election)

김신형*,유예원*,박지은*,노상희*,손남례**

(Kim Sin Hyeong,* Yu Ye Won*, Park Ji Eun*, Noh Sang Hui*, Nam Rye Son**)

요 약

본 연구는 온라인 정치 뉴스 댓글 공간에서 대규모 언어모델(LLM) 기반의 생성형 인공지능(AI) 댓글이 개입하는 현상을 탐지하고 그 사회적 함의를 분석한다. 이를 위해 2022년 제20대 대통령 선거 기간 동안 수집된 네이버 뉴스 댓글 데이터를 활용, Google Gemini 2.5 Flash와 OpenAI GPT-4를 이용해 생성한 AI 댓글과 인간 댓글을 구분하는 분류 모델을 구축하였다. 모델 학습에는 한국어에 특화된 사전학습 언어모델인 KOELECTRA와 KCELECTRA를 활용했으며, 특히 KCELECTRA 모델은 99% 이상의 높은 정확도를 기록했다. 이 모델을 2025년 제21대 대통령 선거 댓글 데이터에 적용한 결과, 전체 댓글 중 약 1.46%~3.80%가 AI 댓글로 분류되었다. 이는 실제 선거 여론 공간에 AI 기반의 인위적 개입이 존재했음을 시사하며, 탐지된 수치는 최소치로 해석할 필요가 있다. 본 연구는 대선이라는 특정 정치적 맥락 속 LLM 기반의 AI 개입을 실증적으로 분석한 최초의 연구라는 점에서 학술적 의의가 있으며, 향후 공정한 여론 환경 조성을 위한 지속적인 모니터링 및 탐지 기술의 고도화 필요성을 강조한다.

■ 중심어 : 대통령 선거 댓글 ; 생성형 인공지능 ; AI 댓글 탐지 ; 여론 조작 ; 대규모 언어모델

Abstract

This study investigates the potential intervention of AI-generated comments in online political news discussions and examines their implications for democratic discourse. Using comment data collected from the 20th Korean presidential election (2022), we built a classification model to distinguish between human-written and AI-generated comments. AI-generated comments were produced using Google Gemini 2.5 Flash and OpenAI GPT-4, and Korean-specific pretrained language models such as KoELECTRA and KcELECTRA were employed for training. Experimental results show that the KcELECTRA model achieved an accuracy of over 99%. Applying the trained model to comment data from the 21st presidential election (2025) revealed that approximately 1.46% to 3.80% of the comments were classified as AI-generated. This finding indicates that artificial intervention using generative AI was present in the electoral discourse, though the detected rate likely represents a minimum estimate. As one of the first empirical studies to analyze AI intervention in the context of a national election, this work underscores the importance of advancing detection techniques and continuous monitoring to safeguard fair and transparent public opinion formation in the era of generative AI.

■ keywords : Presidential Election Comments ; Generative Artificial Intelligence ; AI-Generated Comment Detection ; LLM

Ⅰ. 서 론

현대 사회에서 다양한 의견은 온라인 매체를 통해 실시간으로 공유된다. 그중 온라인 뉴스 댓글은 정치적 의견 형성의 공론장으로 자리 잡았다. 한국리서치 여론조사에 따르면, 전체 응답자의 55%가 댓글이 뉴스 독자의 생각에 영향을 미친다고 답했으며, 이는 댓글이 전혀 영향을 미치지 않는다는 응답(14%)의 약 3.9배에 달하는 높은 수치다. 전체 응답

* 준회원, 전남대학교 빅데이터융합학과
** 정회원, 전남대학교 소프트웨어중심대학사업단
이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2021-0-01409)
접수일자 : 2025년 08월 21일
수정일자 : 2025년 09월 09일
게재확정일 : 2025년 09월 24일
교신저자 : 손남례 e-mail : nrson72@gmail.com

자의 42%는 댓글이 많은 뉴스를 검색하거나 선택해 읽고, 27%는 뉴스를 읽기 전에 댓글을 먼저 확인한다고 응답하였다[1]. 이는 댓글이 다른 독자의 생각에 강한 영향력을 미치고 있음을 보여준다.

댓글이 여론 형성에서 차지하는 역할이 커질수록 이를 인위적으로 조작하려는 가능성이 커지고 있다. 소셜봇은 소셜 미디어에서 존재하며 텍스트를 자동 생성하고 인간처럼 행동하는 소프트웨어 에이전트를 말한다. 이들은 뉴스 확산, 정치 여론 형성, 여론 조작 등 사회 및 정치적 영역에서 인간 형태를 모방하며 악용될 수 있다. 최근에는 ChatGPT와 같은 대규모 언어 모델(LLM)의 확산이 인간과 같은 자연스러움을 지닌 글을 대량 생산이라는 특성이 있어 문제의 심각성을 키웠다.

OpenAI 보고서는 중국, 러시아, 이란 등 일부 국가·조직이 AI 도구를 활용해 정치 관련 게시물과 댓글을 대량 생산하여 특정 입장을 강화하거나 여론을 조작한 정황을 제시한다[2]. 또한 2025년 대한민국의 대통령 선거를 앞둔 국면에서도 유사한 우려가 제기되었는데, Cybrea(2024)는 2024년 4 - 5월 TikTok과 X(구 Twitter) 활동 계정 중 약 29%를 가짜 계정으로 추정하고, 이들이 선거 관련 허위 정보 유포에 관여했다고 분석하였다[3]. 이처럼 기술 발전은 조작의 비용을 낮추고 정교함을 높이는 방향으로 작동하여 댓글 공론장의 정당성을 위협한다.

이러한 문제로 이를 탐지하려는 시도가 이어져 왔다. GPT 계열 텍스트 탐지, 소셜 미디어 내 가짜 뉴스 판별 등 다양한 접근이 제시되었고[4,5] 국내에서도 XDAC (XAI-Driven Detection and Attribution of LLM-Generated Comments)는 한국어 뉴스 댓글 데이터 셋을 기반으로 인간 댓글과 LLM 기반 AI 댓글을 분류하고 생성 주체 모델을 식별하는 시스템을 구축하였다[6]. 그러나 기존 연구는 몇 가지 한계를 갖는다. 첫째, 해외 연구에서 보고된 GPTZero나 OpenAI ‘AI 텍스트 탐지기’의 사례처럼 탐지기의 오판율이 높아[7] 실제 적용에는 어려움이 따른다. 둘째, XDAC과 같은 국내 모델은 학술적 의미가 크지만, 아직 상용화 단계에 이르지 못해 실제 선거 환경에서 즉각적으로 활용되기 어렵다. 따라서 기존 연구는 탐지 가능성을 제시하는 수준에 머무르는 경우가 많고, 실제 정치적 이벤트에서 AI 개입을 실증적으로 확인한 연구는 부족하다. 특히 선거와 같은 민감한 정치적 맥락에서는 일부 우려와 의문이 제기되었을 뿐, 이를 실증적으로 규명한 연구는 거의 이루어지지 않아 여전히 부족한 실정이다. 이러한 한계를 보완하기 위해 본 연구를 진행하고자 한다.

댓글은 선거와 같은 정치적 이벤트 시기에는 유권자의 의견 표출과 토론이 집중되는 주요 장이 된다. 이에 본 연구는 제21대 대통령 선거에서 댓글에 AI 개입이 실제로 존재했는지, 나아가 그 개입이 공론장의 의견 형성에 어떠한 흔적을 남겼는지를 확인하고자 했다. “제20대 대통령 선거 시기의 인간 댓글과 AI 댓글로 학습한 모델이 제21대 대통령 선거 댓글 속 AI 존재 가능성을 식별할 수 있는가?”라는 핵심 질문을 중심으로 실험을 설계했다. 2022년 제20대 대선 기간에 수집한 네이버 뉴스 댓글을 기반으로 인간 댓글과 AI 생성 댓글을 구분하는 분류 모델을 학습한다. 이후 해당 모델을 2025년 제21대 대선 기간의 실제 댓글 데이터에 적용하여 AI 개입 가능성을 정량적으로 평가한다. 따라서 본 연구는 공론장의 투명성을 검토하며 향후 AI 생성 글이 지니는 가능성과 위협성을 평가하는 연구가 될 것으로 기대된다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 고찰하고, 3장에서는 데이터 수집과 전처리 과정을 기술한다. 4장에서는 실험 설계 및 결과를 제시하고, 마지막으로 5장에서는 결론과 향후 연구 방향을 제언한다.

Ⅱ. 관련 연구

1. 대규모 언어모델(Large Language Model)

대규모 언어모델(LLM)은 방대한 말뭉치를 기반으로 사전 학습된 심층 신경망으로, 텍스트 생성, 요

약, 분류 등 다양한 자연어 처리(NLP) 작업에 활용된다. 대표적으로 GPT-3는 소수의 예시만으로도 다양한 작업을 수행할 수 있는 Few-shot 학습 능력을 입증하며 LLM의 잠재력을 보여주었다[8]. 본 연구에서는 AI 댓글 생성을 위해 Google Gemini 2.5 Flash와 OpenAI GPT-4를 사용하였으며, 두 모델은 Transformer 아키텍처를 기반으로 복잡한 문맥을 효과적으로 이해하고 자연스러운 한국어 문장을 생성할 수 있도록 설계되어 있다[9]. 그러나 LLM 기반 텍스트는 문법적 완결성과 문체적 일관성을 갖추고 있어 인간이 작성한 댓글과의 구분이 어렵다. 이러한 특성은 허위 정보 확산과 여론 조작에 악용될 위험성을 높인다[5, 10]. 실제로 LLM이 생성한 텍스트는 인간이 속을 정도로 자연스러울수록 탐지가 더욱 어렵다는 점이 보고된 바 있다. 또한 LLM의 사회적·윤리적 위험성을 분석한 연구에서는 대량 생성된 AI 텍스트가 민주적 공론장을 왜곡하고 정보 신뢰성을 저해할 수 있다는 점을 경고하였다[11].

OpenAI(2023)[9] 보고서와 Cybrea(2024)[12]의 분석에 따르면, 일부 국가와 조직은 LLM을 활용해 정치적 선전과 선거 관련 허위 정보를 대규모로 생산 및 유포한 것으로 나타났다. 이러한 맥락에서 LLM 기반 텍스트 탐지는 단순한 기술적 과제를 넘어, 공정한 여론 형성과 민주주의의 투명성을 지키기 위한 필수적 사회적 대응으로 간주되어야 한다. 본 연구는 이러한 문제의식에 따라, 국내 대통령 선거라는 정치적으로 민감한 맥락에서 실제 AI 댓글 개입 가능성을 정량적으로 탐지하는 것을 주요 목표로 한다.

2. 학습 모델 및 성능 특성

AI 댓글 탐지를 위해 주로 활용되는 모델은 전통적인 시퀀스 모델인 LSTM과 Transformer 계열의 사전학습 언어모델이다.

(1) LSTM

LSTM(Long Short-Term Memory)은 RNN 계열의 순환 신경망으로, 문장 내 단어들의 순서와 장기 의존성을 효과적으로 학습할 수 있다[13]. Chang & Masterson(2020)은 정치적 텍스트 분류에 LSTM을 적용하여 약 87%의 정확도를 기록하였으며, 순차적 패턴 인식과 문맥 정보 유지에 강점을 보였다[14].

(2) BERT

KLUE-BERT는 한국어 언어 이해 벤치마크(KLUE) 데이터셋을 기반으로 학습된 BERT 계열 모델로, 문서 분류와 문장 추론 등 다양한 자연어 처리 과제에서 우수한 성능을 발휘한다[15]. 특히 혐오 발언 탐지나 댓글 분류와 같은 한국어 비정형 텍스트 분석 분야에서 CNN이나 Bi-LSTM 등 전통적인 모델보다 F1-score가 높다[16,17].

(3) ELECTRA 계열 모델

KoELECTRA는 ELECTRA 구조를 기반으로 한국어 데이터에 특화된 사전학습을 수행한 모델로, K-MHaS와 같은 혐오 발언 데이터셋에서 높은 성능을 기록하며 비정형 텍스트 분류에 적합한 모델로 평가된다[18]. KcELECTRA는 KoELECTRA를 기반으로 추가 사전학습을 거쳐 한국어 댓글 및 소셜미디어 텍스트 분석에 최적화된 모델이다. 이모티콘, 속어, 줄임말 등 실제 댓글에서 자주 사용되는

모델	특징	장점	한계
LSTM	RNN 구조, 장기 의존성 학습	단어 순서·문맥 반영, 비교적 단순하고 해석 용이	긴 문장 처리에 취약, 병렬화 한계
BERT	Transformer 기반, 양방향 문맥 학습	다양한 NLP 태스크에서 SOTA 성능, 자연어 이해에 강함	학습 비용이 크고, 긴 문장 처리에 제약
KoELECTRA	한국어 말뭉치 기반 학습	한국어 텍스트 분류에 강력, 학습 효율성 우수	구어체, 댓글 데이터 적용에 한계
KcELECTRA	댓글·SNS 데이터 추가 학습	비정형 텍스트(속어, 이모티콘 등) 처리에 특화	문맥적 성능 향상에 제한적

표 1. 모델별 특징과 한계점

표현을 효과적으로 처리할 수 있는 장점을 지니며, Kim et al.(2025)의 SNS 기반 그루밍 대화 탐지에 활용되어 높은 정확도를 기록하였다[19].

이러한 모델들의 특성과 장단점을 정리하면 표 1과 같다. 이처럼 생성형 AI가 작성한 텍스트를 탐지하기 위해 다양한 모델을 시도하고 있으며, 일상적 텍스트에서 AI가 얼마나 사용되었는지를 파악하려는 움직임이 활발해지고 있다[20]. 특히 한국어 댓글 연구는 감정이나 특징적 표현을 파악하는 데 강점을 보여, 댓글의 감정 및 정치적 성향을 주제로 활용하는 연구 성과가 주로 보고되었다[21, 22]. 그러나 이러한 연구는 특정 시기의 데이터나 실험 환경에 머물렀으며, 이를 기반으로 다른 시기 데이터에 변형 적용하거나 AI 댓글 사용 여부를 직접 판별하는 연구 진행은 미흡하다. 더불어 실제 선거와 같은 민감한 정치적 상황에서 AI 개입을 정량적으로 탐지한 사례는 드물다. 따라서 본 연구는 이러한 한계를 보완하여, 국내 대선이라는 실질적 사회·정치적 맥락에서 AI 댓글 개입 가능성을 검증하고자 한다.

Ⅲ. 데이터셋 구축

1. 제20대 대통령 선거 뉴스 기사 댓글

(1) 기사 수집

제20대 대통령 선거 뉴스 기사 데이터셋은 후보자 등록 1일 전인 2022년 2월 12일부터 개표 4일 후인 2022년 3월 13일까지 총 30일을 수집 기간으로 설정하였다. 수집 대상은 네이버 뉴스의 주요 6개 언론사이며, 수집 도구로는 Python 환경에서 BeautifulSoup 라이브러리를 사용하였다. 언론사별 '댓글 많은 순' 랭킹 페이지는 BeautifulSoup을 이용하여 HTML 요소를 파싱하였다. 각 페이지에서 날짜별 상위 20개 기사를 대상으로, 기사 제목은 태그에서 추출하였으며, 기사 URL은 <a class="_es_pc_link"> 태그에서 파싱하였다. 이중 정치와 무관한 기사는 연구 목적에 부합하지 않으므로 수동으로 필터링하였으며, 중복 기사를 제거한 뒤 날짜별 상위 5개 기사를 최종 선정하였다. 결과적으로 6개 언론사 × 30일 × 5개 기사 = 900개 기사를 확보하였다. 선정된 900개 기사의 URL을 이용하여 기사 본문과 댓글 통계(성별·연령 분포)를 크롤링함으로써 기사 데이터셋 구축을 완료하였다.

(2) 댓글 수집

1) Human 댓글 수집

각 뉴스 기사의 전체 댓글 영역을 Selenium을 활용하여 스크롤링을 반복함으로써 모든 HTML 블록을 로딩 및 수집하였다. 네이버 '클린봇' 기능은 비활성화하여 필터링되지 않은 원본 댓글을 확보하였다. 수집된 HTML은 BeautifulSoup을 이용하여 태그에 포함된 댓글 본문과 태그에 기록된 공감 수를 추출한 뒤, 데이터프레임 형태로 구조화하였다. 기사별 댓글을 공감 수 기준으로 내림차순 정렬한 뒤, 상위 100개의 댓글을 우선 추출하였다. 이 과정을 통해 총 90,000개의 Human 댓글 데이터를 확보하였으며, 해당 데이터에는 라벨 0을 부여하였다. 본 연구에서는 2022년 시점의 인터넷 댓글을 모두 인간(Human) 작성으로 간주하였다. 이는 당시 생성형 AI 서비스가 상용화되지 않았음을 전제로 한다.

2) AI 댓글 생성

AI 댓글 생성에는 Google Gemini-2.5-Flash와 OpenAI GPT-4를 사용하였다. 육안으로 AI임을 쉽게 구분 가능한 특징은 배제하기 위해 프롬프트 설계 시 다음과 같은 생성 조건을 적용하였다. 댓글 길이를 구간별로 비율을 설정해 생성하여, 길이만으로 AI 여부가 구분되는 것을 방지하였다. 마침표 사용 최소화, 맞춤법 오류 허용, 이모티콘 남발 금지, 생성형 AI가 강조를 위해 사

용하는 '*' 기호 사용 금지 등의 제한 사항을 프롬프트에 명시하였다. 또한, 이전 대화를 기억하는 것이 불가능한 API 호출의 약점을 해결하기 위해 이미 생성된 댓글 목록을 프롬프트에 포함하여 유사 표현의 재생성을 차단하였다. 기사당 100개 댓글을 생성하여 최종적으로 약 90,000개의 AI 댓글 데이터를 확보하였으며, 해당 데이터에는 라벨 1을 부여하였다.

표(2)는 본 논문에서 사용하는 프롬프트 사용 예시이다.

```
"""주어진 기사를 읽고 그에 대한 한국어 댓글을 조건과 요구사항에 맞게 작성해줘.
<기사 제목>
{title}
<기사 본문>
{text}
<기사 정보>
- 실제 댓글 작성자 통계
- 성별: 남자 {male}%, 여자 {female}%,
- 연령대: 10대 {age_10}%, 20대 {age_20}%, 30대 {age_30}%, 40대 {age_40}%, 50대 {age_50}%, 60대 이상 {age_60over}%
[댓글 생성 원칙]
- 총 {cnt}개 댓글 작성
- 댓글 길이는 최소 {min_len}, 최대 {max_len} 사이에서 다양하게
- 가장 중요: 2022년도의 인터넷 사용자처럼 작성할 것
- 말투, 성향, 의견, 길이 다양하게
- 마침표(.), '*' 기호는 최대한 지양
- 각 댓글은 줄바꿈으로 구분, 번호나 하이픈 등으로 시작하지 말 것
...(중략)...
- 혐오표현 허용, 맞춤법 오류도 허용
- 이미 생성된 댓글과 중복되는 구조, 표현, 단어 사용 금지
[이미 생성된 댓글 리스트]
{chr(10).join(dup_comments)}"""
```

표 2. 프롬프트 사용 예시

그러나 이러한 프롬프트 설계는 실제 환경의 다양성을 충분히 반영하지 못한다는 한계가 있다. 본 연구에서는 특정 제약을 부과한 단일 프롬프트를 사용해 댓글을 생성하였다. 이로 인해 AI 댓글 데이터가 특정 패턴에 편향되는 결과가 발생할 수 있다. 또한, 학습 모델의 일반화 능력을 제한하였을 가능성이 존재한다.

(3) 데이터 분석

그림(1)은 댓글 길이 분포 그래프이다. 제시된 전반적인 분포 형태는 인간 댓글과 AI 댓글이 유사하나, 인간 댓글에서 300자에 해당하는 댓글 수가 눈에 띄게 많다. 이는 네이버 뉴스 플랫폼의 댓글 글자 수 제한(최대 300자)으로 인해 해당 구간의 빈도가 높게 나타난 결과이다. 길이 정보만으로 로지스틱 회귀를 통해 라벨을 예측한 결과, Accuracy 55%로 단순 길이 정보만으로는 두 집단을 구분할 수 없음을 알 수 있다.

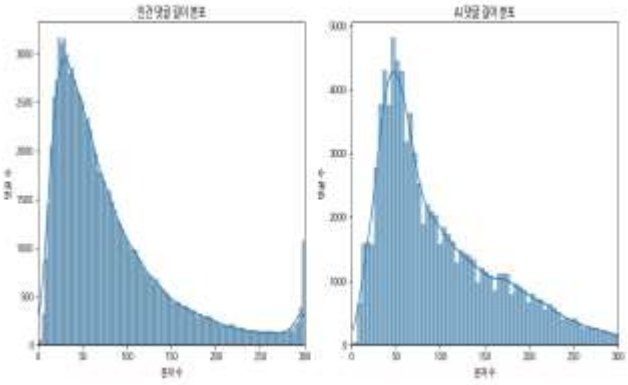


그림 1. 댓글 길이 분포 그래프

그림 (2)는 인간 댓글과 AI 댓글의 주요 단어를 시각화한 워드클라우드이다. 인간 댓글에서는 정치적 인물명과 강한 의견 표현이 크게 나타난 반면, AI 댓글은 비교적 일반적인 형용사와 부사, 완곡한 어휘의 빈도가 높게 나타났다. 이러한 차이는 두 집단의 어휘 선택 경향을 시각적으로 명확하게 보여준다.

그림 2. 워드클라우드

그림 (3)은 감정 분석 결과이다. 그림 (3)에서

정치 관련 뉴스 댓글의 특성상 부정적인 감정이 전체적으로 압도적인 비중을 차지하며, AI 댓글 또한 부정적인 감정을 중심으로 작성된 경우가 많았다. 프롬프트 설계에서 다양한 감정 톤을 요구했음에도 불구하고, 정치 이슈의 맥락이 부정적 표현을 유도했기 때문으로 해석된다.

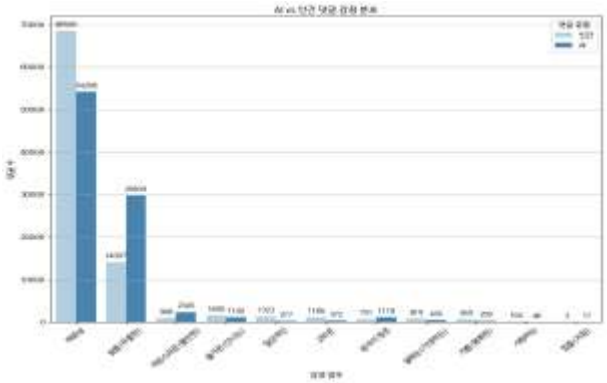


그림 3. 감정 분석 결과

표 (3)는 KoGPT2 언어모델을 사용하여 각 댓글의 Perplexity를 산출한 결과이다. Perplexity는 언어모델이 주어진 문장을 예측하는 데 필요한 불확실성을 수치화한 지표로, 값이 클수록 해당 텍스트는 모델이 '예측하기 어려운' 문장으로 인식된다.

Label	mean	std	Q1	Q2	Q3
Human	560.28	3672.14	138.98	238.27	454.58
AI	221.94	839.39	91.03	135.40	212.92

표 3. 라벨별 Perplexity 비교

그림 (4)는 Human 댓글과 AI 댓글의 Perplexity 분포를 비교한 결과이다. AI 댓글의 Perplexity는 평균 221.94(표준편차 839.39)로 Human 댓글의 평균 560.28(표준편차 3672.14)보다 현저히 낮게 나타났으며, 사분위 범위(Q1, Q3) 또한 Human 댓글(138.98, 454.58)에 비해 AI 댓글(91.03, 212.92)이 더 좁게 분포하였다. 이는 AI가 생성한 문장의 문장 구조와 어휘 선택이 비교적 일관적이며 규칙적인 패턴을 가진다는 것을 시사한다. 반대로 Human 댓글은 구어체, 맞춤법 오류, 비문, 비정형적 어휘 사용 등으로 모델이 예측하기 어려운 패턴을 더 많이 포함하고 있음을 예상할 수 있다.

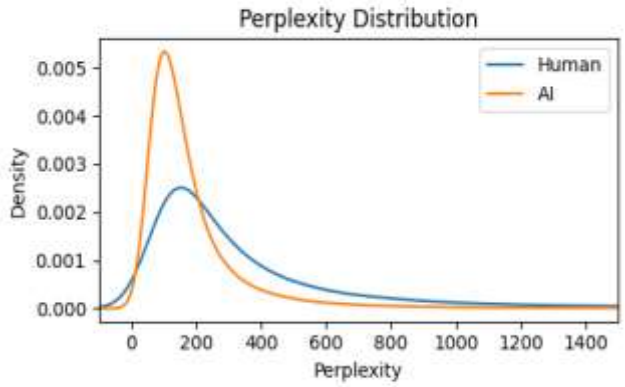


그림 4. Perplexity 분포 그래프

2. 제21대 대통령 선거 뉴스 기사 댓글

제21대 대통령 선거 뉴스 기사 댓글 데이터셋은 제20대 대선 데이터셋과 동일한 절차로 구축되었으나, AI 라벨이 없다는 점이 주요 차이점이다. 수집 기간은 후보자 등록 1일 전인 2025년 5월 9일부터 개표 2일 후인 2025년 6월 6일까지 총 30일로 설정하였다. 다만, 해당 기간은 국제정세(이스라엘-팔레스타인 분쟁, 미국 대선 등) 관련 보도가 빈번하여, 대선 관련 기사의 비중이 상대적으로 낮았다. 최종적으로 844개 기사가 선정되었으며, 각 기사에 대한 모든 댓글을 크롤링하여 총 427,237개의 댓글 데이터를 확보하였다

Ⅳ. 실험 결과 및 분석

본 연구는 2022년 제20대 대통령 선거 기간의 네이버 뉴스 댓글 데이터를 기반으로 AI 생성 댓글과 인간 작성 댓글을 분류하는 모델을 구축하고, 이를 2025년 제21대 대통령 선거 기간의 댓글 데이터에 적용하여 AI의 개입 가능성을 탐색하는 것을 목표로 한다.

1. 실험 환경 및 설정

(1) 실험 환경

실험은 Google Cloud Platform의 Vertex AI Colab Enterprise 환경에서 수행하였다. 런타임은 g2-standard-4 머신 타입과 NVIDIA L4 GPU를 활용하여 구축하였다.

(2) 분류 모델 선정

댓글과 같은 비정형 텍스트를 효과적으로 분류하기 위해 사전학습 언어모델(Pre-trained Language Models)을 중심으로 실험을 설계하였다. 구체적으로 한국어 자연어 처리(NLP)에 널리 활용되는 BERT, KoELECTRA, KcELECTRA를 주요 실험 모델로 채택하였다. 또한, 사전학습 모델과 비교하기 위해 기본적인 딥러닝 시퀀스 모델인 LSTM을 포함시켜, 사전학습 기반 모델의 성능 향상 폭을 함께 분석하였다.

(3) 베이스라인 성능 검증

초기 데이터셋(인간 댓글 9만 개, AI 생성 댓글 9만 개)을 활용하여 네 가지 분류 모델의 성능을 검증하였다. 그 결과는 표 (4)과 같다.

모델	Accuracy
LSTM	0.7237
BERT	0.9871
KoELECTRA	0.9844
KcELECTRA	0.9907

표 4. 모델 베이스라인 accuracy

실험 결과, LSTM은 약 72%의 정확도를 기록하며 상대적으로 낮은 성능을 보였다. 반면, BERT, KoELECTRA, KcELECTRA와 같은 사전학습 언어모델은 모두 98% 이상의 높은 정확도를 달성하였다. 특히 댓글 데이터에 특화된 KcELECTRA는 정확도 0.99로 가장 우수한 성능을 보였다.

그러나 이러한 높은 성능은 모델의 본질적 우수성이라기보다는 훈련 데이터셋의 편향에 기인했을 가능성이 크다. 즉, 모델이 인간 댓글과 AI 댓글 간의 의미적 차이를 학습했다기보다는, AI 댓글의 일관된 문체적 특성이나 인간 댓글의 비정형적 표현(오타, 신조어, 비문 등)과 같은 피상적인 차이를 구분하는 데 치중했을 가능성을 배제하기 어렵다.

(4) 실험 조건 및 하이퍼파라미터 설정

앞선 베이스라인 성능이 과도하게 높게 나타난 문제점을 보완하고, 실제 상황을 보다 현실적으로 반영하기 위해 실험 조건을 다음과 같이 조정하였다.

첫째, 인간 댓글과 AI 댓글 비율을 조정한다. 실제 여론 환경에서는 인간이 작성한 댓글이 압도적으로 많을 것을 가정하여, 인간 댓글과 AI 댓글이 9:1이 되도록 조정한다. 이때 AI 댓글(label=1)은 무작위로 1만 개를 추출하여 사용하였다.

둘째, 테스트 데이터 셋 비율은 0.3으로 조정한다. 이는 모델의 일반화 성능을 보다 신뢰성 있게 평가하기 위해 전체 데이터셋의 30%를 테스트 데이터 셋으로 설정하였다.

베이스라인 성능 검증 시 댓글 데이터의 길이와 모델의 기본 설정을 고려하여 입력 시퀀스 길이는 128, 학습률은 2e-5로 설정하였으며, 배치 사이즈는 64로 설정하였다. 배치 사이즈를 제외한 하이퍼파라미터를 동일하게 설정하였으며, 최종 실험에서 모델의 일반화 성능을 확보하기 위하여 설정한 주요 하이퍼파라미터는 다음 표 (5)와 같다.

epoch	max_length	batch_size	learning_rate
3	128	32	2e-5

표 5. 모델 학습 하이퍼파라미터 설정

먼저 epoch은 3으로 설정하여 불필요한 반복 학습으로 인한 과적합을 방지하고 모델의 안정성을 확보하고자 하였다. 입력 시퀀스의 길이를 의미하는 max_length는 128로 설정하였는데, 이는 댓글 텍스트의 평균 길이를 고려한 값이다. 또한 학습 과정에서의 연산 효율성과 메모리 사용량의 균형을 맞추기 위해 batch size는 32로 설정하였다. 마지막으로, 모델의 수렴 속도와 최종 성능에 중요한 영향을 미치는 learning rate는 2e-5로 지정하여 안정적 학습이 이루어지도록 하였다.

2. 2022년 데이터셋 기반 성능 분석

본 연구는 실제 온라인 환경을 모방하고자 의도적으로 데이터 불균형을 설정하였기에, 모델의 성능

평가는 여러 지표를 종합적으로 분석하는 것이 필수적이다.

특히, 전체 데이터 중 올바르게 예측한 비율을 나타내는 정확도(Accuracy)는 단독으로 신뢰하기 어렵다. 예를 들어, 인간 댓글이 90%를 차지하는 본 데이터셋에서 모델이 모든 입력을 '인간'으로만 판별해도 90%의 정확도를 기록하는 '정확도의 역설'이 발생할 수 있기 때문이다.

따라서 데이터 불균형 환경에서는 소수 클래스인 AI 댓글을 효과적으로 탐지하는 능력이 더욱 중요하다. 이를 평가하기 위해 정밀도(Precision)와 재현율(Recall)을 핵심 지표로 삼는다. 정밀도는 인간 댓글을 AI로 잘못 판단하는 '오탐' 비율을 억제하여 모델 예측의 신뢰성을 확보한다. 반면, 재현율은 실제 AI 댓글을 놓치지 않고 얼마나 잘 '색출'하는지를 측정하는 지표로, 탐지 시스템의 실효성과 직결된다. 이들의 조화 평균인 F1-score는 균형 잡힌 성능을 평가할 때 유용하며, 불균형 데이터셋에서 모델의 전반적인 성능을 비교할 때 효과적인 기준이 된다.

최종으로 조정된 실험 환경에서 BERT, KoELECTRA, KcELECTRA 모델을 학습시킨 후 성능을 평가한 결과는 다음 표(6)와 같다.

모델	Accuracy	Precision	Recall	F1-score	AUC
LSTM	0.7237	0.6477	0.9813	0.7803	-
BERT	0.9905	0.9722	0.9752	0.9737	0.9988
KoELECTRA	0.9915	0.9830	0.9694	0.9761	0.9989
KcELECTRA	0.9911	0.9657	0.9868	0.9759	0.9993

표 6. 모델 성능 평가 지표 비교

LSTM은 기본적인 딥러닝 기반 시퀀스 모델로서 비교 기준으로만 활용하였으며, 주요 분석은 사전학습 언어모델에 집중하였다. 데이터 불균형 조건에서도 세 가지 사전학습 모델은 모두 99% 이상의 높은 정확도를 유지하며, AI 생성 댓글 탐지에 효과적임을 확인할 수 있었다. 특히, 한국어 비정형 텍스트 분석에 적합한 KoELECTRA(0.9915)와 KcELECTRA(0.9911)가 BERT(0.9905)보다 소폭 우수한 성능을 보였다.

정밀도(Precision)와 재현율(Recall) 측면에서 각 모델의 특성이 구분되었다. KoELECTRA는 정밀도 0.9830으로 가장 높아, AI 댓글로 판별된 경우 실제로 AI일 가능성이 가장 높음을 보여주었다. 반면, KcELECTRA는 재현율 0.9868을 기록하여 실제 AI 댓글을 놓치지 않고 탐지하는 능력이 가장 뛰어남을 보였다.

계산 효율성을 평가하기 위해 각 모델의 학습 및 검증 소요 시간을 비교한 결과는 표 (7)과 같다. 3 epoch 기준 평균 계산 시간은 BERT가 754.08초로 가장 짧았으며, KoELECTRA(808.10초)와 KcELECTRA(815.47초)보다 효율적이었다. 즉, ELECTRA 계열 모델이 더 높은 성능을 제공하는 반면, 계산 효율성에서는 BERT가 우세함을 의미한다.

모델	계산시간(초)
BERT	754.08
KoELECTRA	808.10
KcELECTRA	815.47

표 7. 모델별 계산 시간

그림 (5)와 그림 (6)은 각각 KoELECTRA와 KcELECTRA 모델의 혼동행렬(confusion matrix) 결과를 나타낸 것이다.

그림 5. KoELECTRA 혼동행렬

먼저 KoELECTRA의 경우, True Negative(TN)는 26,920건으로 대부분의 인간 댓글을 정확히 식별하였으며, False Positive(FP)는 80건으로 매우 낮게 나타났다. 이는 KoELECTRA가 인간 댓글을 잘

못 AI로 분류하는 경우가 거의 없음을 의미한다. 그러나 False Negative(FN)는 175건으로, 일부 AI 댓글을 놓치는 경향이 관찰되었다. 이러한 결과는 KoELECTRA가 정밀도(Precision) 측면에서 강점을 가지지만, 재현율(Recall)에서는 상대적으로 한계를 보일 수 있음을 시사한다. 즉, KoELECTRA는 “AI로 판별된 경우 실제로 AI일 가능성”을 높이는 데 유리하지만, 일부 AI 댓글을 탐지하지 못할 가능성이 존재한다.

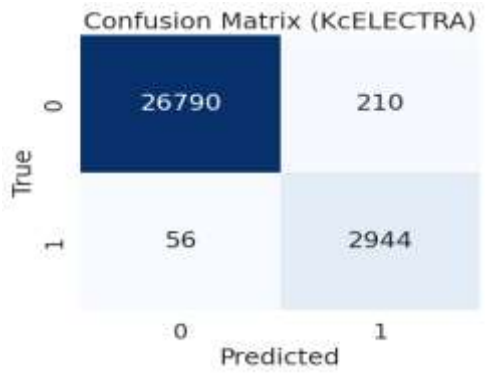


그림 6. KcELECTRA 혼동행렬

반면 KcELECTRA는 True Negative가 26,790건으로 KoELECTRA와 유사한 수준을 보였으나, False Positive가 210건으로 더 많았다. 이는 일부 인간 댓글을 AI로 잘못 분류하는 사례가 KoELECTRA보다 상대적으로 많음을 의미한다. 그러나 False Negative는 56건으로 현저히 낮아, 실제 AI 댓글을 놓치지 않고 탐지하는 능력이 뛰어남을 확인할 수 있었다. 따라서 KcELECTRA는 재현율(Recall) 측면에서 강점을 가지며, 실제 환경에서 AI 개입 가능성을 놓치지 않고 포착하는 데 유리하다.

종합하면, 두 모델은 상반된 장점을 가진다. KoELECTRA는 False Positive를 최소화하여 정밀도가 높다는 점에서 공정성과 신뢰성이 중요한 응용 맥락에 적합하다. 반대로 KcELECTRA는 False Negative를 최소화하여 재현율이 높으므로, 실제 AI 댓글을 최대한 탐지하는 것이 중요한 환경에서 효과적이다. 이러한 차별적 특성은 향후 활용 목적에 따라 모델 선택이 달라질 수 있음을 보여준다.

3. 2025년 데이터셋 분석 결과

제20대 대통령 선거에서 수집한 인간 댓글과 AI 댓글로 학습한 모델을 제21대 대통령 선거 기간의 뉴스 댓글 데이터셋에 적용하여, 실제 AI 댓글 개입 가능성을 탐색하였다.

KoELECTRA 모델을 적용한 결과, 전체 데이터셋에서 AI 댓글로 분류된 비율은 3.80%, 평균 예측 확률은 0.0477로 나타났다. 반면, KcELECTRA 모델을 적용했을 때는 AI 댓글로 분류된 비율이 1.46%, 평균 예측 확률은 0.0181로 산출되었다.

이 수치는 절대적으로 낮아 보일 수 있으나, 단순히 2025년 대선에서 AI 개입이 미미했다고 결론짓기는 어렵다. 그 이유는 훈련 데이터의 특성과 실제 환경 간의 차이에 있다. 본 연구에서 구축한 훈련 데이터는 Gemini와 GPT 모델을 활용해 단일 프롬프트 기반으로 생성되었기 때문에, 특정 패턴에 편향될 가능성이 크다. 반면, 2025년 실제 대선 기간에 사용된 AI 댓글은 다양한 모델과 정교한 프롬프팅 기법을 통해 훨씬 더 인간에 가까운 형태로 생성되었을 가능성이 높다. 이러한 차이로 인해 본 연구에서 탐지된 3.80%와 1.46%라는 값은 실제 개입 수준의 최소치로 해석되어야 한다.

AI가 생성한 것으로 분류된 댓글은 427,237개의의 댓글 중 1.46%(6,237개)~3.80%(16,235개)에 그쳤으며, 이 수치는 최소치임을 감안하더라도 다소 낮게 평가될 수 있다. 이러한 낮은 탐지율의 원인은 다음과 같이 두 가지 주요 관점에서 설명된다.

첫째, 분포 외(Out-of-Distribution, OOD) 데이터 탐지의 한계이다. 모델 학습에 사용된 20대 대선 데이터와 평가에 사용된 21대 대선 데이터는 3년의 시차가 존재한다. 이로 인해 테스트 데이터는 모델이 학습 과정에서 보지 못한 새로운 특성을 지닌 OOD 데이터가 된다. 특히 최신 연구는 모델이 이처럼 '처음 보는 도메인(Unseen Domain)'의 텍스트를 분류할 때 성능이 급격히 하락함을 입증하였다[23]. '21대 대선'이라는 특정 시점의 정치 상황은 변화된 주요 이슈, 인물, 어휘 등으로 인해 20대 대선과는 사실상 다른 도메인으로 간주되며, 이러한 현상이 모델의 일반화 성능을 저하하는 핵심 요인으로 작

용한다.

둘째, 인간과 AI가 생성한 텍스트 간의 언어적 특성 수렴 문제이다. LLM 기술이 고도로 발전함에 따라, AI가 생성한 텍스트와 인간이 작성한 텍스트 사이의 통계적·언어적 차이가 점차 사라지고 있다[23]. 이러한 현상은 특히 짧고 정형화된 표현이 반복적으로 나타나는 온라인 뉴스 댓글 환경에서 더욱 두드러진다. 즉, 인간이 작성한 댓글의 패턴이 오히려 AI 생성 텍스트와 유사해지는 경향이 나타나면서, 언어적 특징에 기반한 분류 모델의 변별력을 약화시킨다.

이러한 점을 종합할 때, 본 연구에서 관찰된 낮은 탐지율은 향후 AI 댓글 탐지 시스템이 변화하는 언어 환경과 LLM 기술에 적응하기 위해 지속적인 재학습과 일반화 성능 강화가 필수적임을 시사한다.

탐지된 비율이 낮더라도 이는 곧 AI 기반 여론 개입의 유의미한 신호가 존재했음을 의미한다. 따라서 본 연구의 결과는 2025년 대선에서 AI의 실제 영향력이 탐지된 수치보다 더 클 수 있음을 시사한다.

결론적으로, 이번 분석은 생성형 AI가 국내 선거라는 민감한 정치적 맥락 속에서 실제로 활용되었을 가능성을 실증적으로 보여준다. 향후 공정한 여론 환경을 보장하기 위해서는 끊임없이 진화하는 AI 기술에 대응할 수 있는 고도화된 탐지 모델의 개발과 지속적인 모니터링 체계 구축이 필수적임을 강조한다.

V. 결론 및 향후 연구

본 연구는 온라인 정치 뉴스 댓글이라는 민감한 여론 공간에서 생성형 인공지능의 개입 가능성을 탐지하고자 하였다. 2022년 제20대 대통령 선거 기간에 수집된 네이버 뉴스 댓글을 기반으로 KoELECTRA 모델을 학습시킨 뒤, 이를 2025년 제21대 대통령 선거 기간의 실제 댓글 데이터에 적용하였다. 실험 결과, AI로 분류된 댓글의 비율은 1.46%~3.80% 수준으로 나타났다. 이는 절대적인 수치만 보았을 때 낮은 편일 수 있으나, 해당 시기에 생성형 AI가 전혀 개입하지 않았다고 단정할 수 없는 수준으로 해석된다. 따라서 본 연구는 국내 대선이라는 정치적 맥락 속에서 AI 기반 여론 개입의 가능성을 실증적으로 확인했다는 점에서 학술적 의의를 가진다. 그러나 본 연구에는 몇 가지 한계가 존재한다. 첫째, 2022년 대선 데이터셋에서 사용된 AI 댓글은 Gemini와 GPT 모델을 통해 단일 프롬프트로 생성된 것이며, 이로 인해 학습 데이터가 특정 패턴에 편향되었을 가능성이 있다. 반면, 2025년 실제 환경에서는 다양한 AI 모델과 정교한 프롬프팅 기법이 활용되었을 가능성이 높아, 탐지율이 실제보다 낮게 측정되었을 수 있다. 따라서 본 연구에서 제시한 1.46%~3.80%라는 수치는 최소치로 해석되어야 한다. 마지막으로 본 연구의 훈련 데이터가 실제 댓글 환경을 충분히 대표하지 못할 수 있다는 점이다. 시간적 격차 속에서 생성형 AI 기술이 빠르게 발전했기 때문에 학습 데이터가 최신 패턴을 반영하는 데 제약이 있으며 동시에 AI 댓글은 단일 프롬프트로 생성한 방식 역시 실제 악의적 활용에서 나타날 수 있는 다양한 변형과 조작을 충분히 담아내지 못한다. 이러한 한계로 인해 탐지 결과는 실제 상황보다 보수적으로 산출되었을 가능성이 크며 다양한 데이터 생성 방식과 시뮬레이션을 통한 검증이 요구된다.

향후 연구에서는 다음과 같은 보완이 필요하다. 첫째, 최신 데이터를 지속적으로 수집하고 더 다양한 AI 모델과 프롬프트 기법을 활용하여 학습용 데이터셋의 다양성과 현실성을 높여야 한다. 둘째, 댓글 텍스트뿐만 아니라 작성 시점, 작성 계정의 활동 패턴 등 메타데이터를 통합적으로 활용하는 다차원 탐지 모델을 구축할 필요가 있다. 셋째, AI 댓글 탐지의 성능을 지속적으로 고도화하기 위해 반(反)탐지형 AI 모델과의 경쟁적 학습(Adversarial Training) 접근을 도입할 수 있다. 마지막으로, 기술적 차원을 넘어 사회·정치적 대응 체계를 마련하고, 공정한 여론 환경을 보장하기 위한 제도적 장치에 대한 연구도 병행되어야 한다.

결론적으로, 비록 탐지된 비율은 낮게 나타났으나 생성형 AI 기반 여론 개입 시도가 실제로 존재했다는 사실은 민주주의 근간인 공정한 선거 과정에 중

대한 위협임을 보여준다. 앞으로 더욱 정교해질 AI 기술의 잠재적 악용 가능성을 고려할 때, 학계와 사회 전반에서 이에 대한 지속적인 경각심과 대응이 필수적이다.

REFERENCES

[1] 한국리서치, "[기획] 뉴스 기사 댓글에 대한 인식," https://hrcopinion.co.kr/en/archives/20815, 2022 (accessed Aug., 19, 2025).

[2] 오픈AI, "러시아·중국 등 챗GPT로 인터넷 여론 조작, 활동 차단해," 한국무역협회 뉴스, https://www.kita.net/board/totalTradeNews/totalTradeNewsDetail.do?no=84028&siteId=2, 2024 (accessed Aug., 19, 2025).

[3] 글로벌이코노믹, "대선 여론 30%는 가짜 계정…AI 딥페이크·中 개입설까지 번졌다," https://m.g-enews.com/article/Global-Biz/2025/06/20250630110353 6846fbbec65dfb_1, 2025 (accessed Aug., 19, 2025).

[4] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending Against Neural Fake News," *Advances in Neural Information Processing Systems (NeurIPS 2019)*, vol. 32, pp. 9051 - 9062, Vancouver, Canada, Dec. 2019.

[5] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, "Automatic Detection of Generated Text is Easiest when Humans are Fooled," *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pp. 1808 - 1822, Seattle, USA (Online), Jul. 2020.

[6] W. Go, H. Kim, A. Oh, and Y. Kim, "XDAC: XAI-Driven Detection and Attribution of LLM-Generated News Comments in Korean," *Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pp. 22728 - 22750, Vienna, Austria, Jul. 2025.

[7] 네이트뉴스, "[샷!] '챗GPT 안 썼다. 억울하다'…AI 역설," https://news.nate.com/view/20250407n01805, 2025 (accessed Aug., 19, 2025).

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 33, pp. 1877 - 1901, 2020.

[9] OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, 2023 (accessed Aug. 19, 2025).

[10] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.S. Huang, et al., "Ethical and Social Risks of Large Language Models," arXiv:2112.04359, 2021.

[11] 사이브레아, "선거 관련 SNS 계정 분석 보고서," 내부 보고서, 2024년

[12] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735 - 1780, Nov. 1997.

[13] C. Chang and M. Masterson, "Using Word Order in Political Text Classification with Long Short-term Memory Models," *Political Analysis*, vol. 28, no. 3, pp. 395 - 411, Jul. 2020.

[14] KLUE-BERT, https://huggingface.co/klue/bert-base, 2021 (accessed Aug., 13, 2025).

[15] S. Park, J. Park, Y. Kim, J. Lee, H. Oh, J. Lim, et al., "KLUE: Korean Language Understanding Evaluation," arXiv:2105.09680v4, pp. 1 - 76, Nov. 2021.

[16] Korean Hate Speech Dataset, https://github.com/kocohub/korean-hate-speech, 2020 (accessed Aug., 10, 2025).

[17] J. Lee, S. Park, Y. Kim, J. Lee, H. Kim, and J. Lee, "K-MHaS: A Multi-label Hate Speech Detection Dataset in Korean Online News Comment," arXiv:2208.10684v3, pp. 1 - 9, Sep. 2022.

[18] S. Kim, B. Lee, M. Muazzam, J. Moon, and S. Rho, "Deep Learning-Based Natural Language Processing Model and Optical Character Recognition for Detection of Online Grooming on Social Networking Services," *Computer Modeling in Engineering & Sciences,* vol. 143, no. 2, pp. 2079 - 2108, May 2025.

[19] 이예솔, "화장품 리뷰를 활용한 사전학습모델 성능 비교 연구: BERT, RoBERTa, ELECTRA를 중심으로," *한양대학교 석사학위논문,* 2024년

[20] 엄기홍, 김대식, "온라인 정치 여론 분석을 위한 댓글 분류기의 개발과 적용: KoBERT를 활용한 여론 분석," *한국정당학회보,* 제20권, 제3호, 167 - 191쪽, 2021년

[21] 임인재, 박윤정, 이세영, 금희조, "제20대 대통령 선거 기사에 대한 댓글 분석: 반시민적 표현을 중심으로," *한국언론정보학보,* 통권 제120호, 147 - 188쪽, 2023년

[22] Y. Li, Q. Li, L. Cui, W. Bi, Z. Wang, L. Wang, L. Yang, S. Shi, and Y. Zhang, "MAGE: Machine-generated Text Detection in the Wild," *Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pp. 36 - 53, Bangkok, Thailand, Jul. 2024.

저 자 소 개

김신형(준회원)

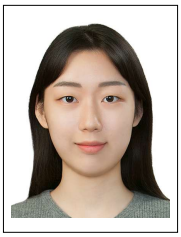
2023년~전남대학교 빅데이터융합학과 학사 재학

<주관심분야 : 자연어처리, 생성형 AI, 빅데이터>

유예원(준회원)

2023년~전남대학교 빅데이터융합학과 학사 재학

<주관심분야 : 자연어처리, 생성형 AI, LLM, 데이터 분석>



박지은(준회원)

2022년~전남대학교 빅데이터융합학과 학사 재학

<주관심분야: 자연어처리, 생성형 AI 연구, 빅데이터 분석>

노상희(준회원)

2022년~ 전남대학교 자율전공학부 학사 재학

<주관심분야 : 빅데이터 분석, 자연어처리, 언론>

손남례(정회원)

2005년 전남대 전산학과 박사 졸업.

2011년~2017년 한국전자통신연구원 연구원

2017년~2021년 호남대 정보통신공학과 교수

2021년~현재 전남대 소프트웨어중심대학사업단 교수

<주관심분야:빅데이터분석솔루션, 전력IT, 딥러닝 등>