Check for updates
특집논문 (Special Paper)
방송공학회논문지 제30권 제5호, 2025년 9월 (JBE Vol.30, No.5, September 2025)
https://doi.org/10.5909/JBE.2025.30.5.706
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

KPF-BERT 기반 가짜뉴스 탐지 시스템

조 예 현[a)], 하 예 린[a)], 임 양 미[a)‡]

KPF-BERT Based Fake News Detection System

Yehyun Cho[a)], Yelin Ha[a)], and Yangmi Lim[a)‡]

요 약

최근 가짜뉴스(Fake News)가 심각한 사회적 문제로 대두되며 가짜뉴스에 대한 법적 정의와 규율이 시도되고 있다. 가짜뉴스 판별을 위해 사람이 직접하고 있는 경우도 있지만, 최근에는 인공지능 기술의 성장으로 인공지능 기술을 활용한 탐지기법들이 발전하고 있다. 국내의 가짜뉴스 탐지 기술은 한국어 기반의 데이터셋 구축의 문제로 국외보다 다소 발전이 느린 편이다. 본 연구에서 한국어 기반의 KPF-BERT 모델을 사용하여 가짜뉴스의 유형 탐지 기능을 포함하였다. 낚시성, 광고성, 자극성 기사를 선정한 이유는 인터넷 위주로의 정보 소비 환경의 변화를 반영하기 위함과 인터넷 뉴스 소비자의 체감 신뢰도와 정보 오인의 가능성이 높은 기사들을 판별하기 위함이다. 이들 유형은 자동화된 알고리즘을 통해 비교적 명확한 언어적·형식적 패턴을 기반으로 탐지할 수 있어, 기술적 접근 가능성과 실질적 사회적 파급력 모두를 고려했을 때 우선적으로 다룰 필요성이 있다. 따라서 본 연구는 낚시성, 광고성, 자극성 기사 유형에 대한 탐지 모델을 구축함으로써, 가짜뉴스의 주요 유형 중 일부에 대한 실효적 대응 방안을 제시하고자 한다.

Abstract

Fake news has recently emerged as a serious social problem, prompting attempts to establish legal definitions and regulations for fake news. While human detection of fake news is sometimes performed manually, the recent advancement of artificial intelligence (AI) has led to the development of AI-based detection techniques. Domestic fake news detection technology has lagged behind internationally due to the difficulty of establishing a Korean-language dataset. In this study, the Korean-based KPF-BERT model was used to detect fake news types. The analysis of clickbait, advertorial, and sensational news focused on reflecting the changing internet-centric information consumption environment and identifying articles with a high perceived trustworthiness and potential for misinformation among online news consumers. These types of articles can be detected using automated algorithms based on relatively clear linguistic and formal patterns. Therefore, given their technological accessibility and potential societal impact, they deserve priority attention. Therefore, this study aims to propose effective countermeasures against some of the major types of fake news by building a detection model for clickbait, advertorial, and sensational articles.

Keyword : Fake news, Clickbait, KPF-BERT model, Deep learning, Detection models

a) 덕성여자대학교 IT미디어공학과(Department of IT Media Engineering, Duksung Women's University)
‡ Corresponding Author : 임양미(Yangmi Lim)
E-mail: yosimi@duksung.ac.kr
Tel: +82-2-901-8350
ORCID: https://orcid.org/0000-0002-3725-0025
· Manuscript July 21, 2025; Revised September 8, 2025; Accepted September 8, 2025.

Ⅰ. 서 론

최근 디지털 플랫폼과 소셜미디어의 급속한 확산은 정보 유통 방식에 큰 변화를 가져왔으며, 이에 따라 가짜뉴스(Fake News)가 심각한 사회적 문제로 대두되고 있다. 이러한 현상은 2016년 미국 대통령 선거 기간 동안 더욱 두드러졌는데, 당시 딥페이크(deepfake) 기술 기반의 가짜뉴스가 주요 언론 보도보다 인터넷 기반의 SNS를 타고 더 많은 사용자 반응을 얻으며 사회적 논쟁의 중심에 놓이게 되었다. 따라서 뉴스기사에 대한 팩트체크가 중요하게 되었고 이에 따른 가짜뉴스의 법적 정의와 규율이 시도되고 있으며[1], 가짜뉴스 판별을 사람이 직접하고 있는 경우도 있지만, 인공지능 기술을 활용한 탐지기법들이 발전하고 있다[2].

인공지능 기반의 가짜뉴스 탐지기법은 데이터를 학습시켜 가짜 여부를 판별한다. 국외의 경우, 가짜뉴스 탐지는 이미 고도화된 기술 모델들이 많이 발표되었다. Bourgonje et al.은 기사 제목과 본문 간의 문장을 분석해 클릭베이트(Clickbait, 낚시성 기사) 탐지를 수행하였고, 약 89.6% 정확도를 보고하였다[3]. 이를 한국어에 적용하기에는 영어에 비해 형태소 분류와 불규칙한 변화, 문장 내에 포함된 단어 수 부족 등으로 한국어 기사에 최적화된 자연어 처리 모델 구축에 부족한 한계가 있어 국내의 가짜뉴스 탐지 기술은 국외보다 다소 발전이 느린 편이다. 또한 클릭베이트 탐지 모델은 선정적, 기만적 제목을 판별하는 것에 중점을 갖고 있다. 이것은 가짜뉴스가 사용자에게 클릭을 유도하기 위해 제목에 사용자를 자극하는 선정성과, 사용자를 속이는 기만적 내용을 주로 담고 있기 때문이다[4]. 한국언론진흥재단(2018)은 온라인 뉴스 환경에서 가짜뉴스로 인식되는 유형들에 대해 '뉴스 형식을 사용한 거짓 정보', '찌라시 정보', '언론사 오보', '선정적 제목 등을 통해 흥미를 끄는 낚시성 뉴스', '광고성 뉴스', '편파적 뉴스', '댓글성 정보' 등으로 분류하여 설문조사를 실시한 바 있다[5]. 일반 사용자들은 특별한 구분 없이 그림 1에서 분류한 형식 대부분을 가짜뉴스로 인식하고 있다.

본 연구에서 그림 1에서와 같은 가짜뉴스의 유형별 탐지 전체를 할 수 없지만, 클릭베이트 탐지 모델에서 사용되는 기만적 내용을 담고 있는 뉴스 데이터 외에 광고성 뉴스, 선정적 내용을 가진 자극성 뉴스 데이터를 확보하여 낚시성, 광고성, 자극성 뉴스를 탐지하였다. 이는 인터넷 위주로의 정보 소비 환경의 변화를 반영하고 인터넷 뉴스 소비자의 체감 신뢰도와 정보 오인의 가능성이 높은 기사들을 탐지하여 상세 정보를 제공하기 위함이다. 이 세 유형은 특히 디지털 뉴스 환경에서 상업적 의도, 감정적 반응 유발을 목적으로 제작되는 경우가 많아 정보의 객관성과 공공성 훼손 가능성이 크다고 보았다. 또한, 이들 유형은 자동화된

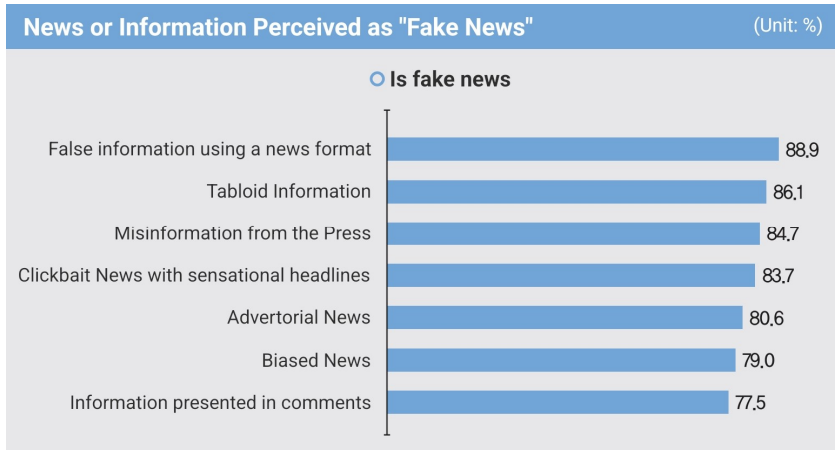

그림 1. 한국언론진흥재단 미디어연구센터 온라인 설문조사 (2018년 3월 26~27일, n=1,060)
Fig. 1. Online survey by the Korea Press Foundation Media Research Center (March 26-27, 2018, n=1,060)

알고리즘을 통해 비교적 명확한 언어적·형식적 패턴을 기반으로 탐지할 수 있어, 기술적 접근 가능성과 실질적, 사회적 파급력 모두를 고려했을 때 우선적으로 다룰 필요성이 있다. 따라서 본 연구는 낚시성, 광고성, 자극성 기사 유형에 대한 탐지 모델을 구축함으로써, 가짜뉴스의 주요 유형 중 일부에 대한 실효적 대응 방안을 제시하고자 한다.

본 연구는 2장에서 딥러닝 기반의 탐지기법 모델에 대한 선행연구를 설명하고, 3장에서는 온라인 뉴스 기사의 낚시성, 광고성, 자극성을 판별하는 탐지 시스템과 모델 제작 과정을 설명한다. 4장에서는 성능 평가를 설명하고, 5장에서 크롬 확장 프로그램, Django 기반 서버 시스템의 실행 과정 및 결과를 설명한다. 6장에서 결론을 설명하고 마무리한다.

Ⅱ. 관련 연구

기존 텍스트 마이닝 기반 가짜뉴스 판별 모델은 주로 TF-IDF, Word2Vec, LSTM 등의 통계적 혹은 RNN 기반 모델을 이용하여 뉴스 제목 또는 본문에서 특징 벡터를 추출하고, 이를 기반으로 분류기를 통해 판별하는 방식이 사용되었다[6]. 하지만 이러한 모델은 문맥 정보를 충분히 반영하지 못하거나, 한 방향으로만 정보를 처리하는 구조적 한계가 있어 문장 내 의미 파악에 어려움이 있다. Devlin et al.은 BERT와 같은 딥러닝 기반의 방법론들을 활용하여 분석 및 탐지 모델을 제안하였다[7]. BERT는 가짜뉴스 탐지 모델 자체가 아니라, 다양한 자연어 처리(NLP) 작업에 활용될 수 있는 사전 학습된 언어 모델이다. 구글이 2018년에 발표한 'BERT(Bidirectional Encoder Representations from Transformers)'는 앞의 단어들을 참조해 다음에 나오는 단어를 예측하는 방식으로 기존의 단방향 언어 모델과 달리 문장에서 예측해야 할 단어 이후의 단어들까지 양방향으로 참조해 의미를 더욱 잘 이해하는 방식으로 학습된 자연어 처리 모델이다. 특히, 단어나 문장 간의 관계성 및 의도 파악이 중요한 뉴스 기사 분석에서는 양방향으로 이해하는 기법이 주요 기술로 자리 잡는다. 국내에서는 Joung 연구에서 콘텐츠 기반 변수(특징) 추출 기법을 사용한 기계 학습(Machine Learning) 분류 모델을 사용하여 SNS를 통해 확

표 1. 기존 연구들과 제안하는 모델에 대한 데이터 타입과 한계점 설명
Table 1. Description of data types and limitations of existing studies and the proposed model

Researcher (Year)	Model / Technique Used	Data Type	Key Features / Contributions	Limitations
Devlin et al. (2018)	BERT	Large amounts of unlabeled plain text data	Introduced bidirectional language model for better context understanding	The model size is large, so it requires a lot of computation and time for training and inference
Joung (2019)	Machine Learning Classification Model Using Content-Based Feature Extraction Techniques	Article body text	Text analysis utilizing sentiment analysis and syntactic features is possible for detecting fake news, and detection is possible using only text without any auxiliary data	Insufficient context understanding; one-directional processing limits semantic analysis
Yoon et al. (2021)	Title + Article input-based model	News headlines and full text	Proposed a classification model using both headline and body	Limited use of pretrained language models
Kumari & Singh (2024)	Multimodal deep learning framework	text, image	Analyze text and images together	Additional computational resources are required for image analysis
Beseiso & Al-Zahrani (2025)	Ensemble technique combining LSTM and CNN	Article body text	High prediction accuracy by understanding contextual meaning through LSTM and extracting important patterns from text through CNN	The model is complex, so the training time is long, and the internal mechanism is difficult to interpret due to the nature of the ensemble model
Proposed model	KPF-BERT, Title + Article input-based model	News headlines and full text	Instead of judging news articles as "real" or "fake," it's possible to classify news by type (clickbait, advertorial, sensational). Each of the three types is comprised of a binary classification model, making it easy to add new types	Problems with securing datasets when adding new type models

산되는 가짜뉴스 탐지를 위해 기사 본문 내용 자체의 특징을 추출하여 분석하였으나, 텍스트 분석에만 초점이 맞춰져 있어 탐지성능에 다소 한계점을 보였다[8]. Yoon et al.은 뉴스의 제목과 본문을 입력데이터로 사용하여 가짜뉴스 여부를 판별하는 모델을 제안하였으며[9], 이들은 사용자의 주장에 대한 팩트체크 기사들을 검색하고, 주장의 내용과 검증 기사 간의 유사성을 분석하여 가짜뉴스 여부를 판단하였다. 최근 Kumari & Singh은 텍스트와 이미지 데이터를 모두 활용하는 멀티모달 딥러닝 프레임워크를 제안하여 텍스트와 이미지 데이터를 동시에 사용하였다[10]. 따라서 기사 내용과 이미지가 서로 불일치하는 가짜뉴스 패턴을 효율적으로 탐지가 가능하다. Beseiso & Al-Zahrani의 연구에서는 콘텍스트 강화 모델로 LSTM(Long-Short-Term-Memory)과 CNN(Convolutional Neural Networks)을 결합한 앙상블(Ensemble) 기법을 사용하여 기존의 텍스트 기반 탐지 모델 대비 예측정확도가 크게 향상되었다[11]. 표 1은 기존 연구들과 제안하는 모델에 대한 데이터 타입과 한계점들에 대해 정리한 것이다.

Joung et al.과 Yoon et al.의 연구를 제외하고는 국외 연구로 한국어 기반의 연구에 직접 대입하기에는 한계가 있다. Han et al.은 한국어 데이터셋 KoBERT 기반 가짜뉴스 탐지 모델을 제안하였다[12]. KoBERT(Korean BERT)는 SKT T-Brain이 2019년 발표한 것으로 Wikipedia와 한국어 기사 등에서 수집한 한국어 문장 데이터를 사전 학습시켰고, 데이터 기반 토큰화(Tokenization) 기법을 사용하여 한국어의 불규칙 특성을 학습시켰다[13]. 이후 한국어 문장 독해 능력이 향상된 'KPF-BERT'도 2023년 발표되었다. KPF-BERT는 구글의 BERT 모델을 한국언론진흥재단이 보유한 빅카인즈 기사 데이터(2000년~2021년, 약 4,000만 건)를 활용해 학습시킨 결과물이다. 이러한 언론 도메인 특화 학습을 통해 KPF-BERT는 일반 목적의 언어 모델과 달리 뉴스 기사 특유의 문체와 용어를 효과적으로 이해하고 분석할 수 있다. 따라서 해당 모델이 가짜뉴스 기사 판별에 가장 적합하다고 판단하였다. 표 2는 한국언론진흥재단이 공개한 'KPF-BERT'와 기존 한국어 학습기반 BERT의 비교 평가 결과이다[14]. 가짜뉴스의 낚시성, 광고성, 자극성을 분류하여 판별한 이유는 가짜뉴스의 주요 유형에 대응하기 위한 실제 방안을 모색하고자 함이다. 이러한 접근은 가짜뉴스의 허위 정보 전달을 넘어, 사용자의 클릭을 유도하는 특정한 목적이 이윤 창출을 위한 것인지 여론조작을 위한 것인지 등의 판별을 차후에 할 수 있기 때문이다.

Ⅲ. 가짜뉴스 탐지 모델

1. 가짜뉴스 탐지 시스템

본 연구에서는 KPF-BERT 기반 딥러닝 모델을 활용하여 온라인 뉴스 기사의 낚시성, 광고성, 자극성을 판별하는 프로그램을 개발하였다. 시스템 제작에 앞서, 모델 학습을 위한 데이터 수집 및 전처리 과정을 수행하였고, 이를 바탕으로 낚시성, 광고성, 자극성 기사 탐지를 위한 각각의 이진 분류 모델을 구축하였다. 이후 완성된 모델을 기반으로 크롬 확장 프로그램과 연동되는 탐지 시스템을 구현하였다. 전체 시스템 구성은 뉴스 기사 URL 추출, 기사 본문 크롤링(crawling) 및 전처리, 모델 분석, 판별 결과를 크롬 확장 프로그램에서 화면에 보여주는 순서로 진행하였다. 그림 2는 판별할 뉴스 기사 데이터의 input/output의 흐름을 설명

표 2. KPF-BERT와 기존 BERT와 비교 평가 결과
Table 2. Comparative evaluation results between KPF-BERT and existing BERT

Category	NSMC Movie Review Sentiment Analysis	KLUE-NLI Natural Language Inference	KLUE-STS Sentence Semantic Similarity	KorQuAD v1 Machine Reading Comprehension
Evaluation Metric	Accuracy	Accuracy	Pearson Correlation	Accuracy / F1
KPF BERT	91.29%	87.67%	92.95%	94.95%
KorBERT(ETRI)	90.46%	80.56%	89.52%	82.00%
KoBERT(SKT)	89.92%	79.53%	86.17%	71.36%
BERT base multilingual	87.33%	73.30%	85.66%	90.02%

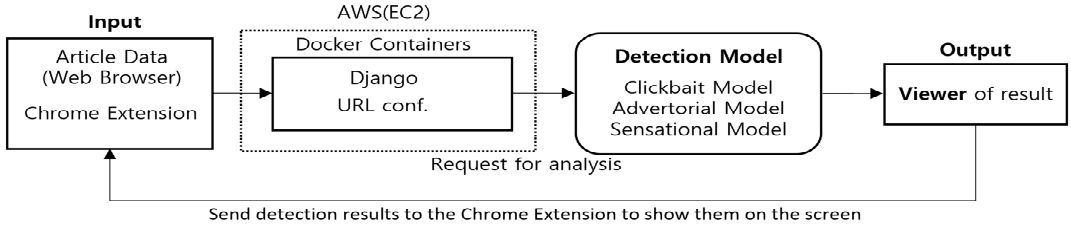

그림 2. 가짜뉴스 탐지 시스템의 데이터 흐름
Fig. 2. Data Flow in Fake News Detection System

한 시스템 구성도이다.

2. 데이터셋 구축

낚시성 기사 탐지 모델의 경우, 한국지능정보사회진흥원(NIA)이 운영하는 AI Hub에서 공개한 '낚시성 기사 탐지 데이터셋'을 학습, 검증, 테스트로 나누어 활용하였다[15]. 해당 데이터는 자연어처리(NLP) 및 딥러닝 기반 탐지 모델 학습을 목적으로 구성된 인공지능 학습용 데이터로, '제목과 본문의 불일치 기사'와 '본문의 도메인 일관성 부족 기사'로 구분되어 있다. 이는 낚시성 기사의 대표적 특징을 구조적으로 반영하고 있어, 본 연구의 낚시성 기사 탐지 모델 구축에 적합하다고 판단하였다.

광고성 및 자극성 기사 탐지 모델의 경우, 별도의 데이터셋을 구축하였다. 이 데이터셋은 광고성, 자극성 기사와 정상 기사로 구성되어 있다. 데이터는 2022년부터 2025년까지 인터넷신문윤리위원회에서 발행한 기사심의결정문(이하 '심의문') PDF에서 수집하였으며, 해당 문서에는 각 심의 대상 기사에 대한 URL과 함께 어떤 조항을 위반했는지가 명시되어 있다[16]. 본 연구에서는 이 조항 정보를 기준으로 다음과 같이 기사 유형을 분류하였다. 광고성 기사는 「인터넷신문 기사심의규정」 중 상업적 상품 및 서비스 홍보, 과장된 정보 등을 다룬 조항에 해당하는 경우로 정의하였다. 구체적으로는 제17조(기사와 광고의 분리), 제17조 1(이용자 보호)을 위반한 기사들을 광고성 기사로 분류하였다. 자극성 기사는 「인터넷신문 기사심의규정」 중 선정적·자극적 표현, 차별적 표현, 혐오 표현 등을 다룬 조항에 해당하는 경우로 정의하였다. 분류 기준은 제5조(선정보도의 지양), 제6조(제목의 원칙), 제11조(차별적 표현 금지), 제13조(범죄보도), 제13조 1(자살보도)을 위반한 기사들을 자극성 기사로 분류하였다[17]. 표 3은 낚시성, 광고성, 자극성 기사 탐지 모델의 정의를 정리한 표이다.

광고성, 자극성 기사를 수집하기 위해 newspaper3k 라이브러리를 사용하여 심의문에 있는 URL에서 기사의 제목과 본문을 크롤링하였다. 정상 기사는 Beautifulsoup 라이브러리를 사용하여 네이버 뉴스의 6개 카테고리(정치, 경제, 사회, 생활/문화, 세계, IT/과학)에서 기사의 제목과 본문을 크롤링하였다.

전처리 과정에서 심의문에 있는 광고성 및 자극성 기사 10,329건 중 약 22.17%에 해당하는 2,289건의 기사가 크롤링에 실패하였는데, 주요 원인으로는 기사가 삭제된 경우가 가장 많았고, 이외에도 인증서 오류, 서버 내부 오류, 네트워크 타임아웃 등이 있었다. 이로 인해 본문이 비어 있거나 추출을 실패한 경우는 학습 데이터셋에서 제외하고, 나머지 유효한 데이터만을 JSON 형식으로 저장하였다. 네이

표 3. 낚시성, 광고성, 자극성 기사 탐지 모델 정의
Table 3. Definitions of Clickbait, Advertorial, Sensational Model

Model	Definition	Dataset Source
Clickbait Model	Determines whether there is a discrepancy between the headline and the body content	Provided by AIHub
Advertorial Model	Detects advertorial and commercially motivated articles	Extracted from review decisions of the Internet Newspaper Ethics Committee (PDF)
Sensational Model	Detects sensational elements such as provocative, violent, or discriminatory expressions	

표 4. 학습 환경 및 하이퍼파라미터
Table 4. Learning environment and hyperparameters

Item	Clickbait Model	Advertorial Model and Sensational Model
Pretrained Model	KPF-BERT	
Tokenizer	Dedicated KPF-BERT Tokenizer	
Epoch	3	
Batch Size	16	
Learning Rate	1e-5	2e-5
Optimizer	AdamW	
Loss Function	CrossEntropyLoss	
Maximum Input Length	512	
Evaluation Metrics	Accuracy, F1 Score	
Model Saving Criterion	Save model with highest validation F1 score	Save model after final epoch
Framework	PyTorch, Hugging Face Transformers	

버 뉴스에서 수집한 정상 기사는 중복 URL 제거, 제목이나 본문 길이가 너무 짧은 기사 제외 등의 전처리를 수행하여 JSON 형식으로 저장하였다. 최종적으로 수집된 데이터는 광고성 기사 4,509건, 자극성 기사 3,531건이고, 정상 기사 데이터는 네이버 뉴스 6개 카테고리에서 각 800건씩, 총 4,800건이다. 광고성 및 자극성 기사는 0으로, 정상 기사는 1로 라벨링하여 두 개의 독립적인 이진 분류 데이터셋(광고성+정상, 자극성+정상)을 구성하였다.

그 후 두 데이터셋을 학습, 검증, 테스트 비율로 7:1.5:1.5로 분할하여 모델 학습에 사용하였다. 모델 입력 단계에서는 BERT 기반 모델의 특성을 반영하기 위해, 제목과 본문은 [제목] [SEP] [본문]의 구조로 입력되도록 하였다. 또한 허깅페이스(Hugging Face, AI스타트업)의 'Tokenizer'를 활용하여 결합한 텍스트를 'WordPiece' 기반 'subword' 단위로 분리하였으며, 입력 길이는 BERT 모델의 최대 길이인 512 Token으로 설정하였다.

3. 모델 구조 및 학습 설정

낚시성, 광고성, 자극성은 각각 독립적인 이진 분류 모델로 구성되며, 모두 한국언론진흥재단에서 공개한 KPF-BERT를 기반으로 구현되었다. 각 모델은 독립적으로 학습되기 때문에 데이터셋 간 규모나 출처 차이가 직접적인 편향 요인으로 작용하지는 않는다. 다만, 낚시성 데이터는 AIHub에서 제공한 대규모 데이터셋을 활용한 반면, 광고성 및 자극성 데이터는 자가 구축한 비교적 소규모 데이터셋이 사용되었기 때문에, 이로 인한 학습 효과 차이가 발생할 수 있음을 인지하고 있다. 이러한 차이를 보완하기 위해 전처리, 라벨 검토, 입력 형식 통일 등 데이터 품질 관리 과정을 통일되게 적용하였다. BERT 모델 자체는 이진 분류를 위한 출력 구조를 포함하고 있지 않기 때문에, 마지막 층에 단일 'Linear Layer'를 추가하여 분류 작업을 수행할 수 있도록 구성하였다. 각 모델의 입력은 [CLS] 제목 [SEP] 본문 [SEP] 형식으로 구성되며, 출력은 'Softmax'를 거쳐 확률값으로 변환된다. 학습에는 'CrossEntropyLoss'를 손실 함수로, 'AdamW'를 최적화 알고리즘으로 사용하였다. 각 모델의 학습 환경 및 하이퍼파라미터는 표 4에 정리되어 있다.

Ⅳ. 성능 평가

본 장에서는 제안한 뉴스 기사 낚시성, 광고성, 자극성 탐지 시스템의 성능 평가 결과를 제시한다. 표 5는 낚시성 기사 탐지 모델의 테스트셋 기준 성능을 나타낸다. 전체 기사 73,344개에 대해 Precision, Recall, F1-score를 측정하였으며, 낚시성/비낚시성 클래스 간 균형 잡힌 분류 성능을 보였다. 정확도는 88.59%이며, Marco 평균과 Weighted 평균 역시 88.58%~88.59% 수준으로 전반적인 안정성을 확인할 수 있다. 표 6은 광고성 기사 탐지 모델로 Precision 99%, Recall 100%로 전체 정확도는 99%로 매우 안정적인 성능을 보였다. 표 7은 자극성 기사 탐지 모델의 테스트셋 성능을 나타낸다. 자극성 기사 탐지 모델 역시 두 클래스 간 균형 잡힌 결과를 나타냈다. 그러나 높은 정확도는 데이터 유출 또는 과적합의 가능성이 있어서, 이를 고려하여 학

습, 검증, 테스트 데이터셋 간의 데이터 중복이나 데이터 편향 등을 검사하였으나 문제는 발견되지 않았다. 높은 정확도는 두 클래스 간 구분이 명확하여 모델이 쉽게 패턴을 학습한 것으로 예측한다. 표 8은 세 가지 유형 정보 탐지 모델의 테스트셋 기준 성능 평가 결과이다.

표 5. 낚시성 기사 탐지 모델의 테스트셋 기준 성능 평가
Table 5. Performance Evaluation of the Clickbait Model Based on the Test Set

Category	Precision	Recall	F1-score	Support
Clickbait	88.88%	88.44%	88.66%	37,006
Non-clickbait	88.28%	88.74%	88.51%	36,338
Accuracy	-	-	88.59%	73,344
macro Average	88.58%	88.59%	88.58%	73,344
weighted Average	88.59%	88.59%	88.59%	73,344

표 6. 광고성 기사 탐지 모델의 테스트셋 기준 성능 평가
Table 6. Performance Evaluation of the Advertorial Model Based on the Test Set

Category	Precision	Recall	F1-score	Support
Advertorial	99%	100%	99%	677
Normal	100%	99%	99%	571
Accuracy	-	-	99%	1,248
macro Average	99%	99%	99%	1,248
weighted Average	99%	99%	99%	1,248

표 7. 자극성 기사 탐지 모델의 테스트셋 기준 성능 평가
Table 7. Performance Evaluation of the Sensational Model Based on the Test Set

Category	Precision	Recall	F1-score	Support
Sensational	100%	99%	99%	538
Normal	99%	100%	99%	564
Accuracy	-	-	99%	1,102
macro Average	99%	99%	99%	1,102
weighted Average	99%	99%	99%	1,102

표 8. 낚시성, 광고성, 자극성 기사 탐지모델 테스트셋 기준 성능 평가
Table 8. Performance Evaluation of the Clickbait, Advertorial, Sensational Model Based on the Test Set

model	accuracy	F1-score
Clickbait	89.1%	88.6%
Advertorial	99.0%	99.0%
Sensational	99.0%	99.0%

V. 크롬 확장 프로그램, Django 기반 서버 구성 및 결과

시스템은 크롬 확장 프로그램, Django 기반 서버, 낚시성·광고성·자극성 기사 탐지 모델로 구성되며, 사용자의 요청에 따라 뉴스를 판별하고 결과를 보여준다. 그림 3은 프

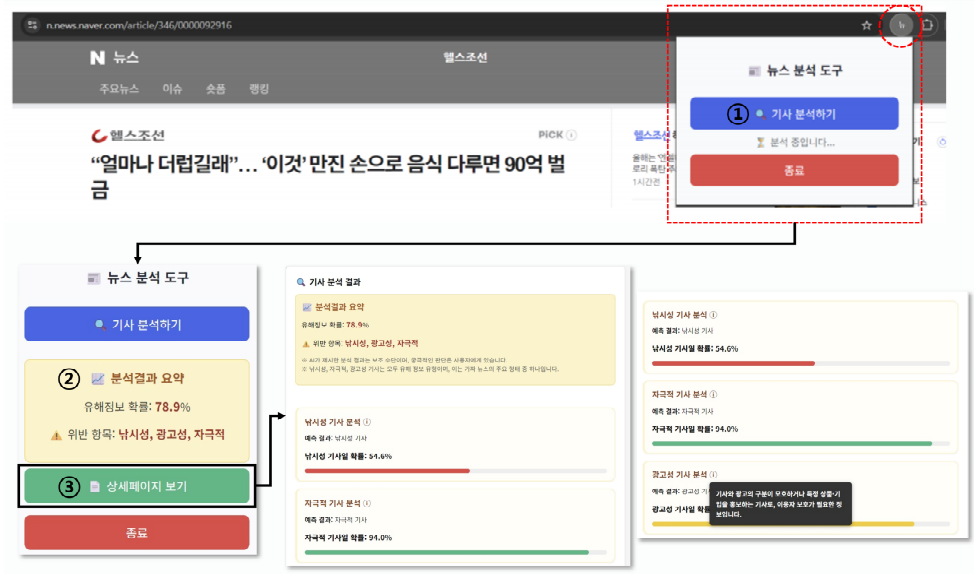

그림 3. 프로그램 실행 결과 화면
Fig. 3. Program execution results screen

로그램 실행 결과 화면이다. 사용자가 크롬브라우저의 상단 오른쪽에 탑재한 확장 프로그램을 실행하기 위해서는 그림 3의 ① '기사 분석하기' 버튼(파란색)을 클릭해야 한다. 클릭 이벤트가 전송되면, 해당 페이지의 URL이 Django REST API 서버로 전송된다. 전송된 URL은 서버 측에서 기사 본문을 크롤링하고 사전 학습된 모델을 통해 뉴스기사 탐지가 수행된다. 결과는 그림 3의 ② 분석결과 요약으로 나타나며, 세 가지 유형별(낚시성, 광고성, 자극성) 확률의 평균값과 각 유형에 대한 위반 항목이 표시된다. 그림 3의 ③ '상세페이지 보기' 버튼(녹색)을 클릭하면, 각 유형별 확률값과 판정 결과를 카드 형태로 제공한다. 프로그램 실행 과정에서의 크롤링 실패를 막기 위해 newspaper3k와 BeautifulSoup 기반의 다중 분기 로직을 적용하였다. 현재까지 개발 환경에서 크롤링 실패는 발생하지 않았으나, 일부 예외 상황에 대해서 오류 메시지를 통해 사용자에게 안내하는 기능을 포함하고 있다.

Ⅵ. 결 론

본 연구에서는 한국어 기반 낚시성, 광고성, 자극성 뉴스기사를 탐지하고 결과를 화면으로 보여주는 시스템을 제안하였다. 낚시성 기사, 광고성 기사, 자극성 기사 탐지를 위한 모델 학습에는 공개 데이터(AIHub) 및 기사심의결정문 기반의 자체 구축 데이터셋을 활용하였으며, 테스트셋 기준 낚시성 기사 탐지 모델은 F1-score 88.6%, 자극성 및 광고성 모델은 각각 99.0%의 높은 정확도를 기록하였다. 특히 낚시성, 광고성, 자극성 기사는 디지털 뉴스 환경에서 클릭 유도, 상업적 의도, 감정 자극 등으로 대표되는 가짜뉴스의 주요 유형으로, 일반 사용자와 플랫폼 운영자 모두에게 높은 주의가 요구되는 정보 형태라는 점에서 우선적인 분류 대상으로 타당성이 있다. 제안된 시스템은 2025년 6월 20일부터 22일까지 뉴스 플랫폼에서 30번의 시뮬레이션 결과, 낚시성은 14건, 광고성은 4건, 자극성은 13건, 정상 12건으로 나타났다. 일부 기사에는 2가지 이상의 유형을 포함하고 있었다. 가장 많이 분포되는 가짜뉴스 유형은 낚시성과 자극성이었다. 자극적인 제목과 내용으로 클릭 수를 높이려는 콘텐츠 제작자의 의도가 맞물린 결과로 예측한다.

본 연구에서 뉴스 사용자가 기사 소비 과정에서 낚시성 제목, 선정적 보도, 기사형 광고 등 가짜 정보를 직관적으로 인식하고 판단할 수 있도록 지원하며, 언론사 및 뉴스 플랫폼의 콘텐츠 필터링 도구로도 활용될 가능성을 지닌다. 향후의 연구에서는 자극성 판단 기준의 주관성으로 발생할 수 있는 언론의 자유 침해, 긴 기사에서 핵심 정보 누락 등의 한계는 여전히 존재하며 이러한 한계를 극복하기 위해 다수 전문가 기반의 자극성 판단 기준 정립과 문단 단위의 핵심 문장을 추출하여 분석에 사용하는 방안 등을 고려할 수 있다. 또한 문장 단위 위험도 평가, GPT 기반 설명 생성 기능의 도입, 응답 속도 개선, 데이터셋 증강 등 외에도 모델 예측의 해석 가능성 향상과 사용자 피드백 기반의 시스템 개선, 미디어 리터러시 교육 도구로의 활용 등 제도적 연계 가능성을 강화할 것이다.

참 고 문 헌 (References)

[1] Y. Kim et al., An Analysis of Fake News Situations and Countermeasures-Focusing on Major OECD Country Cases, report of Korea Communications Commission, KCC-2023-34, December, 2023

[2] S. Kwon, M. Cha, K. Jung, W. Chen and Y. Wang, "Prominent features of rumor propagation in online social media," 2013 IEEE 13th International Conference, pp.1103-1108, Dallas, TX, USA, December 2013.
doi: https://doi.org/10.1109/ICDM.2013.61

[3] P. Bourgonje, J. M. Schneider and G. Rehm, "From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles", Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism, pp.84–89, September, 2017, Copenhagen, Denmark.
doi: https://doi.org/10.18653/v1/W17-4215

[4] Sun Min Lee, "How Do Sensationalism and Deceptiveness of the Headlines Affect Intentions to Use News?", journal of Korean Women's Association for Communication Studies (Media, Gender & Culture), Vol.35 No.3, pp.105-141, September 2020.
doi: https://doi.org/10.38196/mgc.2020.09.35.3.105

[5] J. Kim, Fake news and the role of the media, Press Arbitration Commission, Wn159, pp.14-27, Summer 2021

[6] D. Lim, G. Kim and K. Choi, "Development of a Fake News Detection Model Using Text Mining and Deep Learning Algorithms", Information Systems Review, Vol. 23, No. 4, November 2021.
doi: https://doi.org/10.14329/isr.2021.23.4.127

[7] J. Kim, S. Park, J. Lee, J. Kim and P. Kang, "Development of an

Unsupervised Learning-Based Fake News Generation Framework and Study the Effectiveness of Dataset", Quality Evaluation Metrics, Vol.50, No.1, pp. 36-46, 2024.
doi: https://doi.org/10.7232/JKIIE.2024.50.1.036
[8] H. Joung, Fake News Detectoion Using Content-based Feature Extraction Method, Master Thesis of Ewha Womans University, February, 2019.
[9] S. Yoon, K. Park, J. Shin, H. Lim, S. Won, M. Cha and K. Jung, "Detecting Incongruity Between News Headline and Body Text via a Deep Hierarchical Encoder," In proceedings of the AAAI Conference on Artificial Intelligence Article, No.98, pp. 791-800, January 2019.
doi: https://doi.org/10.1609/aaai.v33i01.3301791
[10] S. Kumari and M. P. Singh, "A Deep Learning Multimodal Framework for Fake News Detection", journal of Engineering, Technology & Applied Science Research, Vol.14, No.5, pp.16527-16533, October 2024.
doi: https://doi.org/10.48084/etasr.8170
[11] M. Beseiso and S. Al-Zahrani, "A Context-Enhanced Model for Fake News Detection", journal of Engineering, Technology & Applied Science Research, Vol.15, No.1, pp.19128-19135, February 2025.
doi: https://doi.org/10.48084/etasr.9192
[12] Y. Han and G. Kim, "A Study on Automated Fake News Detection Using Verification Articles", KIPS Trans. Softw. and Data Eng. Vol.10, No.12 pp.569-578.
doi: https://doi.org/10.3745/KTSDE.2021.10.12.569
[13] KoBERT, https://github.com/SKTBrain/KoBERT, (accessed July. 20, 2025)
[14] KPF-BERT, https://github.com/KPFBERT, (accessed July. 20, 2025)
[15] AI Hub of NIA, https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71338 (accessed Jul. 20, 2025)
[16] Internet Newspaper Ethics Committee, https://inec.or.kr/board/article_result/list (accessed Jul. 20, 2025)
[17] Internet Newspaper Ethics Committee, https://inec.or.kr/article/rules (accessed Jul. 21, 2025)

저 자 소 개

조 예 현

- 현재 : 덕성여자대학교 IT미디어공학전공 학사과정
- ORCID : https://orcid.org/ 0009-0008-8056-7598
- 주관심분야 : 인공지능, 미디어 공학

하 예 린

- 현재 : 덕성여자대학교 IT미디어공학전공 학사과정
- ORCID : https://orcid.org/0009-0003-3919-8050
- 주관심분야 : 인공지능, 미디어 공학

임 양 미

- 1998년 3월 : 큐슈대학교 예술공과대학 정보전달전공 석사
- 2009년 2월 : 중앙대학교 첨단영상대학원 박사
- 2010년 ~ 현재 : 덕성여자대학교 가상현실융합학과 교수
- ORCID : https://orcid.org/0000-0002-3725-0025
- 주관심분야 : 멀티미디어, 인터랙티브아트, 가상현실, 입체영상