ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2025 December: 31(4): 193~216 http://dx.doi.org/10.13088/jiis.2025.31.4.193

생성형 AI 기반 한국어 가짜뉴스 데이터셋 구축 및 탐지 모델 비교 분석 연구*

이상민
국민대학교 경영정보학부
(luke2176@kookmin.ac.kr)

이유진
국민대학교 경영정보학부
(iamlyj03@kookmin.ac.kr)

장우진
국민대학교 경영정보학부
(woojin9n@kookmin.ac.kr)

안현철
국민대학교 경영정보학부
(hcahn@kookmin.ac.kr)

생성형 인공지능(Generative AI)과 대규모 언어 모델(LLM)의 발전으로 인간이 작성한 기사와 거의 구분이 어려운 수준의 텍스트가 대량 생산되면서, 특히 정치 영역에서 선거 후보자나 특정 집단의 이미지를 의도적으로 훼손하는 가짜뉴스가 한국 사회의 정보 신뢰성을 심각하게 위협하고 있다. 가짜뉴스는 의도성 여부에 따라 허위조작정보(disinformation)와 오정보(misinformation)로 구분되지만, 본 연구는 이 중에서도 사회적 파급력이 큰 허위조작정보, 즉 의도적 조작 유형에 집중한다. 지금까지의 가짜뉴스 탐지 연구는 정교하게 설계된 허위조작정보 앞에서 텍스트 내용만으로 진위를 판별하기 어렵고, 소셜 맥락과 확산 양상 등 외부 맥락도 충분히 활용하지 못한다는 한계가 지적되어 왔다. 특히 한국어 환경에서는 영어권에 비해 검증된 데이터셋의 규모와 다양성이 크게 뒤처져 있고, 다수의 LLM이 영어 중심 말뭉치로 사전학습됨에 따라 한국어 텍스트에서는 성능 저하와 언어적 편향, 그리고 실제 한국어 뉴스 데이터셋을 활용한 LLM 기반 탐지 연구에서 보고된 논리 구조나 문법성 같은 콘텐츠 품질에 대한 편향이 확인된다. 이러한 연구 환경을 보완하기 위해 본 연구는 한국어 정치 뉴스 데이터를 선별하여 새로운 데이터셋을 구축한다. 구체적으로, 생성형 AI를 활용하여 기존 뉴스의 핵심 사실(kernel of truth)을 유지한 채 해석과 맥락을 왜곡하는 방식의 한국어 정치 가짜뉴스 텍스트를 생성한다. 이로써 진짜 뉴스, 가짜 뉴스 쌍으로 구성된 데이터셋을 완성한다. 나아가 이 데이터셋을 토대로 한국어 특화 텍스트 임베딩 및 사전학습 언어모델 기반 탐지 모델의 가능성과 한계를 실증적으로 분석한다. 궁극적으로는 생성형 AI 시대에 고도화되는 한국어 정치 가짜뉴스에 대응할 수 있는 한국어 특화 탐지 프레임워크의 필요성과 방향성을 제시하고자 한다.

주제어 : 가짜뉴스 탐지, 생성형 AI, 한국어 가짜뉴스, 한국어 사전학습 언어모델, 텍스트 임베딩

논문접수일 : 2025년 11월 25일 논문수정일 : 2025년 12월 14일 게재확정일 : 2025년 12월 16일
원고유형 : Regular Track 교신저자 : 안현철

1. 서론

최근에는 ChatGPT와 같은 대규모 언어 모델(Large Language Model, LLM)을 포함한 생성형 인공지능(Generative AI) 기술의 발전이 텍스트 생성 능력의 혁신을 가져오면서, 가짜 뉴스 문제에 새로운 차원의 위협을 가하고 있다(김혜윤 등, 2025; Loth et al., 2024). 생성형 AI는 방대한 데이터를 학습하여 인간 수준의 언어 이해 및 생성 능력을 보여주며, 이를 통해 정교한 허위 정보나 가짜 구매 후기/댓글 등을 대량으로 생성할 수 있게 되었다(고상훈, 안현철, 2024; 박준성, 2024). 과거에는 대부분 수동으로 작성되었던 허위 정보와 달리, 생성형 AI가 만들어낸 고도화된

* 이 논문은 2022년 대한민국 교육부와 한국연구재단의 인문사회분야 중견연구자지원사업의 지원을 받아 수행된 연구임 (NRF-2022S1A5A2A01048638).

텍스트는 일반인이 진위 여부를 식별하기 매우 어려워, 가짜 뉴스의 생성 및 확산 위험성을 극도로 높이고 있다(강주영, 송민, 2024; 정원영, 2024). 따라서 이러한 시대적 변화 속에서 사회적 신뢰를 보호하고 정보의 진실성을 유지하기 위해, 급증하고 고도화되는 한국어 가짜 뉴스를 신속하고 정확하게 탐지할 수 있는 자동화된 기술 개발은 매우 시급하고 중요한 연구 과제라 할 수 있다.

가짜 뉴스 문제를 해결하기 위해 학계에서는 전문가 기반 팩트 체크(Fact Check) 서비스부터 인공지능 기반 탐지 방법론까지 다양한 노력이 이루어져 왔다(좌희정 등, 2019). 초기 연구들은 주로 기사 내용(콘텐츠)만을 활용하여 텍스트 분석을 수행했으며, 방법론적으로는 단어의 출현 빈도에 기반한 TF-IDF(Term Frequency-Inverse Document Frequency)나 단어를 벡터로 변환하는 워드 임베딩(Word Embedding) 기법(Word2Vec, Doc2Vec 등)이 사용되었고, 이를 SVM(Support Vector Machine)이나 신경망(Neural Network) 같은 기계학습 분류기와 결합되어 탐지 모델을 구축했다(김지혁, 안현철, 2023; 심재승 등, 2020; 이충열, 2018; 현준서 등, 2024). 이후 딥러닝(Deep Learning) 시대가 도래하며 LSTM(Long Short-Term Memory)이나 CNN(Convolutional Neural Network) 같은 순환 신경망 모델을 이용한 연구가 진행되었으며, 특히 문맥을 양방향으로 참조하여 깊이 이해하는 BERT(Bidirectional Encoder Representations from Transformers)와 같은 사전 학습 언어 모델(PLM)의 등장으로 텍스트 이해 성능이 크게 향상되었다(이태원 등, 2023; 임동훈 등, 2021). 국내에서도 한국어의 고유한 언어적 특성(예: 교착어적 특성, 문법 구조의 특수성)을 효과적으로 처리하기 위해 KoBERT, KcBERT, 그리고 KLUE-BERT와 같은 한국어 특화 모델들이 개발 및 활용되었다(강예나 등, 2025; 김혜윤 등, 2025; 윤호영, 안도현, 2023; 정원영, 2024; 한윤진, 김근형, 2021; Park et al., 2021).

그러나 이러한 콘텐츠 기반의 탐지 연구들은 가짜 뉴스의 조작 정밀도가 높아졌을 때 탐지가 어려워지는 근본적인 한계를 내포한다(한윤진, 김근형, 2021). 이에 따라 많은 연구자들이 기사 내용 외에 외부 정보나 맥락(Context)을 활용하여 정확도를 개선하는 방법을 모색했다. 예를 들어, 소셜 미디어에서 뉴스 확산 패턴의 구조적 정보를 활용하거나, 구글 검색 결과를 벡터화하여 외부 검증 자료로 활용하는 Link2vec 기법을 적용하는 연구 등이 제안되었다(정이태, 안현철, 2022; Shim et al., 2021). 하지만 최근 LLM을 가짜 뉴스 탐지에 직접 활용하려는 시도들은 새로운 도전과 한계를 동시에 보여준다. 대다수의 LLM은 영어권 웹사이트, 뉴스 등을 주 학습 데이터로 사용하여 영어 콘텐츠에 편중되어 학습되었기 때문에, 한국어와 같은 비영어권 언어 환경에서 모델 성능을 평가하거나 검증하는 데 어려움이 발생하며, 이는 한국어 기반 LLM 탐지 연구에 중요한 연구 공백을 형성하고 있다. 실제로 한국어 뉴스 데이터셋(SNU 팩트체크 기반)을 활용한 LLM 기반 탐지 연구 결과는 영어 중심 연구 대비 낮은 59.8%의 정확도를 기록했으며, 또한 LLM이 텍스트의 진실성 여부보다는 논리적 구조나 문법적 오류와 같은 콘텐츠 품질에 민감하게 반응하여, 잘 구성된 기사에 편향될 수 있다는 문제점도 지적되었다. 게다가 국내에서는 LLM 학습에 필요한 한국어 공개 데이터셋이 제한적이고, 생성형 AI가 만들어내는 고도화된 텍스트 자체를 탐지하는 연구 역시 여전히 부족하다(고상훈, 안현철, 2024; 김지혁, 안현철, 2023; 이충열, 2018; 정원영, 2024; 강주영, 송민, 2024; Park et al., 2021).

따라서 본 연구는 생성형 AI의 등장으로 더욱 정교화되어 탐지가 어려워진 한국어 가짜뉴스에 효과적으로 대응하기 위한 생성형 AI 기반 한국어 가짜뉴스 탐지 모델 및 프레임워크를 제안하는 것을 목적으로 하며, 기존 연구가 가진 콘텐츠 기반 분석의 한계와 한국어 환경에서 나타나는 LLM 탐지 성능 저하 및 콘텐츠 품질 편향 문제를 극복하고자 한다. 생성형 AI 시대에 적합한 한국어 가짜뉴스 탐지 연구를 수행하기 위해, 먼저 AI Hub 뉴스 데이터 중 정치 카테고리 기사만을 선별·샘플링을 진행하였다. 그 다음, GPT-4o를 활용하여 원본 기사를 기반으로 정치 분야의 진짜 뉴스와 정치부 기자, 극성 정치 유튜버 등 다양한 페르소나를 반영한 가짜뉴스 텍스트를 생성했다. 이를 정리하여 한국어 가짜뉴스 데이터셋을 직접 구축 및 공개했다. 또한 해당 데이터셋을 대상으로 Mecab+Doc2Vec, KoBERT, KLUE-BERT 등 한국어 특화 임베딩 및 분류 모델을 비교·검증을 진행했다. 이를 통해 각 기법의 성능과 시너지 효과를 실증적으로 분석하였다. 나아가 생성형 AI의 강력한 언어 이해 및 추론 능력을 활용하면서도 LLM이 기존 모델이 포착하지 못했던 한국어 가짜뉴스의 언어적·맥락적 특성을 효과적으로 학습할 수 있는 새로운 LLM 기반 탐지 체계를 구축했다. 또한 실증 분석 결과를 토대로 한국어 가짜뉴스 탐지 환경에서 최적의 성능을 보이는 모델 조합을 제시함으로써, 그 성능과 실효성 및 범용적 적용 가능성을 검증하고자 한다. 본 연구는 한국어 환경에서 LLM 활용 연구의 확장에 기여하고, 고도화되는 허위 정보에 대한 체계적이고 실효적인 대응 방안을 제시함으로써, 정보 신뢰성 강화를 위한 학문적 및 실무적 시사점을 제공할 것이다.

본 논문의 구성은 다음과 같다. 제 2장 이론적 배경에서는 가짜뉴스 탐지 전반에 대한 이론과 함께, 한국어 환경에서의 가짜뉴스 탐지 연구를 정리하고, 텍스트 임베딩 기법의 개념과 특징을 소개한 뒤, 선행연구 검토를 통해 본 연구가 다루고자 하는 연구 공백을 도출한다. 제 3장 연구 모델에서는 한국어 가짜뉴스 탐지를 위한 데이터셋 구축 과정과 제안하는 가짜뉴스 분류 모델의 구조 및 구현 방법을 구체적으로 제시한다. 제 4장 실험결과에서는 실험에 활용된 데이터의 특성과 실험 설계 방법을 설명하고, 실험을 통해 도출된 결과를 제시하며 제안 모델의 성능을 분석한다. 마지막으로 제 5장 결론에서는 연구 결과를 종합하여 정리하고, 본 연구의 학술적·실무적 시사점과 한계, 그리고 향후 연구 방향을 제안한다.

2. 이론적 배경

2.1. 가짜뉴스의 정의 및 유형

가짜뉴스는 내용의 허위성과 생산·유포 과정에서의 의도성을 기준으로 허위조작정보(Disinformation)와 오정보(Misinformation)로 구분된다. 허위조작정보는 특정한 경제적·정치적 이익을 목적으로 고의적으로 잘못된 정보를 조작·유포하여 대중을 오도하거나 혼란을 조장하는 행위를 의미하며, 정보 설계 단계에서부터 체계적인 기만 의도(deceptive intent)가 내재되어 있다는 점이 핵심 특성으로 제시된다(좌희정 등, 2019). 반면 오정보는 정보 제공자가 내용이 허위임을 인지하지 못한 상태에서 비의도적으로 잘못된 정보를 전달하는 경우를 의미하며, 적극적인 기만 의도가 수반되지 않은 실수·오류에 기반한 정보 전달의 사례에 해당한다(고인석, 2025).

가짜뉴스 탐지 연구는 정보의 허위성, 기만 의도, 오도 효과(misleading effect) 등 복수의 차원을 고려하여 발전해 왔으며, 자동 탐지 방법론은 크게 세 가지 축으로 구조화된다(고인석, 2025). 이는 뉴스 본문과 제목 등 텍스트 특성을 활용하는 내용 기반 기법(Content-based methods), 소셜 네트워크 상의 외부 맥락 정보를 활용하는 맥락 기반 기법(Context-based methods), 그리고 두 접근을 결합한 혼합(Hybrid) 기법이다(이태원 등, 2023). 내용 기반 탐지는 구현이 단순하나, 뉴스의 조작 정밀도가 높아질수록 진위 여부를 식별하기 어렵다는 한계를 드러내고 있으며, 이를 보완하기 위해 혼합형 탐지 모델이 대안으로 제시되고 있다(박성수, 이건창, 2019; 정이태, 안현철, 2022; 정동인, 2023; 한윤진, 김근형, 2021; 현윤진, 김남규, 2018).

본 연구는 생성형 인공지능이 악의적인 목적으로 정교하게 생성할 수 있는 가짜뉴스 중, 특히 선거 후보자의 이미지를 실추시키는 등 정치적 목적을 가지고 양산되는 허위조작정보 유형의 가짜뉴스 탐지에 초점을 맞춘다.

2.2. 가짜뉴스 탐지 연구 동향

2.2.1. 기존 탐지 기술의 발전과 한계

콘텐츠 기반 탐지는 뉴스 콘텐츠 자체에서 언어적 특징(어휘 자질, 구문론적 자질)을 추출하여 가짜뉴스 여부를 분류하는 방법이다(박성수, 이건창, 2019; 이태원 등, 2023). 초기 한국어 가짜뉴스 탐지 연구에서는 단어의 빈도수에 기반하여 문서 특성을 파악하는 TF-IDF와 같은 희소표현(Sparse Representation) 방식이 주로 활용되었으나, 이는 문장의 문맥 정보를 반영할 수 없다는 한계를 가졌다(한윤진, 김근형, 2021).

이러한 한계를 극복하기 위해 소셜 컨텍스트 기반 탐지가 등장했는데, 이는 뉴스 콘텐츠 정보뿐만 아니라 소셜 미디어에서의 사용자 행동, 출처 정보, 확산 네트워크 구조 등을 통합하여 활용하는 방식이다(김지혁, 안현철, 2023). 소셜 미디어에서 정보가 워낙 빠르게 전파되는 특성으로 인해, 가짜 뉴스 탐지의 필요성은 매우 높고 중대한 현안으로 인식된다.

또한, 내용 기반 탐지에서는 뉴스의 내용 외에 감성 변화 패턴을 추가적인 통찰력으로 활용하는 방법이 제안되기도. 이는 뉴스를 긍정/부정으로 단순 분류하는 대신 문장 단위에 따른 감성 변화를 분석하여, 허위 정보가 독자의 신뢰를 훼손하기 위해 과도한 불안감을 부추기는 특성을 탐지에 활용한다(이태원 등, 2023).

2.2.2. LLM의 등장과 허위정보의 고도화

최근 대규모 언어 모델(LLM)과 같은 생성형 인공지능 기술이 발전하면서, 인간이 작성한 글과 인공지능이 생성한 글의 구분이 더욱 어려워졌다. LLM은 방대한 양의 언어를 학습하여 텍스트를 생성하며, 정교하게 생성된 가짜 댓글이나 가짜 리뷰는 향후 더욱 발전할 것으로 예상된다(정원영, 2024).

AI가 생성한 텍스트는 매우 자연스러워 인간이 작성한 텍스트와 구별하기 어렵다는 점이 다수의 설문 조사에서 확인되었으며, 이러한 특성 때문에 기존 탐지 모델은 정교하게 생성된 AI 기반 콘텐츠를 효과적으로 식별하는 데 한계를 보이고 있다. 특히 기존 연구에서 활용된 인간 고유의 언어적 특성이나 직관적 판단을 기반으로 구축된 모델은 이러한 고도화된 허위 정보에 충분히 대응하지 못하는 것으로 나타났다(강주영, 송민, 2024; Park, 2024).

2.3. 텍스트 임베딩 기법과 분류모델

일반적으로 가짜 뉴스 탐지는 참/거짓을 판별하는 이진 분류(Binary Classification) 문제로 다루어지며, 성능 향상을 위해 다양한 기계 학습 및 딥러닝 기법들이 활용되고 있다.

2.3.1. 텍스트 기반 탐지 기법

가짜뉴스 탐지 연구에서 가장 기본적으로 사용되는 접근은 텍스트 분석 기반 방법이다. 이 접근은 뉴스 기사를 구성하는 문자, 단어, 문장, 문단 등 다양한 수준의 언어적 요소에서 특징을 추출하여, 해당 문서가 가짜뉴스인지 여부를 분류하는 데 활용된다(박성수, 이건창, 2019). 초기 연구에서는 주로 TF-IDF(Term Frequency – Inverse Document Frequency)를 사용하여 문서의 특성을 벡터화하는 방식이 적용되었으나, TF-IDF는 개별 단어의 출현 빈도에 기반한 희소 표현을 사용하기 때문에 문맥 정보를 충분히 반영하지 못한다는 구조적 한계를 가진다(심재승 등, 2020; 한윤진, 김근형, 2021).

이러한 한계를 보완하기 위해 문맥 정보를 보다 효과적으로 포착할 수 있는 밀집 표현 기반 워드 임베딩(Word Embedding) 기법이 제안되었다. Word2Vec은 유사한 문맥에서 사용되는 단어는 의미적으로도 유사하다는 분포 가설(Distributional Hypothesis)에 기반하여 단어를 연속적 벡터 공간에 매핑하는 방식이며, 텍스트의 의미적 구조를 효과적으로 반영할 수 있다는 점에서 널리 활용되었다(심재승 등, 2020). Word2Vec은 중심 단어를 주변 단어로부터 예측하는 CBOW(Continuous Bag of Words) 방식과, 중심 단어로부터 주변 단어를 예측하는 Skip-gram 방식 두 가지로 학습할 수 있으며, 일반적으로 Skip-gram 방식이 희귀 단어 표현 및 의미 학습에서 더 우수한 성능을 보이는 것으로 알려져 있다(김지혁, 안현철, 2023; 임동훈 등, 2021).

이후 Le와 Mikolov(2014)는 Word2Vec을 확장하여 문단 또는 문서 전체를 단일 벡터로 표현할 수 있는 Doc2Vec을 제안하였다. Doc2Vec은 문서 내 단어들의 분포적 특성과 문서의 고유한 주제 정보를 반영해 문장·문단·문서 단위의 임베딩을 학습하는 방식으로, PV-DM(Distributed Memory)과 PV-DBOW(Distributed Bag of Words)라는 두 가지 학습 구조를 갖는다. PV-DM은 주변 단어와 함께 문서 ID를 입력으로 사용하여 다음 단어를 예측하는 방식으로 문맥의 흐름과 문서 주제를 반영할 수 있으며, PV-DBOW는 문서 ID만을 입력으로 사용하여 단어 예측을 수행하므로 단어 순서와 무관하게 전체 문서의 의미적 정보를 요약하는 데 유리하다(김지혁, 안현철, 2023).

가짜뉴스 탐지 분야에서도 Doc2Vec 기반 표현은 텍스트 내용 기반 탐지에 활용되어 왔다. 예를 들어, Doc2Vec으로 문서를 벡터화한 뒤 문서 수준에서 핵심 정보를 요약하여 분류에 활용하는 모델이 제안된 바 있으며, Doc2Vec 임베딩에 품사 태그 수나 전체 단어 수 등 기사 구조를 반영하는 변수를 추가하여 탐지 성능을 개선하려는 연구도 수행되었다(김지혁, 안현철, 2023; 정이태, 안현철, 2022).

이와 같은 텍스트 표현 방식을 입력으로 하여 적용된 분류 모델은 매우 다양하다. 전통적인 기계학습 기반 분류기로는 서포트 벡터 머신(Support Vector Machine, SVM)과 로지스틱 회귀(Logistic Regression)가 널리 사용되었으며, 딥러닝 접근에서는 LSTM(Long Short-Term Memory), Bi-LSTM (Bidirectional LSTM, 그리고 CNN(Convolutional Neural Network) 등이 적용되어 텍스트 기반 탐지의 성능을 향상시키고자 하였다(이소현, 2017; 이충열,

2018; 이태원 등, 2023; 임동훈 등, 2021; 현준서 등, 2024).

2.3.2. LLM 기반 임베딩과 한국어 특화 모델

최근 텍스트 분류 연구는 대규모 언어 모델(LLM)을 활용한 임베딩 기법을 중심으로 전개되고 있다. BERT(Bi-directional Encoder Representations from Transformers)는 양방향 Transformer 인코더 구조를 기반으로 문장 내 단어가 앞·뒤 문맥과 맺는 관계를 동시에 학습하는 언어 모델로, 자연어 이해 작업 전반에서 높은 성능을 보이고 있다. BERT는 마스크드 언어 모델(Masked Language Model, MLM)과 다음 문장 예측(Next Sentence Prediction, NSP)이라는 두 가지 비지도 학습 과제를 통해 사전 학습이 이루어진다(강병희, 신승협, 2024).

특히 한국어는 교착어적 특성과 복잡한 형태소 구조를 가지므로, 한국어의 언어적 특성을 효과적으로 반영한 특화 언어 모델의 필요성이 제기되어 왔다(조예현 등, 2025). KoBERT, KLUE-BERT 등 한국어에 특화된 언어 모델은 형태소 단위 토크나이징과 한국어 말뭉치 기반 사전 학습을 통해, 기존 다국어 모델 대비 한국어 자연어 처리 작업에서 더 우수한 성능을 보이는 것으로 보고된다(정원영, 2024).

텍스트 분류 작업에서는 이와 같은 임베딩을 입력으로 활용하는 지도학습(supervised learning) 기반 기계학습 기법이 널리 사용된다. 이는 라벨이 부여된 학습 데이터로부터 분류 모델을 구축하고, 학습된 모델을 새로운 텍스트 데이터에 적용하여 범주를 예측하는 방식이다. 이러한 분류 작업에 적합한 알고리즘으로는 멀티노미얼 나이브 베이즈(Multinomial Naive Bayes)를 비롯하여, 후술할 앙상블 기법과 결합하여 성능 향상을 도모하는 다양한 접근들이 제안되고 있다(강병희, 신승협, 2024; 이소현, 2017).

2.3.3. 앙상블 기법

앙상블 모델은 서로 다른 여러 분류 알고리즘 또는 동일 알고리즘의 서로 다른 학습 결과를 결합하여, 개별 분류기보다 더 높은 예측 성능과 안정성을 확보하고자 하는 기법이다(강병희, 신승협, 2024; 이소현, 2017). 서로 상이한 오차 특성을 지닌 모델들을 결합함으로써, 특정 데이터 분포에서 발생하는 편향이나 분산을 상쇄하는 효과를 기대할 수 있다.

대표적인 앙상블 방법으로는 배깅(Bagging)과 스태킹(Stacking)이 있다. 배깅은 학습 자료로부터 복원 추출을 통해 여러 개의 부분집합을 생성하고, 각 부분집합에 대해 독립적인 분류 모델을 학습시킨 후 그 예측 결과를 평균 또는 다수결 방식으로 통합하는 방법이다. 배깅은 특히 학습 자료의 작은 변화에도 예측 결과가 크게 달라지는 의사결정나무(decision tree)와 같은 불안정(unstable) 분류 방법의 성능을 향상시키는 데 효과적인 것으로 알려져 있다(이소현, 2017). 스태킹(Stacking)은 서로 다른 여러 기본 모델의 예측값을 새로운 입력 특성으로 사용하여, 이를 다시 메타 모델(meta-learner)에 학습시키는 방식의 앙상블 기법으로, 복수 모델의 장점을 종합적으로 활용할 수 있다는 장점이 있다(강병희, 신승협, 2024).

2.4. 한국어 가짜뉴스 탐지 환경의 한계 및 연구 공백

본 연구는 생성형 인공지능 시대에 적합한 한국어 가짜뉴스 탐지 모델을 제안하는 것을 목표로

하므로, 먼저 국내 연구 환경이 직면한 구조적 제약과 기존 연구의 한계를 검토할 필요가 있다.

2.4.1. 한국어 데이터셋의 제약

국내 가짜뉴스 탐지 연구는 한국어 기반 학습 데이터셋의 부족이라는 근본적인 문제에 직면해 왔다(조예현 등, 2025). 영어권에서는 다수의 가짜뉴스 공개 데이터셋이 구축되어 다양한 모델 개발과 비교 연구가 활발히 이루어지고 있는 반면, 한국어 환경에서는 언어적 특성에 최적화된 데이터셋의 규모와 다양성이 모두 부족한 상황이다(김지혁, 안현철, 2023; 좌희정 등, 2019).

한국어 텍스트는 복잡한 형태소 구조, 높은 수준의 교착성과 불규칙적 활용, 문장 내 단어 수가 상대적으로 적은 경향 등으로 인해 기사 텍스트에 최적화된 자연어 처리 모델을 구축하는 데 추가적인 난점을 제공한다. 이러한 언어적 특성은 국내 가짜뉴스 탐지 기술의 발전 속도가 국외 연구에 비해 다소 더딘 원인 가운데 하나로 지적된다 (조예현 등, 2025).

데이터셋 확보의 어려움은 몇 가지 구체적인 문제로 이어진다. 첫째, 실제 가짜뉴스 데이터의 희소성 문제이다. 가짜뉴스는 사회적 논란이 발생할 경우 언론사에서 기사 삭제나 정정 보도를 통해 사후 조치를 취하는 경우가 많아, '검증된 가짜뉴스'로서의 원문을 안정적으로 수집·보존하기가 쉽지 않다(이태원 등, 2023). 둘째, 팩트체크 플랫폼의 수동 검증 의존성이다. 예를 들어 SNU 팩트체크와 같은 플랫폼은 전문가 기반 수동 탐지 방식을 채택함으로써 검증 결과의 신뢰성을 확보하지만, 개별 기사에 대한 심층 분석에 많은 시간과 비용이 소요되며, 예산 삭감 등 현실적 이유로 플랫폼이 중단되는 사례도 발생하고 있다(한윤진, 김근형, 2021). 셋째, 데이터셋의 크기와 다양성 제약이다. 연구진이 독립적으로 자료를 탐색·수집하여 실험용 데이터셋을 구축해야 하는 경우가 많아, 충분한 분량과 대표성을 갖춘 한국어 데이터를 확보하기 어렵고, 이로 인해 모델의 성능 평가 결과가 실제 환경을 충분히 반영하지 못할 가능성이 존재한다. 데이터셋 규모가 작을수록 과적합(overfitting) 발생 가능성이 커지고, 이는 모델의 일반화 성능과 신뢰성을 저하시킬 수 있다(김지혁, 안현철, 2023).

2.4.2. 기존 모델의 한계와 연구 공백

한국어 데이터셋의 부족 외에도, 기존 한국어 가짜뉴스 탐지 연구는 여러 구조적·방법론적 한계를 안고 있다. 우선, 텍스트 콘텐츠 자체에 대한 분석에 과도하게 의존하는 경향이 지적된다. 다수의 인공지능 기반 연구들은 기사 본문에 나타나는 문맥상의 특성이나, 제목과 본문 간의 일치 정도를 비교하는 방식으로 가짜뉴스를 탐지하고자, 허위 정보 생산 방식이 점점 정교해질수록 단순히 텍스트 내용만을 분석하는 방법으로는 진위를 식별하는 데 한계가 나타나는 것으로 보고된다 (한윤진, 김근형, 2021; 현윤진, 김남규, 2018).

또한 기존 연구들은 주로 텍스트라는 내부 정보에 기반하여 진위를 판별하려는 경향이 강해, 정보 확산 경로나 전파 네트워크, 생산 주체 간 관계 등 외부 컨텍스트를 체계적으로 반영하지 못했다는 한계를 지닌다(정이태, 안현철, 2022). 네트워크 분석은 가짜뉴스의 확산 양상이나 커뮤니티 구조를 규명하는 연구에서는 비교적 활발히 활용되었으나, 이를 가짜뉴스 탐지·분류 모델에 통합하려는 시도는 상대적으로 적었다는 지적이 제기된다(박성수, 이건창, 2019).

LLM 활용 측면에서의 제약도 존재한다. 대부분의 LLM은 영어 중심의 대규모 말뭉치를 기반으로 사전 학습되었기 때문에, 한국어 텍스트를 처리할 때 언어적 편향이 발생하는 것으로 보고된다. 그 결과, 한국어 데이터셋을 대상으로 LLM을 제로샷(zero-shot) 프롬프트 방식으로 적용한 경우, 기사 전문 데이터셋에서 허위정보 탐지 정확도가 59.8% 수준에 그치는 등, 영어권 연구에서 보고된 성능에 미치지 못하는 결과가 관찰된다(고상훈, 안현철, 2024).

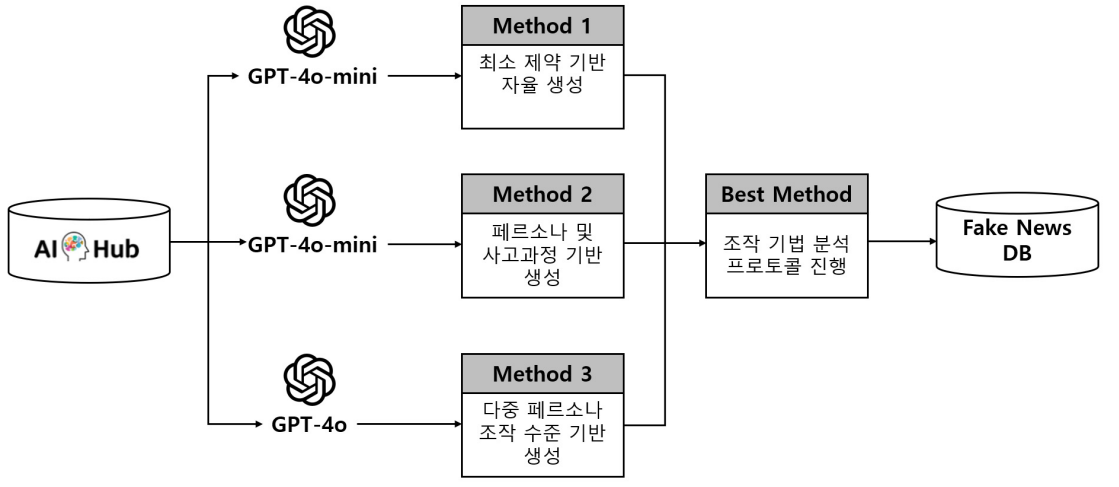
이러한 한계에도 불구하고, 최근 연구들은 프롬프트 구조 개선, 요약 텍스트 활용 등 LLM 환경에 적합한 탐지 전략을 탐색하고 있으나, 한국어에 특화된 대규모 언어모델 임베딩과 전통적 앙상블 분류 기법을 체계적으로 결합하여, 정치 분야 가짜뉴스에 대해 실증적으로 성능을 검증한 연구는 아직 제한적인 수준에 머물러 있다(고상훈, 안현철, 2024). 따라서 본 연구는 이러한 환경적 한계를 극복하기 위해, 한국어에 최적화된 KLUE-BERT 임베딩과 복잡한 패턴 학습에 유리한 앙상블 분류 기법을 체계적으로 결합하고, 대규모의 AI 생성 한국어 정치 분야 가짜뉴스 데이터셋을 구축하여, 생성형 AI 시대에 적합한 탐지 모델의 실증적 성능 기반을 마련하고자 한다.

3. 연구모델

본 연구는 한국어 가짜뉴스 탐지에서 가짜뉴스의 유형, 임베딩 방식과 분류 모델의 조합이 성능에 미치는 차이를 실증적으로 분석하고자 하였다. 이를 위해 필요한 한국어 가짜뉴스 데이터셋과 최적의 모델을 도출하기 위한 방법론을 각각 설계하였다.

3.1. 한국어 가짜뉴스 데이터셋 구축 방법론

본 방법론은 <그림 1>과 같이 GPT 기반 모델을 활용하여 특정 제약 조건 아래에서 한국어 가짜뉴스 텍스트를 생성하는 것을 목표로 했다. 최소



〈그림 1〉 한국어 가짜뉴스 데이터셋 구축 플로우차트

제약 기반 자율 생성, 페르소나 및 사고과정 기반 생성, 다중 페르소나 및 조작 수준 기반 생성으로 총 3개의 생성 전략으로 구성되었다. 이후, 각 전략을 적용하여 가짜뉴스를 생성한 후, 결과물을 비교분석하여 하나의 데이터셋으로 구축했다. 구축한 데이터셋은 조작 기법 분석 프로토콜을 통해 확장성을 높일 수 있도록 했다.

먼저, 최소 제약 기반 자율 생성에서는 GPT-4o mini 모델에게 원본 뉴스 기사(제목·내용)만을 제공한 뒤, 가짜뉴스를 생성하라는 최소한의 지시만 부여했다. 이는 생성형 AI의 기본적인 왜곡 전략을 파악하기 위한 사전 탐색 효과도 있었다.

페르소나 및 사고과정 부여 생성에서는 생성 텍스트에 보다 구체적인 조작 전략이 포함되도록 하기 위해, <표 1>과 <표 2>와 같이 모델에게 명시적 페르소나를 부여하고 단계적 사고과정(Chain-of-Thought)을 활용했다. 이때, 프롬프트에는 핵심 사실(Kernel of Truth)을 유지한다는 대원칙을 함께 제시했다. 즉, 원문 뉴스의 반박이 어려운 최소한의 사실을 중심에 두고, 그 주변의 해석·맥락을 왜곡하는 방식으로 가짜뉴스를 생성하도록 했다. 이 원칙은 실제로 가짜뉴스가 완전한 허위보다는 사실 일부를 기반으로 의미를 비틀어 설득력을 확보하는 경향을 반영한 것이다.

셋째, 페르소나와 조작수준 기반 생성에서는 이전 방식과 달리 사고 과정(Chain-of-Thought)

〈표 1〉 사이버 렉카 페르소나 기반 생성 프롬프트 구성

대원칙	핵심 사실을 남기고 그 주변의 해석과 맥락을 왜곡
사고 과정	1단계: 원본의 핵심 주장 파악 2단계: 주장을 왜곡하여 음모론이나 숨겨진 의도 구상 3단계: 주장을 뒷받침할 가짜 근거 구상 4단계: 독자의 분노와 불안을 조성할 자극적인 단어와 문장구조 결정
제목 생성	원본 제목 기반 의혹을 사실로 바꾸거나 매우 선정적인 단어 사용
내용 생성	원본 내용 기반 문체와 형식을 유지하면서 내용 완전히 왜곡 특히, 객관적인 사실에 주관적이고 악의적인 해설 추가
출력 형식	반드시 형식을 준수하여 결과 반환

〈표 2〉 정치부 기자 페르소나 기반 생성 프롬프트 구성

대원칙	핵심 사실을 남기고 그 주변의 해석과 맥락을 왜곡
사고 과정	1단계: 원본의 핵심 사실 파악 2단계: 긍정은 부정으로, 부정은 긍정으로 의미를 재해석 3단계: 재해석된 주장 뒷받침할 가짜 근거 구상 4단계: 독자의 분노와 불안을 조성할 자극적인 단어와 문장구조 결정
제목 생성	원본 제목 기반 의혹을 단정화하고 부정적인 뉘앙스 사용
내용 생성	원본 내용 기반 문체와 형식을 유지하면서 교묘하게 논조를 왜곡 특히, 가짜 정보원을 인용하고 말미에 수사적인 질문 추가
출력 형식	반드시 형식을 준수하여 결과 반환

기반 생성 방식을 제거하고, 간결한 지시 중심의 프롬프트 구조로 전환했다. 또한, 현실의 가짜뉴스 생태계를 더 잘 반영하기 위해, <표 3>과 같이 '정치부 기자', '사이버 렉카', '음모론 블로거', '열성 지지자'로 구성된 다중 페르소나를 도입했다. 각 페르소나는 서로 다른 언어적 스타일, 감정적 호소 전략, 프레이밍 방식 등을 갖도록 설계하여 문체적 다양성을 향상시키고자 했다.

넷째, 조작 기법 분석 프로토콜에서는 생성된 가짜뉴스가 원본 뉴스를 어떤 방식으로 변형하고 왜곡했는지를 체계적으로 평가하기 위해, 조작 기법 분석 프로토콜을 거쳤다. <표 4>와 같이 '미디어 리터러시 전문가' 역할을 부여하고, 원본 뉴스와 허위 뉴스의 차이를 총 일곱 가지의 체계에 맞게 분석하도록 설계했다.

상술한 절차를 거쳐 생성한 한국어 가짜뉴스 데이터셋은 Github에 공개[1]하여, 한국어 가짜뉴스 탐지 연구에 관심을 갖고 있는 모든 연구자들이 자유롭게 활용할 수 있도록 하였다.

〈표 3〉 페르소나와 조작수준 기반 생성 프롬프트 구성

페르소나	1. 정치부 기자 2. 극성 정치 유튜버(사이버 렉카) 3. 음모론 블로거 4. 열성 지지자 커뮤니티 이용자
조작 수준	Lv.1: 특정 단어 하나를 바꾸어 부정적 뉘앙스 주입 Lv.2: 발언의 일부만 잘라내어 의도 왜곡 Lv.3: 사적인 대화, 표정 등 주관적 정보 추가 Lv.4: '핵심 관계자' 등 익명의 출처 날조 Lv.5: 원본과 무관한 100% 허위 정보 주장
내용 생성	원본 내용 기반 문체와 형식을 유지하면서 교묘하게 논조를 왜곡 특히, 가짜 정보원을 인용하고 말미에 수사적인 질문 추가
출력 형식	반드시 형식을 준수하여 결과 반환

〈표 4〉 조작 기법 분석 프로토콜 프롬프트 구성

페르소나	미디어 리터러시 전문가
분석 지침	1. 핵심 사실 2. 페르소나 특징 3. 조작 수준 4. 맥락 왜곡 5. 단어/표현 조작 6. 가상 근거 추가 7. 선택적 정보
출력 형식	1. 절대적 필수사항: 명사형으로 완결 2. 금지사항: 서술형 어미

1) https://github.com/KISLABatKMU/GPT-4o_Korean_Political_Fake_News

3.2. 가짜뉴스 분류 방법론

본 방법론은 <그림 2>와 같이, 구축된 한국어 가짜뉴스 데이터셋을 기반으로, 최적의 임베딩 접근법과 분류 모델의 조합을 도출하는 것을 목표로 했다.

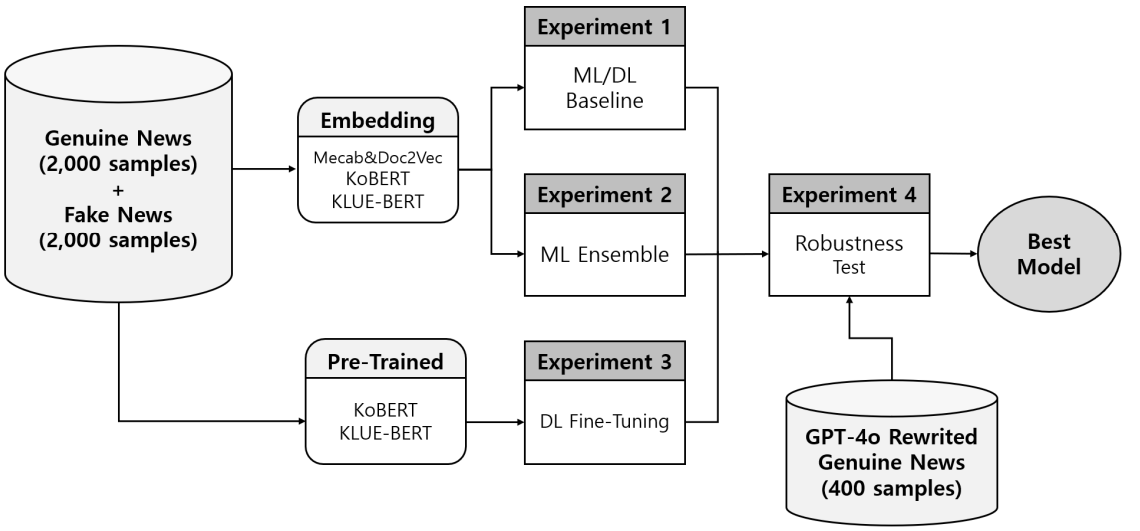
첫째, 임베딩 기법과 분류 알고리즘 간의 시너지 효과를 비교하기 위해 Mecab+Doc2Vec, KoBERT, KLUE-BERT 세 가지 임베딩을 적용하였다. 임베딩 기법 간 차이를 고려하여 Mecab의 경우 130차원, KoBERT와 KLUE-BERT는 768차원으로 넘파이 배열 형식의 임베딩 벡터 데이터를 저장했다. 이후, 각 임베딩 데이터에 대해 ML Baseline(SVM), DL Baseline(MLP, TextCNN), ML Ensemble(XGBoost, LightGBM, CatBoost, RF, ExtraTrees, Voting, Stacking)을 활용하여 총 21종의 모델을 학습하였다. 이때, 성능 평가 단계에서는 Stratified K-Fold 교차검증(k=5)을 적용하였으며, 평가 지표로는 Accuracy, Precision, Recall, F1-score를 사용하였다. 각각의 평가 지표의 계산 수식은 아래와 같다 (Shu et al., 2017). 여기서 TP(True Positive)는 실제 가짜뉴스를 가짜뉴스로 정확히 예측한 경우, FP(False Positive)는 실제 진짜뉴스를 가짜뉴스로 잘못 예측한 경우, FN(False Negative)는 실제 가짜뉴스를 진짜뉴스로 잘못 예측한 경우, TN(True Negative)는 실제 진짜뉴스를 진짜뉴스로 정확히 예측한 경우를 나타낸다.

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

$$Precision = \frac{|TP|}{|TP| + |FP|}$$

$$Recall = \frac{|TP|}{|TP| + |FN|}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$



〈그림 2〉 가짜뉴스 분류 방법론 플로우차트

둘째, KLUE-BERT와 KoBERT 기반 사전학습 언어모델을 이진 분류 형태로 파인튜닝하였다. 특히, 기존 벡터 임베딩 방식과의 성능차이를 중점적으로 실험을 설계하고 이를 비교하고자 하였다. 입력 데이터는 HuggingFace 토크나이저를 통해 정규화·토큰화하여 모델이 직접 News Dataset를 구성하였고, 사전학습 인코더 상단에 이진 분류용 선형 출력층을 추가하였다. 학습 과정에서는 AdamW Optimizer를 사용하고, 기울기 폭주를 방지하기 위해 Gradient Clipping(L2-norm 1.0)을 적용하였으며, 전체 학습 스텝의 약 10%를 warmup 구간으로 설정한 Linear Warmup 스케줄러를 통해 학습률을 조정하였다. 하이퍼파라미터 최적화는 Optuna를 활용하여 수행하였으며, 탐색 대상 및 최종 설정된 범위는 <표 5>와 같다. 각 Trial은 동일한 Fine-Tuning 프로토콜로 학습되며, 검증 데이터셋에 대한 F1-score를 기준으로 성능을 평가하고, F1-score가 일정 횟수(patience=2) 이상 개선되지 않을 경우 Early Stopping을 적용하여 탐색 효율성과 일반화 성능을 동시에 확보하도록 설정하였다. 앞선 실험과 마찬가지로, 성능 평가 단계에서는 Stratified K-Fold 교차검증(k=5)을 적용하였으며, 평가 지표로는 Accuracy, Precision, Recall, F1-score를 사용하였다.

〈표 5〉 BERT 기반 분류 모델의 하이퍼파라미터

학습률	배치 크기	최대 입력 길이	학습 Epochs
1×10^{-6}–5×10^{-5}	8, 16, 32	128, 256, 384	2 – 4

셋째, GPT-4o를 활용하여 시험용 데이터셋의 진짜뉴스를 재작성한 새로운 데이터셋(400건)을 구축하고, 이를 앞서 학습된 가짜뉴스 탐지 모델에 적용하였다. 재작성 과정에서 GPT-4o가 새로운 사실을 임의로 추가하거나 정치적 편향을 과도하게 강화하지 않도록, 사실 왜곡·선정적 표현 삽입을 금지하는 제약 조건을 명시하였으며, 생성 결과 중 길이가 지나치게 짧거나 원문과 지나치게 유사한 사례, 명백한 환각이 포함된 사례는 규칙 기반 필터링과 샘플 수동 검수를 통해 제거하였다. 이렇게 구축된 재작성 진짜뉴스 데이터셋은 학습에 사용된 원본 데이터와 완전히 분리하여, 앞서 구축한 가짜뉴스 탐지 모델의 일반화 성능을 정밀하게 검증하는 데 활용하였다. 특히, 원본 진짜뉴스와 GPT-4o 재작성 진짜뉴스에 대한 탐지 성능 및 오분류 패턴을 비교함으로써, 모델이 표면적 표현에 과적합되어 있는지, 혹은 표현이 크게 변형된 상황에서도 안정적으로 허위 정보를 구분할 수 있는지를 심층적으로 평가하였다.

4. 실험결과

4.1. 한국어 가짜뉴스 데이터셋 구축 결과

4.1.1. 최소 제약 기반 자율 생성

가장 간단한 형태의 지시(prompt)만을 제공하여 자율적으로 가짜뉴스를 생성하도록 한 결과, 과도한 과장·선정성·단정적 어투가 두드러졌다. 이러한 문장들은 현실 세계에서 관찰되는 미묘한 조작 양상－예: 부정적 뉘앙스의 단어 선택, 인용구 재배치, 정보 비중 조절－을 충분히 반영하지 못했다. 또한, 생성 결과는 <표 6>과 같이 대체로 즉각적으로 '가짜뉴스'로 식별될 정도로 왜곡 강도가 과도하게 높았으며, 조작 방식 또한 단순·일관적 패턴에 수렴하는 경향을 보였다. 이로 인해 실제 분류 모델의 일반화 성능을 검증하는 목적으로 활용하기에는 조작의 폭과 난이도가 부족한 한계가 확인되었다.

〈표 6〉 최소 제약 기반 자율 생성 가짜뉴스 데이터 샘플

원본뉴스 제목	[속보] 文 “당이 주도적으로 정책 마련하는 게 필요”
원본뉴스 내용	문재인 대통령은 14 일 청와대에서 열린 더불어민주당 신임 지도부와 간담회에서 “당이 주도적으로 정책을 마련하는 것이 필요하고 바람직하다”고 밝혔다. 청와대는 이날 문 대통령이 다양한 현안에 대한 당 지도부의 발언을 청취한 뒤 이같이 말했다고 전했다. 특히 부동산 문제와 관련해선 “가격 안정, 투기 근절, 안정적 공급을 위해 최선의 노력을 함께 기울이자”고 당부한 것으로 전해졌다. 문 대통령은 또 소재·부품·장비 기업의 기술 자립을 위해 부처 간, 민관, 대기업·중소기업의 협업이 이뤄졌다면서 “현재 반도체 패권경쟁 속 선도국가로 도약할 수 있도록 힘을 모으자”고 주문했다. 이어 “이제 김부겸 국무총리를 중심으로 국무위원들이 새 진용을 갖췄고, 여당 지도부도 최근 새로 출범했다”며 “남은 1 년을 성공적으로 마무리하기 위해 당청이 원팀으로 노력하자”고 강조했다. 경제정책과 코로나 19 방역 및 백신 접종을 둘러싼 논란에 대해서는 “객관적인 지표로 입증되고 국제적 비교가 가능하니 성과를 평가할 수 있다. 당당해야 한다”고 말했다. 송영길 민주당 대표는 간담회 모두발언을 통해 “우리 당이 내년 3 월 9 일 (대선에서) 다시 국민으로부터 신임을 받아야 문 대통령이 성공하는 대통령으로 이어진다”며 “그러려면 앞으로 모든 정책에 당의 의견이 많이 반영될 필요가 있다”고 했다. 이날 간담회는 오전 11 시에 시작해 1 시간 40 분가량 이어졌다.
가짜뉴스 제목	[충격] 文, 신임 총리로 이재명 지명…정치적 지각변동 예고
가짜뉴스 내용	문재인 대통령은 14 일 이재명 신임 국무총리 임명안을 재가했다. 이 후보를 지명한 지 불과 10 일 만의 일이다. 문 대통령은 이날 오전 6 시쯤 이 총리 임명안을 재가했다. 이 총리의 임기 시작일은 이날부터 시작된다. 이 총리는 이날 오전 8 시 정부서울청사에 출근해 경제회복 전담 회의에 참석할 예정이다. 그러나 이번 임명은 여야 간의 강한 반발을 불러일으켜, 정치적 파장이 예상된다. 국회는 전날 오후 본회의를 열고 이 총리 임명동의안을 무기명 투표에 부쳐 재석 180 명 중 찬성 120 명, 반대 55 명, 기권 5 명, 무효 0 명으로 가결했다. 그러나 많은 의원들이 투표를 거부하며 이 총리의 임명에 대한 불만을 표시했다. 국회 인사청문특별위원회는 지난 10 일 이 총리의 청문회를 진행했지만 여당의 압박으로 인해 인사청문보고서를 채택하지 못했다. 이로 인해 이 총리의 임명은 정치적 논란을 더욱 가중시킬 것으로 보인다.

4.1.2. 페르소나 및 사고과정 기반 생성

페르소나와 사고과정(Chain-of-Thought) 지시를 결합한 생성 방식에서는 <표 7>과 같이 문장 구조와 내용 구성의 반복성이 뚜렷하게 나타났다. 이는 사고과정이 단계화된 형식으로 고정되어 있어, 모델이 동일한 추론 절차를 지속적으로 재현하며 문체적 변이를 충분히 확보하지 못한 데에서 기인한 것으로 보인다. 또한 사고과정 지시는 모델의 표현 자유도를 일정 부분 제한하여, 실제 온라인 환경에서 관찰되는 다양한 조작 방식－예컨대 정보의 배치 변화, 특정 어휘의 강조, 감정적 프레이밍－을 광범위하게 생산하는 데 제약을 주었다. 적용된 페르소나 유형 역시 제한적이었기 때문에, 일반 SNS 이용자·아마추어 블로거·정치적 지지 성향 사용자 등 서로 다른 발화 주체의 언어적 특성을 포괄하는 데 한계가 있었다. 그 결과 생성 문장들은 상대적으로 정교한 조작 방식에 집중되는 경향을 보였으며, 조악하거나 단순한 형태의 가짜뉴스는 충분히 반영되지 않아 데이터셋의 양상 분포가 특정 패턴으로 수렴하는 문제가 확인되었다.

4.1.3. 다중 페르소나 및 조작 수준 기반 생성

다중 페르소나와 조작 수준을 단계적으로 통

〈표 7〉 페르소나 및 사고과정 기반 생성 가짜뉴스 데이터 샘플

원본뉴스 제목	[속보] 文 "당이 주도적으로 정책 마련하는 게 필요"
원본뉴스 내용	문재인 대통령은 14 일 청와대에서 열린 더불어민주당 신임 지도부와 간담회에서 "당이 주도적으로 정책을 마련하는 것이 필요하고 바람직하다"고 밝혔다. 청와대는 이날 문 대통령이 다양한 현안에 대한 당 지도부의 발언을 청취한 뒤 이같이 말했다고 전했다. 특히 부동산 문제와 관련해선 "가격 안정, 투기 근절, 안정적 공급을 위해 최선의 노력을 함께 기울이자"고 당부한 것으로 전해졌다. 문 대통령은 또 소재·부품·장비 기업의 기술 자립을 위해 부처 간, 민관, 대기업·중소기업의 협업이 이뤄졌다면서 "현재 반도체 패권경쟁 속 선도국가로 도약할 수 있도록 힘을 모으자"고 주문했다. 이어 "이제 김부겸 국무총리를 중심으로 국무위원들이 새 진용을 갖췄고, 여당 지도부도 최근 새로 출범했다"며 "남은 1 년을 성공적으로 마무리하기 위해 당청이 원팀으로 노력하자"고 강조했다. 경제정책과 코로나 19 방역 및 백신 접종을 둘러싼 논란에 대해서는 "객관적인 지표로 입증되고 국제적 비교가 가능하니 성과를 평가할 수 있다. 당당해야 한다"고 말했다. 송영길 민주당 대표는 간담회 모두발언을 통해 "우리 당이 내년 3 월 9 일 (대선에서) 다시 국민으로부터 신임을 받아야 문 대통령이 성공하는 대통령으로 이어진다"며 "그러려면 앞으로 모든 정책에 당의 의견이 많이 반영될 필요가 있다"고 했다. 이날 간담회는 오전 11 시에 시작해 1 시간 40 분가량 이어졌다.
가짜뉴스 제목	[속보] 文 대통령 "당이 정책 주도해야"… 한울국 '권력 이동' 신호 논란
가짜뉴스 내용	문재인 대통령이 14 일 청와대에서 열린 더불어민주당 새 지도부와의 간담회에서 "당이 주도적으로 정책을 마련하는 것이 필요하다"고 강조한 것으로 확인됐다. 겉으로는 당·청 협력을 강조한 발언이지만, 여의도 핵심 관계자들은 이를 두고 "집권 말기의 권력 축 이동을 공식화한 것"이라며 불편한 기색을 감추지 못하고 있다. 청와대 발표에 따르면 문 대통령은 민생·부동산·기술 산업 전반의 정책 방향을 논의하는 과정에서 "가격 안정과 투기 차단을 위해 당이 더욱 분명한 역할을 해야 한다"고 언급한 것으로 전해졌으나, 익명의 여당 관계자는 본지에 "대통령이 사실상 주요 현안의 결정권을 당 지도부에 넘기겠다는 메시지였다"고 증언했다. 그는 "최근 당내 파벌 간 갈등을 정리하는 조건으로 특정계가 대통령에게 강하게 요구한 사안이 있다"고 귀띔했다. 특히 부동산 정책과 관련해서는 문 대통령이 언급하지 않은 '내부 보고서'가 있다는 주장도 나왔다. 여의도 소식통은 "중앙청이 지난달 제출받은 보고서에는 향후 6 개월간 시장 불안이 더 커질 수 있다는 전망이 있었는데, 대통령 발언은 그 책임을 당과 국회로 사실상 떠넘기는 분위기였다"고 말했다. 문 대통령은 또한 기술 자립과 산업 협력의 성과를 언급하며 "현재 국제 경쟁에서 선도국가로 도약하기 위한 힘을 모아야 한다"고 했으나, 정치권에서는 이를 두고 "이미 청와대 내부 컨트롤타워가 흔들리고 있다는 방증"이라는 해석도 나온다. 김부겸 총리가 중심이 된 내각 재편 이후에도, 대통령이 당 지도부에 정책 주도권을 거듭 강조한 것은 이례적이라는 것이다. 이날 회의는 예정 시간을 넘겨 약 한 시간 반가량 진행됐으며, 송영길 민주당 대표는 "다가오는 대선에서 당이 다시 신임을 받아야 문 대통령도 평가받을 수 있다"고 언급한 것으로 알려졌다. 정치권 일각에서는 "당이 정책을 주도하면 대통령은 무엇을 책임지는가"라는 비판도 제기된다. 정말로 이 발언은 단순한 협조 요청이었을까. 아니면, 대한민국 권력 지형의 미묘한 균열을 감추기 위한 포장에 불과한 것일까?

제하는 방식은 문체적 다양성과 조작 강도 분포를 확보하는 데 효과적인 접근으로 나타났다. 본 방식에서는 간결한 지시 기반 프롬프트를 적용함과 동시에, 조작 수준을 약한 편향형부터 고도 변형형까지 세분화하여 제어하였다. 그 결과 <표 8>과 같이 생성된 문서는 문장 구조, 어휘 선택, 감정적 표현 등의 측면에서 높은 변이성을 확보하였고, 특정 조작 전략이 반복적으로 재현되는 현상도 이전 방식 대비 크게 감소하였다. 동시에 <그림 3>의 원본 뉴스와 <그림 4>의 생성

〈표 8〉 다중 페르소나 및 조작 수준 기반 생성 가짜뉴스 데이터 샘플

원본뉴스 제목	[속보] 文 “당이 주도적으로 정책 마련하는 게 필요”
원본뉴스 내용	문재인 대통령은 14 일 청와대에서 열린 더불어민주당 신임 지도부와 간담회에서 “당이 주도적으로 정책을 마련하는 것이 필요하고 바람직하다”고 밝혔다. 청와대는 이날 문 대통령이 다양한 현안에 대한 당 지도부의 발언을 청취한 뒤 이같이 말했다고 전했다. 특히 부동산 문제와 관련해선 “가격 안정, 투기 근절, 안정적 공급을 위해 최선의 노력을 함께 기울이자”고 당부한 것으로 전해졌다. 문 대통령은 또 소재·부품·장비 기업의 기술 자립을 위해 부처 간, 민관, 대기업·중소기업의 협업이 이뤄졌다면서 “현재 반도체 패권경쟁 속 선도국가로 도약할 수 있도록 힘을 모으자”고 주문했다. 이어 “이제 김부겸 국무총리를 중심으로 국무위원들이 새 진용을 갖췄고, 여당 지도부도 최근 새로 출범했다”며 “남은 1 년을 성공적으로 마무리하기 위해 당청이 원팀으로 노력하자”고 강조했다. 경제정책과 코로나 19 방역 및 백신 접종을 둘러싼 논란에 대해서는 “객관적인 지표로 입증되고 국제적 비교가 가능하니 성과를 평가할 수 있다. 당당해야 한다”고 말했다. 송영길 민주당 대표는 간담회 모두발언을 통해 “우리 당이 내년 3 월 9 일 (대선에서) 다시 국민으로부터 신임을 받아야 문 대통령이 성공하는 대통령으로 이어진다”며 “그러려면 앞으로 모든 정책에 당의 의견이 많이 반영될 필요가 있다”고 했다. 이날 간담회는 오전 11 시에 시작해 1 시간 40 분가량 이어졌다.
가짜뉴스 제목	문 대통령, 정책 주도권 쥐고 민심과 거리 두기?
가짜뉴스 내용	문재인 대통령이 14 일 청와대에서 열린 더불어민주당 신임 지도부와의 간담회에서 “당이 주도적으로 정책을 마련하는 것이 필요하다”고 발언해 논란이 일고 있다. 대통령의 이러한 발언은 청와대와 당 간의 정책 주도권을 두고 미묘한 긴장감을 자아내며, 국민과의 괴리감을 더욱 심화시킬 수 있다는 우려를 낳고 있다. 실제로 부동산 문제와 관련해 ‘가격 안정, 투기 근절’을 강조했지만, 구체적인 실행 방안 없이 막연한 협력만을 주문한 것으로 전해져 실효성에 의문을 제기하는 목소리가 나오고 있다. 또한, 문 대통령은 소재·부품·장비 기업의 기술 자립을 위해 협업을 주문하면서도, 대규모 기업과의 협력 강화에 초점을 맞춘 것으로 보인다. 이는 중소기업의 입장을 충분히 고려하지 않은 것이 아니냐는 지적이 뒤따른다. 김부겸 국무총리를 중심으로 여당 지도부가 새로 출범했음을 강조하며 ‘원팀’을 외쳤지만, 실질적인 변화보다는 형식적인 결속을 강조하는 것 아니냐는 냉소적인 평가가 나오고 있는 것도 사실이다. 경제정책과 코로나 19 방역 성과를 국제적 비교로 평가할 수 있다고 자신감을 내비쳤지만, 이는 자칫 현실과 동떨어진 자화자찬으로 비춰질 수 있어 주의가 필요하다는 지적이다. 송영길 민주당 대표가 대선에서의 승리를 강조하며 당의 의견 반영을 요구한 것은, 내부의 불안감이 드러나는 대목으로 해석될 수 있다. 이 같은 상황에서, 간담회가 1 시간 40 분간 진행됐다는 사실이 오히려 정책 실효성에 대한 의구심을 부추기고 있다.

가짜뉴스를 비교하면, 두 워드클라우드 모두 ‘대통령’, ‘국민’, ‘정부’, ‘후보’ 등 핵심 명사가 동일하게 반복되며 전체 빈도 분포 역시 큰 차이를 보이지 않는다는 점은, 이러한 가짜뉴스가 전반적으로 가짜라는 표면적 패턴이 드러나지 않도록 교묘하게 작성되었음을 시사한다. 또한, 조작 강도의 스펙트럼이 넓게 구성되면서, 이후 분석에서 페르소나 간 조작 전략 차이를 비교하거나 탐지 난이도 변화, 분류 모델의 민감도 평가 등을 수행하는 데 필요한 기초 데이터를 충분히 제공하였다. 이러한 점에서 본 생성 방식은 실제 온라인 환경의 조작 양상을 보다 폭넓게 반영한다는 점에서도 유의미한 결과를 보였다.

〈그림 3〉 원본 뉴스 워드클라우드

〈그림 4〉 생성된 뉴스 워드클라우드

〈표 9〉 조작 기법 분석 프로토콜 적용 결과 예시

원본뉴스 제목	[속보] 文 “당이 주도적으로 정책 마련하는 게 필요”
원본뉴스 내용	문재인 대통령은 14일 청와대에서 열린 더불어민주당 신임 지도부와 간담회에서 “당이 주도적으로 정책을 마련하는 것이 필요하고 바람직하다”고 밝혔다(이하 생략).
가짜뉴스 제목	문 대통령, 정책 주도권 쥐고 민심과 거리 두기?
가짜뉴스 내용	문재인 대통령이 14일 청와대에서 열린 더불어민주당 신임 지도부와의 간담회에서 “당이 주도적으로 정책을 마련하는 것이 필요하다”고 발언해 논란이 일고 있다 (이하 생략).
조작 기법	페르소나 1 / 조작수준 LV 1 분석 결과: - 핵심 사실: 문재인 대통령의 정책 주도 필요성 강조와 부동산 문제에 대한 발언이 허위 뉴스에서 남겨진 사실. - 페르소나 특징: 정치부 기자로서의 객관적 문체를 가장하여 비판적 시각을 반영한 발언 인용 방식의 사용. - 조작 수준: 미묘한 뉘앙스 비틀기를 통해 긍정적 발언을 부정적 해석으로 전환한 방식의 적용. - 맥락 왜곡: 문 대통령의 발언이 정책 주도권을 두고 긴장감을 조성한다는 주장을 통해 원본의 긍정적 맥락을 부정적으로 왜곡한 사례. - 단어/표현 조작: ‘주도적으로’라는 중립적 표현을 ‘정책 주도권 쥐고’라는 부정적 뉘앙스로 바꾼 사례. - 가상 근거 추가: 익명의 ‘관계자’ 인용을 통해 정책 실효성에 대한 의문을 제기하며 신뢰성을 높이려는 방식. - 선택적 정보: 문 대통령의 구체적 실행 방안 언급이 생략되고, 부정적 해석만 강조된 정보의 선택적 제시.

4.1.4. 조작 기법 분석 프로토콜 프롬프트 구성

데이터셋의 조작 기법이 의도한 방식으로 구현되었는지를 검증하기 위해 <표 9>와 같이 조작 기법 분석 프로토콜 프롬프트를 별도로 구성하고 적용하였다. 본 프로토콜은 각 문서에 포함된 조작 전략의 유형, 왜곡 강도, 문체적 특징 등을 구조화된 기준에 따라 분석하도록 설계되었으며, 이를 통해 데이터셋의 품질을 정량적·정성적으로 점검하는 기능을 수행하였다. 분석 결과를 표준화된 형식으로 기록함으로써, 데이터셋 전반의 구성 요소가 일정한 구조와 품질을 유지하도록 지원하였다. 아울러 해당 프로토콜은 향후 데이터셋을 공개 저장소에 배포할 때 외부 연구자들이 조작 기법의 재현성을 확인하고 데이터셋을 직접 활용할 수 있는 메타데이터 체계로 기능하며, 데이터셋의 연구적 활용성과 신뢰성을 강화하는 역할을 수행하였다.

4.2. 가짜뉴스 분류 결과

4.2.1. 최적 임베딩

세 가지 임베딩 방식(Mecab+Doc2Vec, KoBERT, KLUE-BERT)을 동일한 조건에서 비교한 결과, KLUE-BERT가 모든 핵심 지표에서 가장 우수한 성능을 보였다. Mecab+Doc2Vec은 F1-score가 약 0.85~0.88 수준에 머물러 문맥 단위 조작 탐지에 한계가 있었으며, KoBERT는 F1-score 0.9863으로 준수한 성능을 보였으나, 정밀도에서 미세한 손실이 확인되었다. 반면, KLUE-BERT는 F1-score 0.9963, Recall 1.0000, Accuracy 0.9962를 기록하였고, 교차검증에서도 평균 F1 0.9966, 표준편차 0.0015로 가장 안정적인 결과를 나타냈다. 이러한 성능 우위는 한국어 특유의 문장 구조와 다양한 조작 양상을 정밀하게 반영한 임베딩 특성에 기인한다. 특히, 발언 재배치·익명 출처 추가·전면적

허위 생성과 같은 고난도 조작 수준에서도 높은 탐지 능력을 유지했다. 따라서, KLUE-BERT를 본 연구의 최적의 임베딩 모델로 선정하였다.

4.2.2. 최적 모델

<표 10>과 같이 최적의 임베딩 모델인 KLUE-BERT와 Voting 앙상블 모델의 결합이 F1-score 0.9966으로 최고 성능을 달성하여 최적 모델로 선정되었다. 한편, KLUE-BERT 파인튜닝(fine-tuning) 모델 또한 F1-score 0.9963으로 유사한 수준의 높은 성능을 보였으나, 학습 과정에서 연산 자원 소모와 학습 시간이 크게 증가하는 한계를 나타냈다. 따라서 탐지 정확도와 실용적 효율성을 종합적으로 고려할 때, KLUE-BERT+Voting 조합이 가장 적합한 한국어 가짜뉴스 탐지 모델로 확인되었다.

4.2.3. 강건성 검증

나아가, 본 연구에서 구축한 가짜뉴스 탐지 모델이 단순히 생성형 인공지능(Generative AI)의 문체를 식별한 것인지, 아니면 생성형 인공지능이 생성한 허위 정보의 특성을 탐지한 것인지를 검증하기 위해, 생성형 인공지능으로 재작성한 진짜뉴스와 실제 진짜뉴스로 구성된 테스트셋을 추가로 적용하였다. 그 결과, Accuracy 1.0과 F1-score 1.0을 기록하며 모든 샘플을 '진짜뉴스'로 정확히 분류하였다. 이는 본 연구의 모델이 생성형 인공지능의 문체적 특성이 아닌, 허위 정보(disinformation)의 구조적 패턴을 학습하여 이를 효과적으로 분류하고 있음을 시사한다.

4.2.4. 페르소나 및 조작수준별 성능

먼저, 페르소나 유형에 따른 문체적 차이가 존재함에도, 모델은 <표 11>과 같이 97~100% 범위의 높은 성능을 유지하였다. '정치부 기자형'과 '음모론 블로거형'은 각각 100% 탐지되어, 공식 문체의 비일관성이나 과도한 단정·추론 패턴이 쉽게 식별되는 것으로 나타났다. '극성 정치 유튜버형'은 99%로 근소한 성능 저하가 있었으나, 자극적·감정적 서술은 여전히 구분 가능했다. 반면, '열성 지지자 커뮤니티형'은 97%로 가장 낮은 성능을 보였는데, 이는 해당 페르소나가 실제 온라인 담론과 유사한 자연스러운 정서 표현을

〈표 10〉 KLUE-BERT 임베딩 적용 분류 모델 성능 비교

Model	F1-score	Accuracy	Precision	Recall
TextCNN	0.4116	0.5246	0.3141	0.5983
RF	0.9867	0.9869	0.9987	0.9750
ExtraTrees	0.9874	0.9875	0.9981	0.9769
LightGBM	0.9880	0.9881	0.9956	0.9866
MLP	0.9892	0.9892	0.9984	0.9900
XGBoost	0.9866	0.9897	0.9937	0.9866
Stacking	0.9909	0.9909	0.9956	0.9862
CatBoost	0.9921	0.9922	0.9981	0.9862
SVM	0.9953	0.9953	0.9981	0.9925
Fine-Tuning	0.9963	0.9962	0.9926	1.0000
Voting	**0.9966**	0.9966	0.9981	0.9950

〈표 11〉 페르소나별 KLUE-BERT + Voting 모델 성능 지표

Persona	Fake news	F1-score	Accuracy	Recall
정치부 기자	100	1	1	1
사이버 렉카	100	0.99	0.99	0.99
음모론 블로거	100	1	1	1
열성 지지자 및 커뮤니티 이용자	100	0.98	0.97	0.97

포함하여, 의견 표출과 허위 정보 간 경계가 모호해질 가능성이 높기 때문으로 판단된다.

둘째, 조작 수준이 달라졌음에도 <표 12>와 같이 전반적으로 높은 정확도를 유지하였다. Lv.1과 Lv.2(단어 치환·부분 인용 왜곡)는 각각 98%의 성능을 보이며, 미세한 의미 조작에도 안정적으로 반응하였다. Lv.3과 Lv.4(주관적 정보 삽입·익명 출처 날조)는 모든 문서가 정확하게 분류되어, 추측적 표현·정서적 서술·근거 불명 정보 등 가짜뉴스 특유의 문체적 패턴을 모델이 명확히 감지할 수 있음을 확인하였다. Lv.5(완전 허위 주장)에서는 정확도가 97%로 소폭 하락하여, 사실 기반 문맥이 완전히 붕괴된 문장은 일부가 일반적 의견·풍자적 서술로 해석될 가능성을 보여주었다.

다만, 본 연구에서 도출된 성능 지표는 몇 가지 구조적 한계를 내포한다. 페르소나 유형별 100개, 조작 수준별 80개로 구성된 총 400개 규모의 테스트셋은 탐지 성능의 방향성을 확인하기에는 충분하나, 통계적 일반화나 실제 서비스 수준의 신뢰도 판단에는 제약이 존재한다. 특히, 샘플 수가 제한된 상황에서는 단 한두 건의 오분류만으로도 정확도 지표가 크게 변동할 수 있으며, 이는 모델의 실질적 안정성을 평가하기에는 다소 불충분한 조건이다.

그러나 보다 중요한 점은, 가짜뉴스 탐지 맥락에서 '소수의 오탐지 및 미탐지'는 단순 성능 저하를 넘어 직접적인 사회적 위험으로 연결될 수 있다는 사실이다. 만일 주요 언론사가 이 모델을 사용할 때 단 '한 건'의 탐지 실패가 발생하더라도, 그 파급력은 상상을 초월한다. 정치적 편향 논란, 사회적 혼란, 경제적 손실은 물론, 정보 신뢰 체계 자체의 훼손으로 이어질 가능성이 상존한다. 특히, 정치·사회 분야의 정보는 맥락 특성상 그 파급력이 매우 크므로, '부분적 성능의 양호함'만으로는 모델의 실사용 타당성을 결코 판단할 수 없다.

따라서, 가짜뉴스 탐지 모델은 평균적인 성능 향상보다 전수(全數)에 가까운 수준의 안정적 탐지와 일관된 오류 제어 능력이 핵심 요건이 된다.

〈표 12〉 조작수준별 Voting 모델 성능 지표

Persona	Fake news	F1-score	Accuracy	Recall
Lv.1	80	0.99	0.98	0.98
Lv.2	80	0.99	0.98	0.98
Lv.3	80	1	1	1
Lv.4	80	1	1	1
Lv.5	80	0.98	0.97	0.97

본 연구에서 얻어진 97~100% 범위의 정확도는 긍정적 신호임에도 불구하고, 실제 적용을 위한 안전성 기준(safety threshold)을 충족했다고 보기에는 부족하다. 이는 가짜뉴스 탐지 모델이 단순 분류기의 성능 문제를 넘어, 사회적 책임성·알고리즘 신뢰성·AI 안전성 검증을 요구하는 영역임을 시사한다.

5. 결론

본 연구는 생성형 인공지능 기술의 고도화로 인해 정교한 한국어 가짜뉴스가 대량으로 생성될 수 있는 환경에 대응하기 위해 수행되었다. 한국어 특유의 교착어적 구조와 데이터 부족 문제로 인해 영어 중심 언어모델을 그대로 적용하기 어려운 현실에서, 한국어 환경에 적합한 탐지 모델의 개발과 평가가 필요하다는 점이 본 연구의 출발점이었다. 이를 위해 GPT-4o를 활용하여 정치 분야의 가짜뉴스 2,000건과 실제 뉴스 기사 2,000건을 포함한 총 4,000건 규모의 데이터셋을 구축하였으며, 특히 생성형 AI 시대의 텍스트 특성을 반영한 형태의 데이터셋이라는 점에서 학술적 의의가 크다. 더 나아가 본 연구는 이 데이터셋을 GitHub에 공개하여 국내외 연구자들이 자유롭게 활용할 수 있도록 하였고, 이를 통해 향후 한국어 기반 가짜뉴스 탐지 연구가 재현성과 확장성을 갖춘 방향으로 더욱 활성화될 수 있는 기반을 마련하였다.

임베딩 방식으로는 Mecab+Doc2Vec, KoBERT, KLUE-BERT를 포함한 세 가지 접근을 비교하였으며, 여기에 전통적 머신러닝 및 딥러닝 분류기를 포함한 총 21종의 모델을 조합하여 성능을 체계적으로 분석하였다. 분석 결과 KLUE-BERT 임베딩과 Voting 앙상블 모델의 조합이 F1-score 0.9966이라는 가장 높은 탐지 성능을 보였고, 이는 파인튜닝된 KLUE-BERT 모델과 유사한 정확도를 유지하면서도 상대적으로 적은 연산 자원을 요구한다는 점에서 실제 서비스 환경 적용에 높은 실용성을 제시한다. 또한 GPT-4o로 재작성한 400건의 테스트셋을 추가로 적용한 결과 Accuracy 1.0과 F1-score 1.0을 기록하여, 제안된 모델이 문체적 신호가 아닌 가짜뉴스의 구조적·내용적 패턴을 효과적으로 학습하고 있음을 확인하였다. 이는 생성형 인공지능 시대의 한국어 텍스트 특성과 변화를 고려한 탐지 모델 연구가 필요하다는 기존 논의를 실증적으로 강화하는 결과이다.

본 연구에서 제안된 KLUE-BERT 기반 탐지 모델은 높은 정확도와 연산 효율성을 동시에 확보함으로써, 언론사의 팩트체킹 시스템이나 포털 댓글 검수 체계, SNS 허위정보 모니터링과 같은 실제 서비스 환경에 즉시 적용할 수 있는 실질적 기술적 타당성을 제공한다. 또한 생성형 AI 환경에 최적화된 고품질 데이터셋을 공개함으로써, 국내외 연구 생태계가 재현성과 비교 가능성을 확보한 상태에서 탐지 모델을 개발·확장할 수 있도록 지원한다. 이러한 기여는 궁극적으로 한국어 기반 허위정보 대응 체계 전반의 신뢰성과 효율성을 향상시키는 데 기여할 것으로 기대된다.

그럼에도 불구하고 본 연구는 데이터셋이 정치 분야로 한정되어 있어 모델의 일반화 능력을 충분히 평가하기 어렵다는 한계를 갖는다. 또한 실제 미디어 환경에서 중요하게 작용하는 출처 신뢰도, 전파 경로, 사회적 맥락과 같은 비텍스트 정보가 분석에 포함되지 못했다는 점 역시 개선 과제로 남는다. 향후 연구에서는 데이터셋 범위를 경제·보건 등 다른 사회 핵심 분야로 확장하고, 이유 기반 프롬프트와 설명 가능 인공지능

기법을 통합하여 맥락적 요소를 함께 고려하는 다차원적 탐지 시스템으로 고도화할 필요가 있다. 아울러 제안된 모델이 언론사의 팩트체킹 자동화나 대규모 포털 플랫폼의 댓글 검수 체계 등 실제 서비스 환경에서 안정적 성능을 발휘할 수 있는지 장기간 운영을 통한 실증적 검증이 필요하다. GitHub를 통한 데이터셋 공개는 이러한 후속 연구를 촉진하고, 한국어 기반 허위정보 탐지 생태계의 발전을 가속하는 중요한 토대가 될 것이다.

참고문헌(References)

[국내 문헌]

강병희. 신승협. (2024). 인공지능 기계학습을 이용한 선거 뉴스 프레임 분류: 사회과 선거 교육 방안 제안. *시민교육연구. 56*(3). 209-238.

강예나. 송민채. 신경식. (2025). Aspect-Based Sentiment Classification Using KLUE-BERT with Advanced Embeddings and Graph Convolutional Networks. *지능정보연구. 31*(1). 293-307. http://dx.doi.org/10.13088/jiis.2025.31.1.293

강주영, 송민. (2024). 한국어 가짜 구매후기 생성과 탐지 성능 평가. *지능정보연구, 30*(2), 313-328. http://dx.doi.org/10.13088/jiis.2024.30.2.313

고상훈. 안현철. (2024). 대규모 언어 모델을 활용한 한국어 가짜뉴스 탐지: 한계와 가능성. *지식경영연구. 25*(4). 113 - 127. https://doi.org/10.15813/kmr.2024.25.4.006

고인석. (2025). 가짜뉴스 개념의 세 차원. *철학논총. 120*. 27-51. http://dx.doi.org/10.20433/jnkpa.2025.04.27

김지혁. 안현철. (2023). 품사별 출현 빈도를 활용한 코로나19 관련 한국어 가짜뉴스 탐지. *지능정보연구. 29*(2). 267-283. https://doi.org/10.13088/jiis.2023.29.2.267

김혜윤. 노연수. 박종혁. 박민정. 양병욱. 정윤서. (2025). 임베딩 모델 및 Advanced RAG 기법을 활용한 한국어 텍스트 분류. *응용통계연구. 38*(4). 571-588. https://doi.org/10.5351/KJAS.2025.38.4.571

박성수. 이건창. (2019). 효과적인 가짜 뉴스 탐지를 위한 텍스트 분석과 네트워크 임베딩 방법의 비교 연구. *디지털융복합연구. 17*(5). 137 - 143. https://doi.org/10.14400/JDC.2019.17.5.137

박준성. (2024). Detection of fake reviews in the era of generative AI : a novel approach to leveraging aspect-based sentiment analysis inconsistency. 연세대학교 공학대학원.

심재승, 이재준, 정이태, 안현철. (2020-07-16). 워드 임베딩을 활용한 한국어 가짜뉴스 탐지 모델에 관한 연구. 한국컴퓨터정보학회 학술발표논문집. 제주.

윤호영. 안도현. (2023). 자연어 생성기반 뉴스 보도 패턴 일반화 및 뉴스 구성에 따른 분류 가능성: 소규모 LSTM 생성 데이터를 통한 내용 및 표현 형식 기반 뉴스 유형화 원리 고찰. *커뮤니케이션 이론. 19*(1). 84-123. https://doi.org/10.20879/ct.2023.19.1.084

이태원. 박지수. 손진곤. (2023). CNN 기반 감성 변화 패턴을 이용한 가짜뉴스 탐지. *정보처리학회논문지: 소프트웨어 및 데이터공학. 12*(4). 179 - 188. https://doi.org/10.3745/KTSDE.2023.12.4.179

이충열. (2018). 텍스트마이닝을 활용한 가짜뉴스 분류기의 성능 고찰. 경북대학교 대학원.

임동훈. 김건우. 최근호. (2021). 텍스트 마이닝과 딥러닝 알고리즘을 이용한 가짜 뉴스 탐지 모델 개발. *경영정보학연구. 23*(4), 127-146. http://dx.doi.org/10.14329/isr.2021.23.4.127

좌희정. 임희석. 오동석. (2019). 자동화기반의 가짜 뉴스 탐지를 위한 연구 분석. *한국융합학회논문지. 10*(7). 15-21. https://doi.org/10.15207/JKCS.2019.10.7.015

정동인. (2023). 다중 가짜뉴스 탐지를 위한 데이터 관계 모델링 기법. 중앙대학교 대학원

정원영. (2024). 인공지능 기반 한국어 온라인 뉴스 가짜 댓글 생성 및 탐지. 서강대학교 메타버스전문대학원.

정이태. 안현철. (2022). 그래프 임베딩을 활용한 코로나19 가짜뉴스 탐지 연구 - 사회적 참여 네트워크의 이용 여부에 따른 탐지 성능 비교. *지능정보연구. 28*(1). 197-216. http://dx.doi.org/10.13088/jiis.2022.28.1.197

조예현. 하예린. 임양미. (2025). KPF-BERT 기반 가짜뉴스 탐지 시스템. JOURNAL OF BROADCAST ENGINEERING. 30(5). 706-714. https://doi.org/10.5909/JBE.2025.30.5.706

조하진. 김경호. (2019). 19대 대통령 선거의 SNS 가짜뉴스(fakenews) 네트워크 분석. *디지털콘텐츠학회논문지. 20*(8). 1553-1565. http://dx.doi.org/10.9728/dcs.2019.20.8.1553

현윤진. 김남규. (2018). 뉴스와 소셜 데이터를 활용한 텍스트 기반 가짜 뉴스 탐지 방법론. *한국전자거래학회지. 23*(4). 19-39. https://doi.org/10.7838/jsebs.2018.23.4.019

한윤진. 김근형. (2021). A Study on Automated Fake News Detection Using Verification Articles. *정보처리학회논문지: 소프트웨어 및 데이터 공학 10*(12). 569 - 578. https://doi.org/10.3745/KTSDE.2021.10.12.569

현준서. 유서현. 조재혁. (2024). Bayesian Optimization-HyperBand 를 적용한 효과적인가짜 뉴스 탐지모델의 성능 평가와 분석. *아이씨티플랫폼학회. 12*(4). 55-66

[국외 문헌]

Loth, A., Kappes, M., & Pahl, M. O. (2024). Blessing or curse? A survey on the Impact of Generative AI on Fake News. arXiv preprint arXiv:2404.03021. https://doi.org/10.48550/arXiv.2404.03021

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. arXiv preprint arXiv:1405.4053. https://doi.org/10.48550/arXiv.1405.4053

Park, S., Moon, J., Kim, S., Cho, W., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I. V., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S., Park, L., Oh, A. H., Ha, J., & Cho, K. (2021). KLUE: Korean Language Understanding Evaluation. arXiv preprint arXiv:2105.09680. https://doi.org/10.48550/arXiv.2105.09680

Shim, J. S., Lee, Y., & Ahn, H. (2021). A link2vec-based fake news detection model using web search results. *Expert Systems with Applications*, *184*, 115491.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22-36.

Abstract

A Study on Constructing a Korean Fake News Dataset Using Generative AI and Comparing Detection Models

Sangmin Lee* · Yoojin Lee* · Woojin Jang* · Hyunchul Ahn**

The advancement of generative AI and large language models (LLMs) has enabled the mass production of text nearly indistinguishable from human-written articles. This has led to the proliferation of fake news, particularly intentional disinformation—particularly in the political sphere—intentionally damaging the image of election candidates or specific groups, posing a serious threat to the reliability of information in Korean society. Fake news can be categorized into disinformation and misinformation depending on intent, and this study specifically focuses on disinformation, the intentional subtype with substantial social impact. Previous fake news detection research has been criticized for its limitations: it struggles to determine authenticity based solely on text content when faced with sophisticated disinformation, and it fails to sufficiently leverage external contexts such as social context and dissemination patterns. Particularly in the Korean language environment, the scale and diversity of verified datasets lag significantly behind English-speaking regions. Furthermore, since many LLMs are pre-trained on English-centric corpora, performance degradation and linguistic bias occur when applied to Korean text. Additionally, bias in content quality aspects like logical structure or grammaticality has been observed in LLM-based detection studies using actual Korean news datasets. To address this research environment, this study selects Korean political news data, then uses generative AI to create Korean political fake news texts that distort interpretation and context while preserving the kernel of truth from existing news. It directly constructs a new Korean fake news dataset composed of real-news versus fake-news pairs. Furthermore, based on this dataset, we empirically analyze the potential and limitations of detection models utilizing Korean-specific text embeddings and pre-trained language models. This aims to present the necessity and direction for a Korean-specific detection framework capable of countering the increasingly sophisticated Korean political fake news emerging in the generative AI era.

Key Words : Fake news detection, generative AI, Korean fake news, Korean pre-trained language model, text embedding

Received : November 25, 2025 Revised : December 14, 2025 Accepted : December 16, 2025
Corresponding Author : Hyunchul Ahn

* School of Management Information Systems, Kookmin University
** Corresponding Author: Hyunchul Ahn
School of Management Information Systems, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 02707, Korea
Tel: +82-2-910-4577, Fax: +82-2-910-4017, E-mail: hcahn@kookmin.ac.kr

저 자 소 개

이 상 민

현재 국민대학교 경영정보학부에 재학 중이다. 주요 관심 분야는 AI 기반 사회문제 해결과 데이터 기반 의사결정 모델 설계이며, 이를 실제 환경에 적용하기 위한 AI 기반 서비스 및 제품 기획에 대한 연구 및 프로젝트를 수행하고 있다.

이 유 진

현재 국민대학교 경영정보학부에 재학 중이다. 주요 관심 분야는 데이터 분석과 인공지능 기반 서비스이며, 데이터 기반 의사결정을 지원하는 모델과 실제 서비스 환경에서의 AI 활용에 관심을 두고 있다.

장 우 진

현재 국민대학교 경영정보학부에 재학 중이다. 주요 관심 분야는 AI 기반 데이터 분석과 LLM 기반 자연어 처리이며, 이를 실제 경영 및 비즈니스에 적용하여 인사이트를 도출하는 것에 관심이 크다.

안 현 철

현재 국민대학교 경영정보학부 교수로 재직 중이다. KAIST에서 산업경영학사를 취득하고, KAIST 테크노경영대학원에서 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 주요 관심 분야는 금융 및 고객관계관리 분야의 인공지능 응용, 지능형 의사결정지원시스템, 정보시스템 수용과 관련한 행동 모형 등이다.