JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 2, pp. 357-368, Feb. 2026

Check for updates

경량 텍스트 구조 신호 기반 허위뉴스 탐지

김 현 아 · 조 윤 용*
경기대학교 교양학부 조교수

Fake News Detection Based on Lightweight Text Structural Cues

Hyun-Ah Kim · Yoon-Yong Cho*
Assistant Professor, Department of General Studies, Kyonggi University, Suwon 16227, Korea

[요 약]

본 연구는 의미 해석을 최소화한 텍스트 구조 신호만으로 허위뉴스를 탐지할 수 있는 가능성을 검증하였다. Kaggle True/Fake 뉴스 데이터셋을 사용하여 출처·도메인 꼬리표·저작권 문구를 제거하는 정화(Decontamination) 절차를 거친 후, 단어·문자 n-그램 기반 구조 신호를 입력으로 하는 경량 분류기를 학습하였다. 시계열 분할(q80, q90)에 따른 고정 평가를 통해 데이터 누출과 과대평가를 통제하였으며, 결과는 PR-AUC 0.9768, ROC-AUC 0.9933, Macro-F1 0.9570, Accuracy 0.9785로 나타났다. 상위 10% 검토 구간에서 정밀도 0.995, 재현율 0.659을 달성하여 무오탐에 가까운 탐지 성능을 보였다. 본 연구는 단일 공개 데이터셋(Kaggle True/Fake)에서 출처 표식 제거(정화) 이후에도 구조 신호만으로 높은 분리력이 관측될 수 있음을 확인하고, 해당 설정에서의 보수적 하한 성능을 보고함으로써 고비용 언어모델에 의존하지 않는 경량·비침습형 대응의 가능성을 제시한다.

[Abstract]

This study verifies that fake news can be detected using only structural text cues, without semantic interpretation. Using the Kaggle True/Fake News dataset, decontamination was applied to remove source names, domain marks, and copyright tags for models to learn purely stylistic and syntactic signals. A lightweight word and character n-gram classifier was trained under fixed temporal splits (q80, q90) to prevent data leakage and overfitting. The model achieved a PR-AUC of 0.9768, an ROC-AUC of 0.9933, a macro-F1 of 0.9570, and an accuracy of 0.9785. In the top 10% review range, the precision reached 0.995 and the recall 0.659, with false positives almost eliminated. These findings show that high separability can be observed even after source/boilerplate removal within a single public dataset (Kaggle True/Fake), and a conservative lower-bound performance under fixed temporal splits is reported.

색인어 : 허위뉴스 탐지, 텍스트 구조 신호, 경량 분류 모델, 정화 절차, 시계열 기반 성능 평가
Keyword : Fake News Detection, Text Structural Cue, Lightweight Classification Model, Decontamination Procedure, Temporal Split Evaluation

http://dx.doi.org/10.9728/dcs.2026.27.2.357

Received 12 November 2025; **Revised** 09 December 2025
Accepted 02 January 2026

***Corresponding Author; Yoon-Yong Cho**
Tel: +82-31-249-1472
E-mail: yoonycho@kyonggi.ac.kr

Ⅰ. 서 론

1-1 연구 배경 및 필요성

디지털 뉴스 생태계에서 정보 확산 속도가 가속화되면서 허위뉴스는 사회적 혼란과 신뢰 붕괴를 초래하는 핵심 요인으로 부상하였다. 특히 소셜미디어와 자동화된 유통 구조 속에서 허위정보는 단기간에 대규모로 확산되며, 정치·경제·보건 등 다양한 영역에 영향을 미친다. 이에 따라 최근 연구들은 허위뉴스 탐지가 단순한 사실 검증을 넘어 언어적·구조적 신호를 포착하는 알고리즘적 접근을 요구한다고 지적한다[1],[2]. 그러나 다수의 기존 방법은 대형 사전학습 언어모델에 의존하여 연산비용과 데이터 의존성이 높고, 실시간 적용에는 한계가 있다.

생성형 AI와 대형 언어모델의 확산으로 사람 작성 기사와 유사한 합성 텍스트가 대량 유통되면서 허위정보 대응은 자동화 국면으로 진입하였다. 이로 인해 의미 기반 고비용 모델의 필요성이 강조되었으나, 실제 배치 환경에서는 비용·지연·감사 가능성(auditability) 제약으로 인해 경량 1차 스크리닝 계층의 중요성이 오히려 커지고 있다[3].

현행 허위정보 탐지 연구는 그래프 기반 탐지[4], 텍스트 기반 탐지[5],[6], 그리고 멀티모달 접근으로 구분된다. 그래프 기반 방법은 확산 구조를 활용하지만 계산 비용이 크고, 텍스트 기반 접근은 주제·도메인 편향으로 일반화가 제한된다. 멀티모달 방식은 성능은 높으나 데이터 접근성과 처리 비용 측면에서 실무 적용이 어렵다.

이러한 배경에서 본 연구는 의미 해석에 의존하지 않고 텍스트의 형태적·구조적 신호(structural cues)만으로 허위정보를 식별할 수 있는 가능성을 검증한다. 단어·문자 반복, 구두점 사용, 문장 길이와 같은 문체적 특성은 출처나 주제와 무관하게 비교적 안정적인 분리 신호로 기능할 수 있다.

또한 기존 연구의 무작위 분할(IID) 가정이 시간에 따른 주제·양식 변화를 반영하지 못한다는 점을 고려하여, 본 연구는 기사 게시 시점을 기준으로 한 고정 시계열 분할을 도입한다. 아울러 데이터 정화(Decontamination) 절차를 통해 통신사 명과 도메인 꼬리표 등 출처 식별 신호를 제거함으로써, 모델이 문체·구문 기반 구조 신호만 학습하도록 제한한다.

결과적으로 본 연구는 고비용 언어모델 없이도 경량 구조 신호만으로 실질적인 탐지 성능을 확보할 수 있는지를 검증하고, 시계열 분할과 정화 절차를 통해 실제 배치 단계의 하한 성능을 추정하는 데 목적이 있다. 이는 의미 해석을 최소화한 새로운 탐지 접근을 제시함과 동시에, 저비용·비침습형 허위정보 대응 체계 구축에 실질적 근거를 제공한다.

1-2 연구 목적 및 기여

본 연구의 목적은 뉴스 데이터에서 텍스트 구조 신호에 기반한 허위뉴스 탐지 모델을 설계하고, 이 모델이 의미적 해석이나 복잡한 메타데이터 없이도 일정 수준의 탐지력을 확보할 수 있음을 실증적으로 입증하는 것이다. 이를 위해 Kaggle True/Fake 데이터셋을 사용하여 정화 절차를 선행한 뒤, 단어·문자 n-gram 기반의 벡터 표현을 이용해 경량 분류기를 학습하였다.

연구의 주요 기여는 다음과 같다. 첫째, 출처 식별 편향을 제거한 정화 데이터셋 설계를 통해 모델이 문체·구문적 신호만으로 허위성을 예측하도록 제한함으로써, 기존 딥러닝 기반 탐지 연구의 구조적 한계를 보완했다. 둘째, 시간 기반 고정 분할을 적용하여 모델이 새로운 시점의 뉴스에서도 일관된 분리력을 유지하는지를 평가함으로써, 실제 서비스 환경의 성능 하한을 제시했다. 셋째, 경량 모델 기반 탐지 파이프라인의 효율성을 실증하여, 대규모 언어모델이 아닌 저자원 환경에서도 실시간 탐지가 가능함을 보였다.

결과적으로 본 연구는 “텍스트 구조 신호만으로도 허위정보를 실용적으로 탐지할 수 있다”는 가능성을 실험적으로 검증하고, 허위뉴스 탐지 연구에서 의미적 해석 중심에서 구조적 신호 중심으로의 패러다임 전환을 제안한다.

Ⅱ. 관련연구

2-1 트위터 봇 탐지 연구의 개요

트위터나 페이스북과 같은 소셜미디어에서의 허위정보 확산은 대체로 자동화된 계정(봇, bot)에 의해 증폭된다. 초기 연구들은 인간 사용자와 기계 계정 간의 활동 패턴 차이를 포착하여 비정상 계정을 식별하는 데 초점을 두었다[7]. Ferrara 등은 트윗 간 간격, 리트윗 빈도, 활동 시간대 등의 변동성을 분석해 봇 계정이 인간 사용자보다 주기적이고 기계적인 행동 패턴을 보인다고 보고하였다[8]. 이후 Varol 등은 계정 프로필 특성, 게시 주기, 언어적 다양성, 해시태그 사용 패턴을 종합한 혼합 피처 모델을 제안하여, 봇과 인간 계정의 상호작용 네트워크를 구체적으로 모델링하였다[9]. 이러한 연구들은 허위정보의 유통 매개체로서의 봇 네트워크 구조를 규명하고, 정보 확산의 비정상성을 계량적으로 분석하는 기반을 마련했다는 점에서 중요한 의의가 있다.

최근에는 딥러닝 기반 시계열 모델과 그래프 신경망(GNN)을 이용하여 봇의 자동화 수준, 클러스터링 형태, 주제 중심성 등을 정량적으로 파악하려는 시도도 이루어지고 있다. 예를 들어 Cresci 등은 트윗 내용의 반복성, 시간 동기화 패턴, 계정 간 리트윗 관계를 학습하여 집단적 자동행동(collusive automation)을 탐지하였다. 그러나 이러한 연구들은 여전히 행위 기반 피처에 의존하며, 뉴스 본문이나 게시물의 언어적 구조·문체적 특성을 고려하지 못한다는 한계가 있다. 본 연구는 이러한 기존 접근의 제약을 인식하고, 텍스트 자체의 구조 신호에 초점을 맞추어 허위정보 탐지 가능성을 검증한다.

2-2 메타데이터 기반 탐지 연구

뉴스의 출처, 작성일, URL, 작성자 정보, 게시 플랫폼과 같은 메타데이터는 허위뉴스 판별의 초기 단서로 활용되어 왔다. Shu 등은 뉴스 출처의 신뢰도와 사용자 공유 패턴을 결합하여 허위정보 확률을 추정하는 프레임워크를 제시하였고[10], 이를 통해 "출처 중심 신뢰지표(source credibility index)" 개념을 정립하였다. Pérez-Rosas 등은 기사 작성자, 제목, 주제 메타데이터를 결합한 하이브리드 모델을 제안하였으며[11], 이는 텍스트 내용 분석과 메타 속성 결합이 탐지 성능 향상에 기여함을 보여주었다.

이후 연구들은 SNS 플랫폼의 게시 시간대, URL 도메인 패턴, 링크 길이 등의 비언어적 정보까지 통합하여 확장되었다. 특히 Zhou 등은 게시 시점의 밀집도와 공유 확산 속도를 이용한 시계열 메타 분석으로, 허위정보가 진짜 뉴스보다 단기간에 집중적으로 퍼지는 경향을 확인하였다. 그러나 이러한 접근은 출처 정보가 은폐되거나 조작될 경우 즉각적으로 성능이 저하되며, 출처 식별 자체가 학습 과정에 편향을 유발할 수 있다는 문제를 가진다. 따라서 본 연구는 이러한 메타데이터 의존성을 제거하고, 출처 신호를 완전히 배제한 상태에서 텍스트 구조만을 근거로 탐지를 수행함으로써 메타 정보에 따른 과대평가 문제를 방지한다.

2-3 텍스트 및 그래프 기반 탐지 연구

텍스트 기반 접근은 문장 내 어휘·통사·문체적 패턴을 학습하여 허위정보의 언어적 특징을 식별한다. Conroy 등은 단어 빈도와 문장 길이, 부정적 감정어 비율을 이용한 전통적 텍스트 분류 기법이 허위뉴스 탐지에서도 유효함을 보였고[12], Wang은 LIAR 데이터셋을 활용해 정치 발언의 진위 판별을 위한 문장 수준 분류 모델을 구축하였다[13]. 이후 BERT, RoBERTa 등의 사전학습 언어모델이 등장하면서, 문맥 표현을 통합한 심층 언어표현 기반의 탐지가 활발히 이루어졌다. Zhang 등은 기사 본문과 요약문, 댓글 정보를 통합한 멀티뷰 학습(Multi-View Learning)을 통해 맥락적 상호보완성을 활용하였으며[14], 이러한 방법은 텍스트 단독 기반 모델보다 정밀도와 재현율이 모두 향상되는 결과를 보였다.

한편 그래프 기반 탐지는 뉴스·사용자·출처·댓글 간의 관계 구조를 학습한다. Monti 등은 Graph Convolutional Network(GCN)을 활용하여 노드 간 상호작용을 모델링하였고[15], 이후 GAT(Graph Attention Network)이나 Heterogeneous Graph 모델이 적용되며 뉴스 간 주제 유사도 및 출처 연결망을 반영한 탐지 체계가 제안되었다. 그러나 이러한 접근은 노드 정의와 연결 가중치 설정이 복잡하며, 학습 및 예측 단계에서 고비용 연산을 요구한다. 결과적으로, 실시간 탐지나 자원 제한 환경에서는 활용이 어려운 한계를 지닌다. 본 연구는 이와 달리 복잡한 네트워크 구조 대신 텍스트 내 구조 신호만을 이용하여, 단순·경량 모델로도 실질적 탐지 성능을 확보할 수 있음을 검증한다.

그래프 기반(GCN 계열) 탐지는 사용자–콘텐츠–전파 관계로 구성된 그래프 입력을 전제로 하므로, 텍스트 단독 환경을 가정한 본 연구의 실험 설계와 입력 가정이 다르다. 따라서 본 논문에서는 그래프 모델과의 수치 직접 비교 대신, 텍스트 단독 조건에서 재현 가능한 경량 베이스라인 비교를 중심으로 성능과 비용을 보고한다.

2-4 경량·비침습형 탐지의 최근 동향

최근 연구에서는 대규모 언어모델의 연산비용과 데이터 접근 제약을 완화하기 위해 경량·비침습형 탐지(Lightweight Non-Intrusive Detection) 접근이 주목받고 있다. Horne과 Adali는 단어 길이, 구두점 비율, 대문자 사용 빈도, 가독성 등 단순 문체 지표만으로도 허위뉴스의 문체적 차이를 포착할 수 있음을 보였으며[16], 의미 해석을 최소화한 구조 신호만으로도 유의한 분리력이 가능함을 강조하였다[15]. 다만 이들 연구는 데이터 구성과 분할 방식에서 본 연구의 정화 및 시간 고정 분할과 동일한 설정을 사용하지는 않았다.

한편 Kaliyar 등은 BERT 기반 모델(FakeBERT)을 통해 높은 탐지 성능을 보고하였으나[17], 사전학습 언어모델의 연산비용과 파인튜닝 설정, 데이터 편향 관리가 성능 해석에 중요한 전제가 됨을 함께 시사한다. 이는 입력 표현과 연산비용 측면에서 본 연구의 경량 구조 신호 기반 접근과 명확히 구분된다. 반면 Horne과 Adali의 결과는 고비용 의미 표현 없이도 제목·본문 길이, 단어 반복성, 구두점 및 대문자 사용과 같은 문체·형식적 신호가 효과적인 탐지 축이 될 수 있음을 뒷받침한다[16].

이와 유사하게 Singh 등은 문자·단어 n-그램, 뉴스 길이, 단락 구조, 문장 구두점 패턴을 결합한 경량 모델을 제안하였고, 이러한 비침습적 탐지는 개인정보 이슈나 학습 데이터 편향에 상대적으로 강건하여 공공 영역 적용에 유리함을 보였다. 그러나 기존 연구들은 입력 가정(의미 임베딩, 문체 통계, 전파 그래프 등)과 평가 프로토콜(IID 무작위 분할)이 상이하여, 보고된 AUC나 F1 수치를 직접 비교하기 어렵다. 특히 시간 이동과 출처 노출에 따른 과대평가 가능성을 통제하지 않은 설정에서의 성능은 실제 배치 환경을 반영하기 어렵다. 표 1은 선행연구의 입력 특징과 평가 설정을 논문에 보고된 범위 내에서 요약한 것이다.

본 연구는 이러한 경량화 흐름을 계승·확장하여, 의미 기반 표현학습 없이도 구조적 신호만으로 실질적인 분리력이 가능한지를 정화와 시간 고정 분할 하에서 실증적으로 검증한다.

2-5 본 연구의 차별점

기존 허위정보 탐지 연구는 대체로 대형 언어모델이나 다중모달 결합 구조에 의존하여 높은 성능을 달성했지만, 이러한 방식은 막대한 연산 자원과 대량의 주석 데이터, 그리고

표 1. 대표 허위뉴스 탐지 연구의 입력 특징 및 평가 설정 요약
Table 1. Summary of inputs and evaluation protocols in representative fake-news detection studies

Study	Input features	Model	Dataset	Split / Protocol	Reported performance	Comparability to this work
Horne & Adalı (2017) [16]	Stylometric features: title/body length, word/sentence complexity, stopword ratio, punctuation, uppercase ratio, lexical diversity, repetition	Linear SVM, Random Forest	BuzzFeed, PolitiFact, news articles	IID random split	Accuracy ≈ 0.76–0.78 (dataset-dependent)	Uses lightweight *style/format* cues similar in spirit, but relies on handcrafted statistics and IID split; no temporal split or decontamination
Rashkin et al. (2017) [13]	Lexical & stylistic word patterns	Logistic / SVM	News + fact-checking corpora	IID	Accuracy ≈ 0.74–0.76	Text-only lexical style, but no temporal setting; focuses on topic/style mixtures
Kaliyar et al. (2020)[17] – FakeBERT	Contextual semantic embeddings (BERT)	Fine-tuned BERT	ISOT / Kaggle Fake News	IID random split	Accuracy ≈ 0.98, F1 ≈ 0.98	Heavy PLM, semantic-dependent; high cost, IID protocol → not directly comparable to temporal lower-bound setting
Monti et al. (2019) [9]	User–content propagation graph + text	GCN / geometric DL	Twitter/ FakeNewsNet	Graph-based split	AUC ≈ 0.92–0.95	Requires social interaction graphs unavailable in text-only setting; different input assumption
This work	Decontaminated word/char n-grams (structural cues only)	Lightweight linear (SGD/LogReg), SVM, LightGBM	Kaggle True/Fake	Fixed temporal split (q80/q90)	PR-AUC 0.9768, ROC-AUC 0.9933, Macro-F1 0.9570	Focuses on *lower-bound* performance under time shift and source removal; directly comparable only to models under identical protocol

클라우드 기반 인프라에 대한 접근을 전제로 한다. 이에 따라 실제 언론사나 공공기관 등 제한된 자원 환경에서의 실무 적용성은 낮으며, 모델의 재현성과 비용 효율성 측면에서도 현실적 제약이 존재한다. 또한 대규모 언어모델 기반 접근은 학습 데이터의 편향을 증폭시키거나, 특정 도메인·언어에 과도하게 최적화되어 외부 환경에서의 성능 저하가 빈번하게 발생한다.

이에 반해 본 연구는 출처 식별이 불가능한 정화(Decontaminated) 텍스트를 기반으로 한 구조 신호 중심 탐지를 수행한다는 점에서 근본적인 차별성을 지닌다. 정화 절차를 통해 뉴스의 통신사명, 도메인 꼬리표, 저작권 문구 등 출처를 직접적으로 드러내는 모든 신호를 제거하고, 모델이 문체·구문·어휘 분포 등 비의미적 구조 신호에만 반응하도록 제한하였다. 이러한 설계는 모델의 성능을 과대평가하지 않으면서, 실제 환경에서 기대할 수 있는 최소 성능을 투명하게 제시한다는 점에서 의미가 크다.

특히 본 연구는 세 가지 측면에서 차별화된다. 첫째, 뉴스 데이터의 시계열적 변동을 반영한 고정 시계열 분할(q80, q90) 구조를 도입하여, 시간 흐름에 따른 주제 이동이나 언어 변화가 성능에 미치는 영향을 평가 가능하게 하였다. 둘째, 사전 규칙 기반 정화 절차를 적용하여 데이터 적응적 파라미터 추정 없이 일관된 전처리를 수행함으로써, 누출(leakage)을 원천적으로 차단하였다. 셋째, 문체·구문 중심의 피처 설계를 통해 의미 해석을 완전히 배제하면서도 문장 구조, 어휘 패턴, 구두점 사용 등에서 나타나는 '구조적 신호(structural cue)'를 체계적으로 반영하였다.

또한 기존 연구들이 특정 도메인 또는 특정 플랫폼 설정에 묶이는 경우가 많은 반면, 본 연구는 정화–벡터화–고정 분할이라는 절차를 명시하여 다른 코퍼스에도 동일한 방식으로 적용 가능한 재현 가능한 프로토콜을 제시한다. 다만 본 논문에서 보고하는 실증 결과는 단일 코퍼스 내 하한 성능이며, 외부 코퍼스 일반화는 별도의 교차 검증을 통해 확인되어야 한다.

이러한 접근은 허위정보 탐지의 패러다임을 '고비용 의미 기반 모델(semantic-based)'에서 '저비용 구조기반 모델(structure-based)'로 전환시키는 실증적 근거를 마련한다. 나아가 본 연구의 파이프라인은 학습·검증·보고 절차가 완전히 문서화되어 있어, 재현 가능(reproducible) 하고 투명성(auditability) 을 확보한 탐지 체계로 평가될 수 있다. 이는 허위정보 대응의 학문적 신뢰성과 실무적 효율성을 동시에 충족시킨다는 점에서 의의가 있다.

결과적으로, 본 연구의 설계는 단순히 알고리즘적 성능 개선을 넘어 정책적 적용성과 비용 효율성을 중심으로 한 새로운 연구 방향을 제시한다. 공공기관, 언론사, 교육기관 등에서 실시간 뉴스 검증이나 대규모 정보 모니터링 시스템을 구축할 때, 본 연구의 결과는 경량·비침습형 자동 검증 체계의 표준 프로토타입으로 활용될 수 있으며, 향후 다양한 언어·플랫폼 환경에서의 확장 적용 가능성을 제공한다.

Ⅲ. 연구 방법

3-1 데이터셋과 분할(고정)

본 연구는 Kaggle 공개 True/Fake 뉴스 데이터셋의 True.csv와 Fake.csv를 결합하여 사용한다. 관측 단위는 기사 문서이며, 허위(Fake)를 양성(1), 진짜(True)를 음성(0)으로 정의하였다. 전체 표본은 44,898문서로, 허위 23,481건과 진짜 21,417건으로 구성되며 전체 양성 비율 π는 0.523이다. 분석에는 title과 text를 사용하고, date 열은 분할을 위한 시간 정보로만 표준화하여 활용하였다. 날짜 결측은 극히 소수로, 해당 문서는 학습 분할에 포함시켜 총합을 보존하였다. 데이터는 약 3개년 기간을 포괄하며, 분위수 기준으로 시계열을 구분한다.

표 2. 데이터 구성과 분위수 기반 시간 고정 분할 요약(라벨 분포 포함)
Table 2. Data summary and quantile-based fixed temporal split (with label prevalence)

Item	Value
Total N	44,898
Fake (1)	23,481
True (0)	21,417
Overall Positive Rate π	0.523
Date Range	2015-03-31 → 2018-02-19
Split Boundaries	q80=2017-10-13, q90=2017-11-17
Train N (Fake rate)	36,080 (0.6129)
Dev N (Fake rate)	4,348 (0.1589)
Test N (Fake rate)	4,470 (0.1510)

현실적 일반화를 위해 시간 축 기준의 고정 분할을 적용하였다. 누적 분포의 하위 80% 지점(q80)을 학습–개발 경계로, 하위 90% 지점(q90)을 개발–테스트 경계로 설정하였다. 이에 따라 학습 분할은 q80 이전 구간(및 날짜 결측)을, 개발 분할은 q80–q90 구간을, 테스트 분할은 q90 이후 구간을 포함한다. 분할별 문서 수는 학습 36,080건, 개발 4,348건, 테스트 4,470건이며, 허위 비율은 각각 0.6129, 0.1589, 0.1510으로 나타났다. 즉, 과거 구간에서 허위 비중이 높고 최근으로 갈수록 진짜 기사 비중이 증가하는 시간적 편차가 확인된다.

이러한 분포는 무작위(IID) 분할 대비 보수적인 평가를 유도하며, 4장에서 보고하는 PR-AUC와 운영 지표(P/R@k)는 시간 이동을 고려한 실무 배치 성능의 하한으로 해석된다. 개발 분할에서 확정한 임계값과 보정은 테스트 분할에 변경 없이 적용하고, 모든 전처리·벡터화 파라미터는 학습 분할에서만 적합한 뒤 개발·테스트에는 적용만 한다. 재현성을 위해 라벨 정의, 분할 규칙(q80, q90), 난수 시드, 전처리 사양을 문서에 고정 기록하였다.

허위(Fake)를 양성(1)으로 정의하였다. 날짜 결측치는 학습 분할에 귀속시켜 총합 보존을 유지하였다. 시간 기반 분할은 분류기의 시간 이동 강건성을 점검하기 위한 설계이며, 개발 분할에서 선택한 임계값·보정 절차는 테스트 분할에 그대로 적용한다.

본 데이터는 출처 표식이 라벨과 강하게 공변하는 구조적 특성이 있어, 모델이 출처 식별만으로 과대 성능을 보일 수 있다. 본 연구는 이러한 위험을 선제적으로 낮추기 위해, 벡터화 이전 단계의 정화 규칙을 전 분할에 동일 적용하여 출처·저작권·크레딧·도메인 꼬리표를 제거한 텍스트만 분석에 사용한다. 이로써 보고되는 수치는 출처 노출 여부나 플랫폼 변화에 상대적으로 둔감한 문체·구문 기반의 최소 성능으로 해석된다.

3-2 입력 표현과 전처리

본 연구는 텍스트만을 사용하되 의미 해석에 의존하지 않는 구조 신호를 일관되게 추출하기 위해, 모든 문서를 title과 text를 줄바꿈으로 결합한 단일 문자열로 구성하였다. 결측은 빈 문자열로 대체하되 두 필드가 모두 결측인 문서는 제외하였다. 정규화 단계에서는 유니코드 표준화(NFC), 공백 압축, URL·이메일 제거, 연속 구두점의 단일화만을 수행하고, 숫자와 대부분의 구두점은 보존하여 주장·양식적 패턴을 유지하였다. 토큰화는 공백 기반의 단순 절차를 따르며, 의미 사전·어간추출·표제어화는 적용하지 않았다. 이러한 최소주의 전처리는 3.1절에서 확정한 고정 시간 분할 하에서 누출 없이 재현 가능한 비교를 보장한다. 모든 통계량과 변환 파라미터(어휘집, IDF, 정규화)는 학습 분할에서만 적합하고 개발·테스트 분할에는 적용만 하며, 학습 분할에 한해 20토큰 미만의 극단적 단문을 제외하여 과도한 잡음을 줄였다.

문서 표현은 두 계열로 병행한다. 첫째, 단어 1–2그램 TF–IDF로 어휘 빈도·역빈도를 반영해 문장 구조와 어휘 분포의 차이를 포착한다. 둘째, 문자 3–5그램 TF–IDF로 철자·기호·띄어쓰기 패턴의 반복성을 포착하여 오탈자·표기 변형에 대한 견고성을 확보한다. 두 표현 모두 벡터 크기를 제한하여 소논문 규모의 계산 자원을 전제로 한 경량 학습을 가능하게 한다. 어휘집과 IDF는 학습 분할에서만 구축하고, 개발·테스트에는 동일 사전을 고정 적용한다. 분할 이전 단계에서 수행한 완전중복 제거와 달리, 유사중복(표현만 약간 다른 문서)의 처리는 별도의 탐지 기법을 사용하지 않고 시간 분할의 특성으로 흡수하며, 이는 최신 구간 성능의 보수적 추정이라는 본 연구의 목적에 부합한다. 결과적으로, 본 절의 설계는 의미 해석을 배제한 구조 신호만으로도 상위 검토 큐에서의 정렬 개선이 가능한지를 검정하기 위한 최소 충분 조건을 제공한다.

본 연구의 정화(Decontamination)는 “출처 식별(agency/domain/boilerplate) 신호”가 라벨과 공변하여 과대평가를 유발할 수 있다는 점을 통제하기 위한 전처리이다. 이를 위해 (i) 통신사/보도기관 표식, (ii) 데이트라인(도시명+ 기관표식 + 구분자), (iii) 크레딧 문구(Reporting/Editing by …), (iv) 저작권/도메인 꼬리표(©, all rights reserved, *.com 등)를 사전 고정된 규칙 기반 정규표현식으로 제거하였다.

표 3. 텍스트 입력 표현 및 전처리 요약
Table 3. Summary of text input representation and preprocessing

Item	Setting
Document composition	title+ newline + text; impute missing fields with empty strings; drop if both are missing.
Normalization	Unicode NFC; collapse consecutive whitespace; remove URLs/emails; collapse repeated punctuation to one; keep digits and most punctuation.
Tokenization	Whitespace-based simple tokenization (no dictionary, stemming, or lemmatization).
Word TF-IDF	n_gram=(1,2), min_df=3, max_features≈200,000, sublinear_tf=True, smooth_idf=True, L2 normalization.
Character TF-IDF	n_gram=(3,5), min_df=3, max_features≈100,000, sublinear_tf=True, smooth_idf=True, L2 normalization.
Fit-apply scope	Fit all vocabularies/IDF/scaler parameters on the trainsplit only; apply to dev/test.
Short-text handling	In the trainsplit, exclude documents with token length < 20 (noise mitigation); apply the same exclusion rule to dev/test.
Auxiliary variables	subjectunused in the base model (leakage/bias risk); reported only as an ablation in §4.

표 4. 정화 규칙 요약
Table 4. Summary of decontamination rules

Rule ID	Category	Pattern examples (illustrative)	Removed example
R1	Agency marks	₩b(?:reuters|associated press|ap)₩b	associated press
R2	Dateline tags	^[A-Z][A-Z .-]{2,}₩s*₩((?:reuters|ap)₩)₩s*[—–]₩s*	WASHINGTON (Reuters) –
R3	Credits line	(?i)₩bReporting by₩b.*?(?:;₩s*₩bEditing by₩b.*)?$	“Reporting by …; Editing by …”
R4	Copyright	(?i)©₩s*₩d{4}.*?all rights reserved₩.?	“© 2017 … All rights reserved.”
R5	Domains/URLs	(?i)₩b(?:https?://|www₩.)₩S+|₩b₩S+₩.(?:com|org|net|edu|gov)₩b	“example.com”, “https://…”
R6	Emails	(?i)₩b[A-Z0-9._%+-]+@[A-Z0-9.-]+₩.[A-Z]{2,}₩b	“abc@xyz.com”

정화 규칙은 분할 이전에 확정되어 연구 전 과정에서 변경하지 않았고, train/dev/test 전 분할에 동일 적용하였다. 또한 규칙은 데이터로부터 어떤 파라미터도 추정하지 않는 비학습(Non-learning) 절차이므로 전처리 단계에서의 통계적 누출 위험이 없다. 정화 규칙의 요약과 예시는 표 3-3에 제시한다.

3-3 모델과 평가 설계

본 절의 목적은 3.1절에서 확정한 분위수 기반 시간 분할(q80·q90)과 3.2절의 텍스트 구조 신호 입력을 고정한 상태에서, 경량 분류기의 성능을 공정하게 평가하는 데 있다. 모든 학습은 q80 이전 구간에서만 수행하고, q80–q90 구간의 개발 분할에서는 필요한 선택을 한 번만 확정한 뒤, q90 이후 테스트 분할에는 조정 없이 결과만 산출하여 실제 배치 초기의 하한 성능을 추정한다.

학습 단계에서는 단어·문자 n-그램 TF-IDF 벡터화기와 분류기를 학습한다. 개발 단계에서는 규제 강도 C와 라벨 기반 표 작성을 위한 확률 임계값을 각각 한 번만 선택하며, 테스트 단계에서는 이를 그대로 적용해 지표를 계산한다. 테스트 결과를 참조한 재조정은 허용하지 않는다.

기본 분류기는 해석성과 안정성을 고려해 L2 로지스틱 회귀로 설정한다. 계산 비용이 낮고 n-그램 계수를 통해 구조 신호의 기여를 해석할 수 있으며, 시간에 따른 주제·문체 변화 환경에서도 과적합 위험이 낮다. 클래스 가중은 균형으로 고정하고, 규제 강도 C는 개발 구간의 평균정밀도(PR-AUC)를 기준으로 소수 후보 중에서 선택한다.

구조 신호 기반 분리력이 특정 분류기에 종속되지 않음을 확인하기 위해, 동일 입력 표현과 시계열 분할 조건에서 선형 SVM과 LightGBM을 보조 비교로 포함한다. 이때 정화 규칙, 분할, 피처 표현은 고정하고 분류기만 교체하며, 비교 지표는 PR-AUC와 상위 k% 운용 지표로 통일한다.

본 연구는 대형 사전학습 언어모델과의 최고 성능 경쟁을 목표로 하지 않는다. 대신 정화와 시계열 고정 분할을 적용한 조건에서, 구조 신호만으로 상위 검토 큐(Top-k%)에서 확보 가능한 정밀도와 업무량 효율을 재현 가능한 기준선으로 제시하는 데 초점을 둔다.

입력 표현은 단어 1-2그램 TF-IDF, 문자 3-5그램 TF-IDF, 그리고 두 표현의 단순 결합으로 구성한다. 개발 단계에서 점수를 다시 학습하는 후기 결합이나 스태킹은 미래 분포 적합 위험이 있어 사용하지 않는다. 벡터화기의 어휘집과 IDF 가중치는 학습 구간에서만 적합하고, 개발·테스트 구간에는 그대로 적용하여 실제 배치 초기의 불리함을 반영한다.

임계값은 라벨이 필요한 표 작성을 위해서만 개발 구간에서 Macro-F1이 최대가 되는 값으로 한 번 정해 테스트에 적용한다. PR-AUC와 상위 k% 기준의 정밀도·재현율은 임계값과 무관하므로 예측 점수의 순위만으로 계산하며, 순위 동점 시에는 고정 규칙을 적용해 재현성을 확보한다.

확률 절대값 해석이 필요한 경우에 한해 단조 보정을 개발 구간에서 한 번 적합해 테스트에 고정 적용하되, 본문 결과에는 반영하지 않고 부록으로 제시한다. 누출 방지를 위해 출처 관련 보일러플레이트 제거 규칙은 사전에 고정된 정규표현식

표 5. 모형 및 학습·평가 설정 요약
Table 5. Summary of models and training/evaluation settings

Item	Setting
Input representation	Word TF-IDF(*n-gram=(1,2), min_df=3, max_features≈200,000, sublinear_tf=True, L2 normalization*)Character TF-IDF(*n-gram=(3,5), min_df=3, max_features≈100,000, sublinear_tf=True, L2 normalization*)Word+Charfeature union (concatenation).
Classifier	Logistic regression (L2 penalty, solver=lbfgs, class_weight="balanced").
Hyperparameter selection	Select (C) from a small grid by maximizing AP (PR-AUC) on the devsplit.
Threshold rule	Fix the decision threshold at the devmacro-F1 optimum and apply it to test.
Fit-apply scope	Fit vocabularies/IDF/scaler parameters on trainonly; apply to dev/test.
Repetition & reporting	Report mean ± SD over 3 random seeds; testevaluation performed once.
Primary & auxiliary metrics	Primary: PR-AUC, Macro-F1. Auxiliary: Accuracy, ROC-AUC. Operational: (P@k, R@k) for (k={1%, 5%, 10%}).
Leakage control	Perform deduplication beforesplitting; document all choices with the split held fixed.

으로만 처리하고, 주제 표식은 학습에서 제외한다. 또한 라벨 무작위화와 상수 예측을 통해 평가 위생을 점검한다.

PR-AUC는 상위 구간 선별력을 요약하고, 상위 k% 지표는 실제 검토량과 직접 연결되는 운영 지표로 해석한다. Macro-F1은 임계값 참고용으로만 사용하며, 본문에는 테스트 결과만 제시하고 개발 수치는 부록으로 이관한다. 모든 표와 그림에는 분할 경계(q80, q90), 난수 시드, 선택된 C와 임계값을 명시해 재현성을 확보한다.

정화 규칙은 연구 시작 시 사전 고정된 규칙 기반(비학습) 처리로 정의되며, 어떤 파라미터도 학습 데이터에서 추정하지 않으므로 train→dev/test로의 통계적 누출 가능성이 없다. 모든 전처리 중 학습에 의존하는 요소(예: 표준화 계수)는 train에서만 적합하고 dev/test에는 적용만 한다.

3-4 평가 및 보고 절차

3-1절에서 시간 기준 분할(q80, q90)을 고정하고, 3.2-3.3절에서 입력 표현과 모델·임계값 선택 규칙을 확정했으므로, 본 절에서는 이 설정을 유지한 채 지표 산출과 보고 방식을 규정한다. 모든 선택은 학습·개발 단계에서 잠그고, 테스트 구간에는 그대로 적용하여 실제 배치 초기의 보수적 하한 성능을 재현 가능하게 제시한다.

주요 해석 지표는 평균정밀도(AP, PR-AUC)이다. 시기별 양성비가 변하는 데이터 특성상 ROC보다 PR 곡선이 상위 구간 선별력을 직접 반영하므로, 무작위선은 전체 양성비(π)로 표시해 모든 표와 그림에 함께 제시한다. PR-AUC는 임계값에 의존하지 않아, 개발 구간에서 고정한 임계값을 유지한 상태에서도 순위 기반 효율 비교가 가능하다.

운영 지표로는 P@k와 R@k를 사용한다. "상위 k% 검토" 가정은 실제 검토량과 직접 연결되며, 테스트 문서 수가 고정되어 있으므로 각 k별 검토량과 적중·오탐 건수를 함께 보고한다. 점수 동률 시에는 확률 값, 내부 정렬 키, 문서 ID 순으로 고정 규칙을 적용해 재현성을 확보한다.

확률 임계값은 개발 구간에서 Macro-F1이 최대가 되는 값으로 한 번만 정해 테스트 구간에 그대로 적용한다. 확률 절대값 해석이 필요한 경우에 한해 단조 보정(Platt 또는 Isotonic)을 개발 구간에서 한 번 적합해 테스트에 고정 적용하되, 본문 결과에는 반영하지 않고 부록으로 제시한다.

불확실성은 동일 설정 반복 학습을 통한 평균·표준편차와, 테스트 예측쌍 고정 하의 층화 부트스트랩으로 추정한 PR-AUC 및 Macro-F1의 95% 신뢰구간으로 제시한다. 모델 간 차이는 동일 재표본의 쌍대 차이 분포로 평가하며, 모든 표에는 분할 경계(q80, q90), 난수 시드, 선택된 C, 임계값, 보정 여부를 명시한다.

누출과 과대평가는 절차로 통제한다. 벡터화기의 어휘집과 IDF는 학습 구간에서만 적합하고, 정화는 사전에 고정된 정규표현식으로만 수행한다. subject 열은 학습에서 제외하며, 라벨 무작위화와 상수 예측을 통해 평가 위생을 점검한다.

보고는 재현성을 최우선으로 하여 본문에는 테스트 결과만 제시하고, 개발 수치와 중간 선택 과정은 부록으로 이관한다. 본 결과는 Kaggle True/Fake 뉴스(2015–2018) 단일 코퍼스에서 정화와 시계열 고정 분할을 적용해 출처 기반 과대평가를 통제한 텍스트 구조 신호 기반 하한 성능으로 해석되며, 다른 코퍼스나 최신 합성 텍스트로의 일반화는 후속 검증 과제로 남긴다.

Ⅳ. 결 과

4-1 테스트 구간 주결과

본 절은 분위수 기반 시간 분할(q80·q90)에서 고정된 테스트 구간(q90 이후)에 대해, 정화(Decontamination)된 텍스트만으로 학습·평가한 최종 성능을 보고한다. 정화는 벡터화 이전 단계에서 통신사명, 데이트라인, 저작권·도메인 꼬리표 등 출처 표식을 제거하여 출처 식별에 따른 과대평가를 차단하고, 문체·구문·어휘 분포에 해당하는 2차적 구조 신호만으로 분리력을 측정하기 위한 절차이다(규칙은 연구 시작 시 고정되어 전 분할에 동일 적용). 입력은 §3-2의 단어 1-2그램이며, 분류기는 HashingVectorizer(32,768차원, L2)와

표 **6.** 평가 및 보고 체계 요약
Table 6. Summary of evaluation and reporting protocol

Item	Specification
Primary metric	PR-AUC (AP); no-skill line shown at positive-class prevalence (₩pi).
Secondary & auxiliary	Macro-F1 (threshold fixed from dev), Accuracy, ROC-AUC.
Operational metrics	Precision@k, Recall@k ((k = 1%, 5%, 10%)).
Thresholding	Select once on the devsplit by maximizing Macro-F1 → fix for test.
Calibration	Platt scaling / Isotonic (see appendix); not reflected in main reported numbers (reliability diagram shown separately).
Repetition & uncertainty	Train with 3 random seeds; report mean ± SD. Compute 95% CIs via 1,000 stratifiedbootstrap resamples on the fixed testpredictions.
Leakage control	Fit vectorization/normalization on trainonly; apply to dev/test. Perform deduplication beforesplitting.
Reporting	Provide bilingual (Korean/English) captions for tables/figures; state split, hyperparameters, threshold, and seeds; no redistribution of raw data.

SGD(로지스틱 손실, class_weight=balanced)로 구성하였다. 임계값은 개발 구간에서 Macro-F1 최대화로 선택한 값을 테스트에 고정 적용하였다(thr=0.60). 테스트 집단의 양성비는 $\pi \approx 0.151$이다.

정화 후 전집 성능은 무작위선(π)을 크게 상회하였다(PR-AUC=0.9768, ROC-AUC=0.9933, Macro-F1=0.9570, Accuracy=0.9785). 운영 지표는 "상위 k% 검토" 기준으로 정보성이 낮은 1% 구간을 제외하고 k={5%,10%,15%,20%}에 대해 보고한다. Linear SVM과 LightGBM은 동일 입력과 시계열 분할 조건에서 별도 검증하였으며, PR-AUC과 ROC-AUC 모두 기준선(Logistic) 대비 ±0.003 이내로 유사하였다. 이에 따라 표 7에는 대표성 있는 Logistic 결과만 제시한다.

표 7. 테스트 구간 성능(정화 후) — 1차 지표
Table 7. Test-split performance after decontamination — Primary metrics

Model / Input	PR-AUC	ROC-AUC	Macro-F1	Accuracy
SGD (Word 1-2gram, Hashing)	0.9768	0.9933	0.9570	0.9785

표 8. 테스트 구간 성능(정화 후) — 운영 지표(상위 k% 검토)
Table 8. Test-split performance after decontamination — Operational metrics (top-k review)

(k)	Precision@(k)	Recall@(k)	Workload	TP	FP	FPR
5%	**1.000**	**0.331**	100	100	0	0.0000
10%	**0.995**	**0.659**	200	199	1	0.0006
15%	**0.713**	**0.709**	300	214	86	0.0506
20%	**0.573**	**0.758**	400	229	171	0.1007

다만 본 연구의 정화는 출처명·도메인·크레딧 등 "직접적인 출처 식별 신호"를 제거하는 데 초점을 두며, 그 결과는 출처 의존 과대평가를 완화한 상태에서의 분리력을 보여준다. 그러나 Kaggle True/Fake는 수집·편집 과정에서 문체 규범(예: 보도체 관용구, 데이트라인 관례, 문장 구성) 자체가 라벨과 간접적으로 공변할 수 있어, 정화 이후에도 데이터셋 특유의 문체 편향이 남아 있을 가능성을 배제할 수 없다. 따라서 본 절의 결과는 "구조 신호만으로도 분리력이 관측될 수 있다"는 실증과 함께, 단일 데이터셋 내 하한 성능으로 해석되어야 하며, 외부 코퍼스/최근 생성형 텍스트로의 일반화는 별도 검증이 필요하다(§3-4, §5).

표 9. 동일 입력·동일 시간 분할에서의 경량 베이스라인 직접 비교
Table 9. Direct comparison of lightweight baselines under the same input and temporal split

(k)	Precision@(k)	Recall@(k)	Workload	TP	FP	FPR
5%	**1.000**	**0.331**	100	100	0	0.0000
10%	**0.995**	**0.659**	200	199	1	0.0006
15%	**0.713**	**0.709**	300	214	86	0.0506
20%	**0.573**	**0.758**	400	229	171	0.1007

4-2 임계값·업무량 운용 분석

본 절은 4-1의 정화(Decontamination) 이후 결과를 상위 k% 검토 전략으로 재배치하여, 실제 업무량 대비 적중·오탐·거짓양성률의 교환관계를 정량화한다. 분할과 전처리, 임계값 선택(thr=0.60 고정)은 §3-3과 §4-1의 설정과 동일하다. 테스트 집단은 파일럿 기준 $N \approx 2{,}000$이며, 양성비는 $\pi \approx 0.151$이다. 정보성이 낮은 1% 구간은 생략하고, 보고 단위를 k={5%,10%,15%,20%}로 고정하였다.

정량 결과는 다음과 같다. 상위 5%와 10% 구간은 오탐이 사실상 발생하지 않는 초반 필터를 형성한다. 상위 15%와 20%로 확장하면 재현율은 단조 증가하되 정밀도는 완만히 저하되며, 그 하락폭은 여전히 베이스레이트(π) 대비 높은 리프트를 유지하는 범위에 머문다. 특히 k=20%에서조차 정밀도는 π의 약 3.8배 수준이다.

이 결과는 운영 해석을 명확히 제시한다. 첫째, 상위 5-10%만 검토해도 적중이 각각 100건, 199건으로 집약되며 오탐은 0-1건 수준이다. 둘째, 10%에서 15%로 확장하면 추가 검토 100건당 추가 적중이 약 15건, 15%에서 20%로 확

표 10. 상위 k% 운용 시나리오(정화 후) — 업무량·적중·오탐
Table 10. Top-k% operational scenarios (after decontamination) — Workload, TP, FP

(k)	Workload	True Positives	False Positives	Precision@ (k)	Recall@ (k)	FPR
5%	100	100	0	1.000	0.331	0.0000
10%	200	199	1	0.995	0.659	0.0006
15%	300	214	86	0.713	0.709	0.0506
20%	400	229	171	0.573	0.758	0.1007

장하면 추가 검토 100건당 추가 적중이 약 15건으로 유지되지만, 오탐 비용은 각각 86건, 85건 수준으로 증가한다. 셋째, 그럼에도 15%와 20%의 정밀도는 각각 0.713, 0.573으로 π 대비 4.7배, 3.8배의 리프트를 유지한다. 따라서 실무적으로는 정밀도 우선의 초기 단계(5–10%)를 기본 운용으로 삼고, 재현율 확대가 필요할 때 15–20%까지 단계적으로 넓히는 전략이 비용 대비 효과 면에서 합리적이다. 컷 확률을 절대값으로 해석해야 하는 환경에서는 개발 분할 기준의 단조 보정(Platt/Isotonic)을 일회 적용하면 임계값 해석의 안정성을 높일 수 있다. 순위 기반 큐레이션이 가능한 환경에서는 현설정만으로도 충분한 리프트를 확보한다.

종합하면, 정화 이후에도 본 경량 파이프라인은 상위 5–10%에서 거의 무오탐의 초기 큐를 안정적으로 제공하고, 15–20% 확장 시 리프트를 유지한 채 재현율을 의미 있게 증대시킨다. 변수 중요도는 이러한 분리의 근거가 출처명 자체가 아닌 문체·구문·어휘 분포의 차이에 있음을 보여주며, 이는 출처 노출이 제한되거나 변형되는 실제 환경에도 비교적 강건한 운용 가능성을 시사한다.

4-3 재현율 목표 기반 임계값 운용

4-1~4-2에서 확인한 상위 구간의 정밀도 우위와 단조적인 재현율 증가 관계를 이제 실제 정책 임계값으로 옮겨 적는다. 절차는 단순하다. 먼저 달성하고자 하는 재현율 수준을 정하고, 그 재현율에 대응하는 상위 k% 검토 비율을 선택한 뒤, 그 선택이 만들어낼 업무량(검토 건수), 기대 적중(TP), 오탐(FP), 거짓양성률(FPR)을 사전에 계산해 둔다. 분할·정화·벡터화·점수 생성 규칙은 앞 절과 동일하며, 테스트 집단은 파일럿 기준 N≈2,000, 양성비는 π≈0.151이다. 재현율–k의 단조 관계(§4.2)에 따라 실무적으로 의미 있는 세 구간, 즉 R≈0.66, R≈0.71, R≈0.76에 대응하는 운영점을 도출했으며 요약은 표 11에 정리되어 있다.

상위 10%만 검토할 경우 업무량은 200건이며, 적중 199건·오탐 1건으로 관측되었다. 정밀도 0.995, 재현율 0.659, 거짓양성률 0.0006으로, 상위 점수 구간에 문체·구문 신호가 강하게 응집되어 있음을 보여준다. 이 구간은 초기 대응 단계에서 거의 무오탐에 가까운 큐 구성이 가능하며, 일반 운영에서도 비용 대비 효과가 가장 크다.

재현율을 높이기 위해 15%로 확장하면 업무량은 300건, 적중 214건, 오탐 86건으로 증가하며 정밀도 0.713, 재현율 0.709를 보인다. 이는 검토 100건당 적중 약 15건 증가와 함께 오탐 약 85건이 추가되는 구간으로, 정밀도는 모집단 양성비(π≈0.151) 대비 약 4.7배의 리프트를 유지한다.

재현율을 약 0.76까지 끌어올리기 위해 20%로 확장할 경우 업무량은 400건, 적중 229건, 오탐 171건이며 정밀도는 0.573으로 낮아지지만, 여전히 양성비 대비 약 3.8배의 리프트를 유지한다. 이는 강한 재현율 목표가 설정된 국면에서 선택 가능한 상한선으로 해석된다.

표 11의 운영점들은 대체 관계가 아닌 단계적 확장 구조를 갖는다. 평시에는 10%를 기본으로 운용하고, 필요 시 15%로 재현율을 보강하며, 불가피한 경우에만 20%까지 확대한다. 각 단계는 "추가 100건당 적중 약 15건, 오탐 약 85건"의 교환비용을 수반하므로, 양성 1건의 편익이 오탐 1건의 비용보다 충분히 큰 경우에만 확장이 정당화된다.

용량 관점에서는 주간 검토 200건·300건·400건이 각각 10%·15%·20% 운영에 대응하며, 이는 예산·인력 계획 시 기대 적중과 오탐을 사전에 추정하는 기준으로 활용할 수 있다. 표 4-3의 수치는 점추정치이므로 반복 학습 분산과 층화 부트스트랩 신뢰구간을 함께 고려하는 것이 바람직하다. 또한 정책 필터를 결합해 모집단을 축소하면, 표 4-3의 운영점은 보수적 하한으로 작동하며 동일 재현율을 더 작은 k에서 달성할 수 있다.

4-4 비용 민감도와 운영점 선택

본 절은 4-2의 상위 k% 운용 결과를 기반으로, 검토 비용과

표 11. 재현율 목표별 권장 운영점(정화 후) — 업무량·정밀도·오탐
Table 11. Recommended operating points by target recall (after decontamination) — Workload, precision, FP

Target Recall	Recommended (k)	Precision@ (k)	Recall@ (k)	Workload	(TP)	(FP)
(₩approx 0.66)	10%	**0.995**	**0.659**	200	199	1
(₩approx 0.71)	15%	**0.713**	**0.709**	300	214	86
(₩approx 0.76)	20%	**0.573**	**0.758**	400	229	171

오탐·미탐 비용을 명시적으로 도입한 비용 민감도 분석을 통해 최적 운영점을 선택하는 절차를 제시한다. 목적은 두 가지다. 첫째, k 확장에 따른 한계 수익(추가 적중)과 한계 비용(추가 오탐·추가 검토)의 교환을 정량화해, 조직의 비용·위험 선호에 맞춘 합리적 결정을 가능하게 하는 것이다. 둘째, 주어진 검토 용량(capacity)이 있을 때 기대 적중과 오탐을 미리 산정해 실행계획으로 바로 전환할 수 있게 하는 것이다. 분할, 전처리, 정화 규칙, 임계값(thr=0.60 고정)은 앞 절과 동일하다.

비용 모형은 다음과 같이 정의한다. 한 건의 미탐(양성 놓침) 비용을 c_fn, 오탐 비용을 c_fp, 검토 1건 비용을 c_r라 하자. 상위 k%로 확장할 때의 순이득 증분은 $\Delta U(k) = b\cdot\Delta TP - c_fp\cdot\Delta FP - c_r\cdot\Delta W$로 표현되며, 여기서 b는 양성 1건 적발의 편익(통상 c_fn과 동치로 본다), ΔTP·ΔFP·ΔW는 각각 추가 적중, 추가 오탐, 추가 검토량이다. 확장이 타당하려면 $b/c_fp \geq (\Delta FP + (c_r/c_fp)\cdot\Delta W) / \Delta TP$를 만족해야 한다. 즉, 오른쪽 항은 확장을 정당화하기 위한 최소 편익–비용 비율의 경계값이다.

표 12. k 확장 시 한계량과 편익–비용 경계비율
Table 12. Marginal gains and break-even benefit-cost ratios for expanding k

Expansion interval	ΔTP	ΔFP	ΔW	boundary b/c_fp (c_r=0)	boundary b/c_fp (c_r=0.5·c_fp)
5% → 10%	99	1	100	0.010	0.515
10% → 15%	15	85	100	5.667	9.000
15% → 20%	15	85	100	5.667	9.000

표 12는 4.2의 관측치(5%, 10%, 15%, 20%)를 이용해 인접 k 구간 확장의 한계량을 정리하고, 두 가지 대표적 비용 시나리오에 대한 경계비율을 제시한다: (i) 검토 비용 무시(c_r=0), (ii) 검토 1건 비용이 오탐의 절반(c_r=0.5·c_fp).

해석은 분명하다. 5%에서 10%로의 확장은 검토 비용을 무시하면 $b/c_fp \geq 0.010$이면 충분하므로 거의 항상 정당화된다. 검토 1건이 오탐의 절반 비용일 정도로 비싸다고 가정하더라도 $b/c_fp \geq 0.515$면 확장이 타당하다. 반대로 10%→15%와 15%→20% 확장은 추가 적중 15건을 위해 추가 오탐 85건과 추가 검토 100건이 수반되므로, 검토 비용을 무시해도 b/c_fp가 약 5.67 이상, 검토비가 오탐의 절반일 경우 약 9.0 이상이어야 확장이 합리적이다. 즉, 위기 국면 등에서 양성 한 건의 편익이 오탐 한 건의 비용보다 현저히 클 때에만 15% 이상 확대가 비용 대비 효과를 갖는다.

검토 용량 관점의 계획 수립도 즉시 가능하다. 표 13은 주어진 용량 L에 대응하는 k 선택과 기대 적중·오탐을 제시한다. 값은 4.2의 관측치를 그대로 사용한다.

이 매핑을 사용하면, 예를 들어 용량이 200건일 때 10% 운용이 자연스럽고, 추가 100건의 예산을 확보하는 경우 15%로 확장했을 때의 기대 적중(+15)과 오탐(+85)을 사전에 제시할 수 있다. 더 나아가 조직의 비용·편익 스칼라를 내부 합의로 정하면(예: b/c_fp=6, c_r/c_fp=0.3), 표 12의 경계비율과 비교하여 어떤 확장이 정당화되는지를 즉시 판별할 수 있다. 이 절차는 분기별로 데이터 분포가 변하더라도(이벤트·이슈 집중 등) 운영점 재설정이 공식화되어 있다는 장점이 있다.

요약하면, 5%→10% 확장은 대부분의 비용 시나리오에서 우세하며, 15% 이상 확장은 양성 적발의 편익이 오탐·검토 비용보다 상당히 클 때 선택된다. 본 비용 민감도 분석은 4.2의 k별 성능 요약과 중복되지 않으면서, 실제 정책·운영 의사결정을 직접적으로 지원하는 규칙적 선택 기준을 제공한다.

V. 결 론

본 연구는 벡터화 이전 단계에서 출처 표식과 보일러플레이트를 제거하는 정화(Decontamination) 절차를 거친 뒤, 단어 1–2그램 구조 신호와 경량 선형 분류기(HashingVectorizer + SGD)만으로 허위정보 탐지를 수행하였다. 정화의 목적은 출처 식별에 의한 과대평가를 차단하고, 문체·구문·어휘 분포 등 구조 신호만으로 달성 가능한 최소 성능을 투명하게 제시하는 데 있다. 분위수(q80·q90) 기반 시간 분할을 고정하고 개발 구간에서 선택한 임계값을 테스트 구간에 그대로 적용한 결과, PR-AUC 0.9768, ROC-AUC 0.9933, Macro-F1 0.9570, Accuracy 0.9785를 달성하였다. 상위 검토 시나리오에서 k=5%는 Precision 1.000·Recall 0.331, k=10%는 0.995·0.659, k=15%는 0.713·0.709, k=20%는 0.573·0.758로 관측되었다. 상위 5–10% 구간에서 정밀도가 무오탐에 가까운 수준을 보여, 출처명이 제거된 후에도 뉴스 보도체의 규범적 표현과 클릭 유도형 문체 간의 명확한 분리 축이 존재함을 시사한다.

초기 대응 단계에서는 상위 10% 내 검토만으로 높은 정밀도를 확보해 불필요한 검토 비용과 오탐 위험을 줄일 수 있다.

표 13. 용량 기반 운영점 매핑(정화 후)
Table 13. Capacity-based mapping of operating points (after decontamination)

Weekly review capacity (L)	Review rate (k)	Expected TP	Expected FP	Precision@ k	Recall@ k
100	5%	100	0	1.000	0.331
200	10%	199	1	0.995	0.659
300	15%	214	86	0.713	0.709
400	20%	229	171	0.573	0.758

재현율 확대가 필요할 경우 k를 15% 또는 20%로 확장할 수 있으나, 추가 적중에 비례해 오탐과 비용이 증가한다. 비용 민감도 분석 결과, 5%→10% 확장은 대부분의 시나리오에서 정당화되며, 10%→15% 및 15%→20% 확장은 양성 1건의 편익이 오탐·검토 비용을 충분히 상회할 때만 합리적이다. 조직은 목표 재현율과 오탐 예산을 사전에 정의하고 이에 맞춰 k를 조정함으로써 인력·정책 비용을 통제하면서도 안정적인 성과를 확보할 수 있다.

학문적 기여는 두 가지로 요약된다. 첫째, 출처 표식 제거 후에도 문체·구문 신호만으로 높은 분리력을 확인해 의미 해석을 최소화한 경량 파이프라인의 실효성을 검증하였다. 둘째, 전집 곡선 지표와 상위 k% 기반 업무량 지표를 병기하고, 비용 민감도 틀로 운영점을 산정하는 절차를 제시해 실무 적용 가능성을 높였다.

데이터 최신성 및 생성형 AI 기반 합성 텍스트에 대한 직접 일반화 한계는 §3-4의 "실험 범위(최신성) 제한"에서 선제적으로 명시하였으며, 본 결론에서는 해당 한계를 전제로 결과의 해석·활용 범위를 정리하였다. 향후 연구에서는 정화 원칙과 비용 제약을 유지하면서 문장 길이, 구두점, 대문자 비율 등 저비용 구조 신호의 확장과 시간적 빈도 특성을 결합해 상위 10–20% 구간의 정밀도 저하 완화를 모색할 수 있다. 또한 외부 코퍼스 기반 검증과 분기별 임계값 재보정을 통해 주제·문체 드리프트에 대한 내성을 강화할 필요가 있다.

요컨대, 본 연구는 정화된 텍스트 환경에서도 경량 파이프라인이 상위 소구간에서 무오탐에 가까운 초기 필터를 제공하고, 단계적 확장을 통해 재현율을 현실적으로 끌어올릴 수 있음을 실증하였다. 이는 출처 노출이 제한되거나 변형되는 실제 환경에서 낮은 침습성과 비용으로 정책 목표에 부합하는 탐지를 구현할 수 있음을 보여준다.

참고문헌

[1] S. Vosoughi, D. Roy, and S. Aral, "The Spread of True and False News Online," *Science*, Vol. 359, No. 6380, pp. 1146-1151, 2018. https://doi.org/10.1126/science.aap9559

[2] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "The Spread of Low-Credibility Content by Social Bots," *Nature Communications*, Vol. 9, No. 1, 4787, 2018. https://doi.org/10.1038/s41467-018-06930-7

[3] S. Wang, T. Zhu, B. Liu, M. Ding, D. Ye, W. Zhou, and P. Yu, "Unique Security and Privacy Threats of Large Language Models: A Comprehensive Survey," *ACM Computing Surveys*, Vol. 58, No. 4, 2025. https://doi.org/10.1145/3764113

[4] E. Ferrara, "Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election," *First Monday*, Vol. 22, No. 8, 2017. https://doi.org/10.5210/fm.v22i8.8005

[5] R. Conroy, V. Rubin, and Y. Chen, "Automatic Deception Detection: Methods for Finding Fake News," *Proceedings of the Association for Information Science and Technology*, Vol. 52, No, 1, pp. 1-4, 2016.

[6] W. Y. Wang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver: Canada, pp. 422-426, 2017.

[7] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The Rise of Social Bots," *Communications of the ACM*, Vol. 59, No. 7, pp. 96-104, 2016, https://doi.org/10.1145/2818717

[8] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM 2017)*, pp. 280-289, 2017.

[9] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake News Detection on Social Media Using Geometric Deep Learning," arXiv:1902.06673, 2019. https://doi.org/10.48550/arXiv.1902.06673

[10] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," *ACM SIGKDD Explorations Newsletter*, Vol. 19, No. 1, pp. 22-36, 2017. https://doi.org/10.1145/3137597.3137600

[11] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic Detection of Fake News," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe: NM, pp. 3391-3401, 2018.

[12] N. K. Conroy, V. L. Rubin, and Y. Chen, "Automatic Deception Detection: Methods for Finding Fake News," *Proceedings of the Association for Information Science and Technology*, Vol. 52, No. 12, pp. 1-4, 2015.

[13] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen: Denmark, pp. 2931-2937, 2017. https://doi.org/10.18653/v1/D17-1317

[14] X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," *ACM Computing Surveys*, Vol. 53, No. 5, 109, 2020, https://doi.org/10.1145/3395046

[15] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A Hybrid Deep Model for Fake News Detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge*

Management (CIKM), pp. 797-806, 2017. https://doi.org/10.1145/3132847.3132877

[16] B. Horne, and S. Adali, “This Just in: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, more Similar to Satire than Real News,” in *Proceedings of the International AAAI Conference on Web and Social Media*, Montreal, Quebec: Canada, pp. 759-766, 2017.

[17] R. K. Kaliyar, A. Goswami, and P. Narang, “FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach,” *Multimedia Tools and Applications*, Vol. 80, No. 8, pp. 11765-11788, 2021. https://doi.org/10.1007/s11042-020-10183-2

김현아(Hyun-Ah Kim)

2003년 : 경기대학교 전자계산학과(이학석사)
2009년 : 경기대학교 전자계산학과(이학박사)
2018년～현 재: 경기대학교, 융합교양대학, 교양학부, 조교수
※관심분야 : 이러닝, BPM, 빅데이터, 데이터 마이닝, 머신러닝, 딥러닝 강화학습, IoT

조윤용(Yoon-Yong Cho)

2005년 : University of Missouri 언론학 (석사)
2012년 : University of Oregon 커뮤니케이션 (박사)
2020년～현 재: 경기대학교 자유교양대학 교양학부 조교수
※관심분야 : 디지털미디어, 빅데이터, 소셜미디어, 미디어리터러시 등