사물인터넷융복합논문지 Vol. 12, No. 2, pp. 95-101, 2026 https://doi.org/10.20465/KIOTS.2026.12.2.095

엣지 장치용 경량 LLM 한국어 스팸 탐지 추론 성능 비교

김동주[1], 고석주[2*]
[1]경북대학교 컴퓨터학부 박사과정, [2]경북대학교 컴퓨터학부 교수

Inference Comparison of Lightweight LLMs for Korean Spam Detection on Edge Devices

Dongju Kim[1], Seok-Joo Koh[2*]
[1]Ph.D. candidate, School of Computer Science and Engineering, Kyungpook National University
[2]Professor, School of Computer Science and Engineering, Kyungpook National University

요 약 본 연구는 Jetson Orin Nano 엣지 환경에서 한국어 스팸 문자를 실시간으로 탐지하기 위한 최적 경량 언어 모델을 선정하는 것을 목적으로 한다. 연구 방법으로는 선행연구[2]에서 구축된 KISA 신고 데이터 기반 148,937건의 법률 카테고리 데이터셋과 한국어 특유의 회피 기법을 정규화하는 4단계 전처리 파이프라인(형태소 분석, 표층 정규화, 사전 기반 치환, 토크나이징)을 활용하였다. 이를 바탕으로 Gemma3-1B, TinyLlama-1.1B, DeepSeek-1.3B 세 종의 경량 모델을 비교 평가하였으며, 정보통신망법 제50조에 근거한 4개 법률 카테고리에 대한 Macro-F1 및 카테고리별 Recall을 주요 지표로 채택하였다. 분석 결과, Gemma3-1B는 Macro-F1 0.927, '불법 행위' 카테고리 Recall 0.941을 기록하며 고위험 법률 위반 유형의 미탐지(False Negative)를 최소화하였고, INT4 양자화 후에도 Perplexity 증가율을 4.3% 이내로 유지하며 엣지 배포 가능성을 실증하였다. Gemma3-1B와 TinyLlama-1.1B 간의 성능 차이는 통계적으로 유의미하였다(paired t-test: $p = 0.0031$, Hedges' $g = 1.67$). 결론적으로 기술적 정확도, 법률 카테고리별 탐지 신뢰성, 엣지 배포 가능성의 세 기준을 종합적으로 고려할 때 Gemma3-1B가 한국어 스팸 문자 탐지 시스템 구축을 위한 최적의 모델임을 확인하였다.

주제어 : 한국어 스팸 탐지, 경량 언어 모델, 법률 기반 분류, 엣지 컴퓨팅, 전이학습

Abstract This study aims to identify the optimal lightweight language model (LLM) for real-time Korean spam SMS detection on the Jetson Orin Nano edge platform. The methodology utilizes a legal-category dataset of 148,937 samples and a specialized four-stage Korean preprocessing pipeline—comprising morphological analysis, surface normalization, dictionary-based conversion, and tokenization—established in a prior study. Three lightweight models, Gemma3-1B, TinyLlama-1.1B, and DeepSeek-1.3B, were comparatively evaluated using Macro-F1 and per-category Recall across four legal categories defined under Article 50 of the Act on Promotion of Information and Communications Network Utilization and Information Protection.Experimental results demonstrate that Gemma3-1B achieved a Macro-F1 of 0.927 and an 'Illegal Activity' category Recall of 0.941, effectively minimizing False Negatives for high-risk violations while maintaining a Perplexity increase within 4.3% after INT4 quantization. The performance gap between Gemma3-1B and TinyLlama-1.1B was statistically significant (paired t-test: $p = 0.0031$, Hedges' $g = 1.67$). These findings confirm that Gemma3-1B is the most suitable model for Korean spam detection systems, as it balances classification accuracy, legal-category reliability, and edge deployability.

Key Words : Korean Spam Detection, Lightweight Language Model, Legal-based Classification, Edge Computing, Transfer Learning

이 연구는 교육부 및 한국연구재단의 기초연구사업 지원을 통해 수행되었음 (NRF-2021R1I1A3057509).
*교신저자 : 고석주(sjkoh@knu.ac.kr)
접수일 2026년 04월 02일 수정일 2026년 04월 13일 심사완료일 2026년 04월 20일

1. 서론

스마트폰 보급의 일상화와 함께 스팸 문자의 유통이 급증하고 있다. KISA 통계에 따르면 2023년 한 해 동안 41억 2,800만 건의 스팸이 발송되었으며, 스미싱 피해 신고는 2019년 대비 8배, 금전적 피해액은 36배 증가했다[1]. 자모 분리·특수문자 삽입·은어 변형 등 한국어 특유의 회피 기법이 기존 필터링 시스템을 지속적으로 우회하면서 고도화된 탐지 기술의 필요성이 높아지고 있다. 기존 스팸 탐지 연구의 대부분은 스팸·비스팸 이진 분류에 집중하여, 정보통신망법 제50조가 규정하는 광고성 정보 명시 위반, 불법 행위 연관 전송, 투자 권유 등 세분화된 법적 위반 유형을 직접 식별하지 못한다는 한계가 있다. 또한 BERT 계열의 고성능 모델은 서버 의존적 구조로 인해 실시간 온디바이스 배포에 한계가 있어, 엣지 디바이스에서 자체적으로 동작하는 경량 탐지 모델에 대한 수요가 높아지고 있다. 이에 본 연구는 선행연구[2]에서 구축된 KISA 신고 데이터 기반 148,937건의 법률 카테고리 데이터셋과 한국어 특화 4단계 전처리 파이프라인을 활용하여, Jetson Orin Nano 엣지 환경에서 Gemma3-1B, TinyLlama-1.1B, DeepSeek-1.3B를 비교 평가하고, 법률 카테고리별 탐지 신뢰성·분류 정확도·엣지 배포 가능성을 종합적으로 충족하는 최적 경량 모델을 선정하는 것을 목적으로 한다. 2장에서 관련 연구 검토, 3장에서 비교 모델의 특성 분석, 4장과 5장에서 실험 설계 및 결과를 제시하고, 6장에서 결론을 제시한다.

2. 관련 연구

최근 스팸 탐지 연구는 딥러닝 및 LLM 기반 접근으로 빠르게 이동하고 있다. Ghourabi and Alohaly[3]는 GPT-3 임베딩과 앙상블 학습을 결합하여 트랜스포머 기반 임베딩의 탐지 성능 향상 효과를 보였고, Maqsood et al.[4]은 CNN 기반 딥러닝 프레임워크로 텍스트 전처리와 딥러닝 결합의 효과를 실증하였다. Salman et al.[5]은 다양한 LLM에 대해 zero-shot·few-shot·fine-tuning 전략을 비교하여 fine-tuning이 98.6%의 정확도를 달성하는 등 스팸 탐지에서의 LLM 우수성을 입증하였으며, Dang[6]은 LLM fine-tuning으로 단문 스팸 분류 성능을 개선하였다. 그러나 BERT 계열 모델[7]을 포함한 기존 연구 대부분은 영어 기반 이진 분류에 집중하며, 실시간 엣지 배포와 법률 위반 유형 세분화 분류 체계를 갖추지 못했다. KISA 통계에 따르면 2023년 41억 2,800만 건의 스팸이 발송되었고, 2024년 상반기 탐지량은 전년 동기 대비 68% 증가하는 등 국내 스팸 피해는 지속 심화되고 있다[1]. 자모 분리·은어 변형·특수문자 삽입 등 한국어 고유의 회피 기법은 영어권과 별도의 접근을 요구하며, Lee and Park[8]은 한국어 보이스피싱 데이터 기반 실시간 탐지 모델로 언어 특화 분석의 중요성을 입증하였다. KoBERT[9] 및 경량 한국어 모델 연구[10]는 엣지 배포 가능성을 탐구해 왔으나, 정보통신망법 제50조의 위반 유형을 라벨에 직접 반영한 법률 기반 분류 연구는 드물다. 본 연구의 기초가 된 선행연구[2]는 한국인터넷진흥원(KISA)에 신고된 스팸 데이터를 활용하여 총 148,937건의 법률 기반 데이터셋을 구축하였다. 해당 연구에서는 자모 분리 및 특수문자 삽입 등 한국어 특유의 회피 기법을 정규화하기 위해 1) 형태소 분석, 2) 표층 정규화, 3) 사전 기반 치환, 4) 토크나이징의 4단계 전처리 파이프라인을 제안하였으며, 이는 본 실험의 입력 데이터 표준화 과정에 동일하게 적용되었다.

본 연구는 이를 기반으로 경량 모델의 엣지 배포 가능성을 검증한다. 양자화·지식 증류·가지치기 등 경량화 연구가 활발히 진행되고 있으며[11,12], Frantar et al.[12]의 GPTQ는 INT4 양자화로 성능 저하를 최소화하는 본 연구의 방법론적 근거를 제공한다. Zhang et al.[13], Gemma Team[14], DeepSeek-AI[15]는 각각 TinyLlama -1.1B, Gemma, DeepSeek LLM을 공개하여 1B 규모 경량 모델의 엣지 적용 가능성을 입증하였다. Laskaridis et al.[16]과 Arya et al.[17]은 Jetson Orin에서 배치 크기·시퀀스 길이·양자화 수준이 지연 시간·처리량·Perplexity에 미치는 영향을 실증적으로 분석하였으며, INT4양자화가 소형 모델에서는 오히려 처리 속도를 저하시킬 수 있는 등 trade-off가 존재함을 지적하였다. Groß et al.[18]은 LLaMA·Gemma 등에 대한 임베딩 기반·지시 기반 fine- tuning 비교로 모델 규모와 전략의 상호작용이 중요함을 보였다. 그러나 한국어 법률 카테고리 기반 Recall을 엣지 환경에서 평가한 경량 모델 비교 연구는 아직 미흡하며, 본 연구는 이 간극을 실험적으로 검증한다.

3. 경량 언어 모델 개요

3.1 모델 등장 배경과 특징

본 연구에서 비교 평가한 모델은 Table 1과 같이 파

〈Table 1〉 Overview of Compared Models

Model	Params	Lang. Support	Role	Key Characteristics
Gemma3-1B	1B	140 languages	Evaluation target	Broad expression recognition, multilingual context processing
TinyLlama-1.1B	1.1B	English-centric	Evaluation target	Lightweight architecture, fast inference speed
DeepSeek-1.3B	1.3B	Eng&Chn-centric	Evaluation target	Specialized in structured pattern recognition
BERT-base-multilingual	110M	104 languages	Baseline	Encoder-based, high classification accuracy

라미터 규모 1B ~ 2B 사이의 경량 LLM 모델 3 종과 성능 비교 베이스라인 1 종으로 구성된다. 각 모델의 선정 기준은 한국어 지원 수준, 오픈소스 라이선스 조건, 온디바이스 최적화 가능 여부, 학술 커뮤니티에서의 검증 정도를 적용하였다.

3.1.1 대형 모델의 한계와 엣지 효율성

대형 언어 모델은 서버급 GPU 환경에서 강력한 성능을 발휘하지만, 엣지 디바이스에서는 메모리 및 전력 제한으로 FP32 정밀도 그대로 배포하기 어렵다. 파라미터 규모가 증가하면 추론 시간이 늘어 실시간 처리에 부적합하여, INT4 양자화를 통해 2GB 이상 모델 메모리를 0.5 GB 수준으로 줄이는 것이 엣지 효율성 확보의 대표적인 방법이다. 한국어 스팸 탐지 도메인에서는 경량화의 필요성이 더욱 크다. 한국어 특유의 형태소 구조와 자모 결합 규칙은 변형 가능성이 높아 로컬에서 빠른 전처리와 패턴 복원이 요구되며, 대형 다국어 모델은 비한국어 지식이 자원을 불필요하게 점유한다. 반면 Gemma3-1B처럼 다국어 지원 폭이 넓으면서도 경량 구조를 유지한 모델은 한국어 전용 전처리와 결합 시 탐지 정확도와 엣지 효율성을 동시에 달성할 수 있음이 본 실험에서 확인되었다(Hedges' $g = 1.67$). 법률 기반 분류에서도 대형 모델의 한계가 드러난다. 세분화된 국내 법률 카테고리 학습 시 외부 데이터 영향을 과도하게 받아 분류 오류가 발생할 수 있는 반면, 경량 모델은 KISA 신고 사례 기반 데이터셋으로 학습 범위를 제한함으로써 오류 가능성을 낮출 수 있다. 그 중 Gemma3-1B가 전이학습 효과, 법률 카테고리 분류 정확도, 엣지 배포 가능성 면에서 가장 균형 있는 성능을 보여주었다.

〈Table 2〉 Gemma3-1B vs. TinyLlama-1.1B

Comparison Item	Gemma3	TinyLlama
Parameters	1B	1.1B
Language Support	140 langs	Eng-centric
Macro-F1	0.927	0.901
Overall Accuracy	95.5%	94.2%
Recall- Illegal Activity	0.941	Relatively low
Recall- Investment Solicitation (post-preprocessing)	0.908	0.879
Precision — Legitimate Ad	—	0.955 (strong FP suppression)
Category Confusion Rate	~13%	~15%
Memory after Quantization	0.5 GB	0.5 GB
Perplexity Increase Rate	~4%	~4%
Preprocessing Integration Latency	Slight (mitigable)	None
Update Convenience	Moderate	High
Legal Category Suitability	High	Moderate

3.1.2 Gemma3-1B와 TinyLlama-1.1B 비교

Gemma3-1B와 TinyLlama-1.1B는 파라미터 규모가 유사하여 엣지 배포 가능성 측면에서 비슷하지만, 전이학습 특성과 법률 카테고리 대응 능력에서 차이를 드러낸다. 두 모델의 비교는 Table 2에 정리하였다. Gemma3 -1B는 140개 언어를 지원하는 광범위한 표현 인식력을 바탕으로 투자 권유와 불법 행위 간 법률 카테고리 혼동을 효과적으로 억제한다(오분류율 약 13%). TinyLlama- 1.1B는 추론 속도와 업데이트 편의성에서 우위를 가지나 고위험 카테고리 Recall과 혼동 억제 측면에서 한계가 있다. 전처리 후에도 Gemma3-1B의 카테고리별 Recall이 높게 유지되었으며(Macro-F1 차이: $p = 0.0031$, Hedges' $g = 1.67$), 분류 적합성과 고위험 Recall 유지력을 종합하면 Gemma3-1B가 한국어 스팸 탐지의 모델로 적합하다.

〈Table 3〉 Four-Stage Preprocessing Pipeline

Stage	Processing Content	Tool / Dictionary	Detailed Operations
1	Morphological Analysis	KoNLPy Mecab	Particle & ending separation, compound noun processing
2	Surface Normalization	Rule-based	Special character removal (♣ ★ ▶etc.), shortened URL normalization (bit.ly, naver.me)
3	Dictionary -based Conversion	Slang/Abbrev.Dict.(327) LegalKeywordDict.(156) JamoRestorationDict.(89)	Non-standard expression standardization, legal violation keyword mapping, jamo-split text restoration
4	Tokenizer Application	Gemma3 Tokenizer	Conversion to model input format

〈Table 4〉 Key Hyperparameters

Hyperparameter	Value	Notes
Learning Rate	5×10^{-5}	Candidates:1×10^{-5}, 5×10^{-5}, 1×10^{-4}
Batch Size	8	—
Dropout	0.1–0.3	—
Warmup Steps	500	Adjusted for Jetson memory constraints
Loss Function	Focal Loss	γ=2.0
Focal Loss α	[0.5, 1.5, 2.0, 1.8]	Inverse class frequency-based
Quant. Method	INT4 (GPTQ)	Per-channel quantization
Weight Clipping	$[-3\sigma, +3\sigma]$	Outlier suppression
Calibration Data	10,000 samples	Class-balanced sampling

4. 모델 성능 비교 실험 설계

4.1 전이학습 접근

4.1.1 도메인 적응 및 전처리 효과

전이학습은 선행연구[2]에서 구축한 KISA 신고 데이터 기반 148,937건의 4개 카테고리 데이터셋을 활용했다. 이는 실제 규제기관 신고 사례를 기반으로 법률 라벨링이 명확히 부여되어 도메인 특화 학습 요건을 충족한다. 전처리는 선행연구[2]의 4단계 파이프라인(Table 3)을 적용했으며, 은어·변형 단어로 위법성이 숨겨진 메시지의 표준화하여 모델이 카테고리를 정확히 판별한다.

하이퍼파라미터는 Bayesian Optimization으로 도출하였으며(Table 4), 클래스 불균형 대응을 위해 Focal Loss를 적용하였다. INT4 양자화 후에도 Perplexity 증가는 약 4.3%로 억제되어 성능 안정성이 유지되었다. 모델별 전이학습 효과를 살펴보면, Gemma3-1B는 전처리 결합 후 '투자 권유' Recall이 0.908까지 상승하고 불법 행위 오분류율이 가장 낮아 도메인 적응 시너지가 두드러졌다. TinyLlama-1.1B는 업데이트 적용 시간이 짧

〈Table 5〉 Overall Model Performance Comparison

Model	Acc (%)	Macro-F1		MAT (s)	Notes
		pre-pre processing	post-pre processing		
Gemma3 -1B	95.5	0.927	—	0.29	Highest overall Recall across legal categories
TinyLlama -1.1B	94.2	0.901	—	0.23	Strong FP suppression
DeepSeek -1.3B	94.8	0.893	+3% improvement	0.31	Large effect from special char. & URL normalization
BERT-base-multilingual	96.1	0.941	—	1.85	Baseline; inference time too high for edge deployment

〈Table 6〉 Per-Category Performance Metrics

Category	Metric	Gemma3 -1B	TinyLlama -1.1B	DeepSeek -1.3B
Illegal Activity	Recall	0.941	0.897	—
Illegal Activity	Precision	0.897	—	—
Investment Solicitation	Recall	0.908	0.879	—
Legitimate Advertisement	Precision	—	0.955	—
Non-compliant Advertisement	F1	—	—	0.927
CategoryConfusionRate (Illegal↔Investment)	—	~13%	~15%	—

으나 고위험 카테고리 Recall 향상 폭이 작았고, DeepSeek-1.3B는 정형 패턴 탐지에 강하지만 자모 분리 기반 회피에 추가 전처리가 필요했다.

4.1.2 정확도와 F1-Score 평가

선행연구[2]에서 구축된 4개 법률 카테고리 데이터셋을 테스트셋에 적용한 모델별 성능은 Table 5와 같다.

카테고리별 세부 지표는 Table 6과 같다. Gemma3-1B는 '불법 행위'에서 Recall 0.941, Precision 0.897로 False Negative 비율이 가장 낮았다. TinyLlama-1.1B는 '합법 광고' Precision 0.955로 FP 억제에 강점을 보였으나 '투자 권유' Recall이 0.879로 고위험 탐지에서 한계를 드러냈다. DeepSeek-1.3B는 '광고 비준수' F1=0.927로 높았지만 변형 메시지에서 성능 저하가 두드러졌다.

전처리 적용 후에도 Gemma3-1B의 법률 카테고리별 Recall이 전반적으로 가장 높게 유지되었으며, Macro-F1 차이는 $p=0.0031$, Hedges' $g=1.67$로 통계적으로 유의했다. 실험 재현성 확보를 위해 PyTorch 환경에서 Random Seed=42, 100회 반복 측정으로 추론 시간·메모리·에너지 소비 평균값과 95% 신뢰구간을 산출했다.

〈Table 7〉 Real-Time Inference Performance

Model	TPS (msg/sec)	MAT (sec)	GPU Memory	CPU Usage	Bottleneck
TinyLlama-1.1B	87	0.23	0.5 GB	Low	None
DeepSeek-1.3B	79	0.31*	0.5 GB	Moderate	jamo restoration
Gemma3-1B	74	0.29	0.5 GB	Slightly high	Minor

*Measured including text-variant restoration stage

〈Table 8〉 Model Selection Matrix

Selection Criteria	Priority	Gemma3-1B	TinyLlama-1.1B	DeepSeek-1.3B
Legal Category Recall Retention	①Highest	Best (Illegal 0.941, Investment 0.908)	Investment solicitation limited at 0.879	—
Inter-category Confusion Suppression	②	Confusion rate ~13%	Confusion rate ~15%	—
Korean Evasion Technique Handling	③	Stable Macro-F1 gain after preprocessing	Moderate	Requires additional preprocessing
Edge Deployment Feasibility	④	TPS 74, MAT 0.29 s (meets requirement)	TPS 87, MAT 0.23 s (best speed)	MAT 0.31 s (with variant restoration)
Overall Verdict		**Selected**	**Speed advantage; legal detection limitation**	**Extra preprocessing required for evasion handling**

5. 한국어 스팸 탐지에 경량 모델 적용

5.1 실시간 처리 성능

실시간 처리 성능은 실효성을 좌우하는 핵심 요소로, 엣지 환경에서는 추론 속도와 메모리 효율이 분류 정확도와 직결된다. 세 모델 모두 INT4양자화로 0.5GB로 경량화하였으며(Perplexity 증가 약 4.3%), end-to-end 추론 시간 측정 결과는 Table 7과 같다. 본 연구에서는 전처리 연동에 따른 지연 시간을 구체적으로 파악하기 위해, 전처리 파이프라인의 단계별 소요 시간 비중을 프로파일링하였다. 분석 결과, Stage 1 형태소 분석이 Mecab 엔진 호출 및 복합명사 처리 부하로 인해 약 42%로 가장 높은 비중을 차지하고, 자모 복원 및 은어 사전 매핑이 수행되는 Stage 3 사전 기반 치환이 약 35%, Stage 4 토크나이징과 Stage 2 표층 정규화가 각각 15%와 8%를 기록하였다. 특히 Gemma3-1B의 경우, 다국어 어휘 처리 구조와 Stage 3의 연동 과정에서 소폭의 오버헤드가 발생하나, 멀티쓰레딩 기반 병렬 처리를 통해 실시간 차단 시스템의 요구 요건(MAT 0.29s) 내에서 충분히 완화 가능함을 확인하였다.

TinyLlama-1.1B는 경량 구조 덕분에 가장 짧은 추론 시간을 기록했고, DeepSeek-1.3B는 자모 복원 연동 단계에서 병목이 발생했다. Gemma3-1B는 전처리 연동 시 약간의 지연이 있으나 멀티쓰레딩으로 완화 가능하며, 불법 행위 범주 Recall 유지력이 가장 높아 실시간 환경에서도 안정적이었다. TPS 74건·MAT 0.29초는 실시간 차단 시스템의 최소 처리량 요건을 충족한다.

5.2 모델 선택 기준

모델 선택 기준은 운영 환경과 분류 체계의 요구사항을 동시에 만족하는 방향으로 우선순위에 따라 네 가지를 적용했으며, 모델별 평가 결과는 Table 8과 같다.

첫째, 법률 카테고리별 Recall 유지력이 최우선 기준이다. 고위험 카테고리 False Negative는 법적 책임 위험을 직접 높이므로, Gemma3-1B의 '불법 행위' Recall 0.941, '투자 권유' Recall 0.908은 경쟁 모델 대비 가장 높은 수준이다. 둘째, 법률 카테고리 간 혼동 억제 능력이다. Gemma3-1B는 혼동율 약 13%로 TinyLlama-1.1B (약 15%)보다 낮으며, 이는 광범위한 표현 패턴 이해력에서 기인한다. 셋째, 한국어 회피 기법 대응 능력이다. 자모 분리·특수문자 삽입·은어 치환 등 변형 기법 대응을 위해 전처리 파이프라인과 모델 어휘 처리 구조의 연동이 필요하며, Gemma3-1B는 전처리 전후 Macro-F1 향상 폭이 안정적으로 유지되었다. 넷째, 엣지 배포 가능성이다. INT4 양자화로 메모리 0.5 GB, Perplexity 증가율 4.3% 이내를 달성하며 Gemma3-1B는 이 조건을 충족한다.

이상을 종합하면 Gemma3-1B가 한국어 스팸 탐지 시스템의 핵심 모델로 가장 적합하다. TinyLlama-1.1B는 추론 속도에서 우위를 가지나 고위험 Recall과 혼동 억제에서 한계가 있고, DeepSeek-1.3B는 변형 회피 대응에 추가 전처리가 필요하여 처리 지연이 발생한다.

6. 결론

본 연구는 KISA 신고 데이터 기반 148,937건의 법률 카테고리 데이터셋과 한국어 특화 전처리 파이프라인을 활용하여, Jetson Orin Nano 엣지 환경에서 경량 언어 모델의 한국어 스팸 문자 탐지 적합성을 비교 평가하고 최적 모델을 선정했다. 세 모델은 고유한 장단점을 보였다. TinyLlama-1.1B는 추론 속도(TPS 87건, MAT 0.23초)에서 우위를 보였고, DeepSeek-1.3B는 정형 패

턴 탐지에 강하나 언어 변형 대응에 추가 전처리가 필요했다. Gemma3-1B는 전체 정확도 95.5%, Macro-F1 0.927로 가장 높은 분류 성능을 달성했다. Gemma3-1B는 법률 카테고리별 Recall 유지력에서 결정적 강점을 보였다. '불법 행위' Recall 0.941, '투자 권유' Recall 0.908로 False Negative를 최소화했고, 카테고리 혼동 비율 약 13%로 법적 판정 신뢰성에서도 우위를 보였다 (p=0.0031, Hedges' g=1.67). 이에 본 연구는 Gemma3-1B를 한국어 스팸 문자 탐지 시스템의 모델로 선정한다. Gemma3-1B는 기술적 정확도, 법률 카테고리 탐지 신뢰성, 엣지 배포 가능성을 균형 있게 충족하며 사용자 보호와 법적 준수를 동시에 달성할 수 있는 실용적 기반을 제공한다.

향후 연구는 회피 기법 대응 확장, 실시간 업데이트 통합, 다중 위반 라벨링 구조 확장을 통해 시스템 강건성을 높이는 방향으로 진행될 수 있다.

REFERENCES

[1] KISA, 2024 First Half Spam Distribution Status Report[Internet], https://www.kisa.or.kr/20504

[2] D.Kim, "Development of Dataset and Preprocessing Pipeline for Legal-Based Classification of Korean Spam SMS," in Proceedings of the Korea Internet of Things Society. Conference, Vol.10, No.1, pp.61-63, 2025.

[3] A.Ghourabi and M.Alohaly, "Enhancing Spam Message Classification and Detection Using Transformer-Based Embedding and Ensemble Learning," Sensors, Vol.23, No.8, Article 3861, 2023.

[4] U.Maqsood, S. U.Rehman, T.Ali, K.Mahmood, T.Alsaedi and M.Kundi, "An Intelligent Framework Based on Deep Learning for SMS and E-mail Spam Detection," Applied Computational Intelligence and Soft Computing, Vol.2023, Article ID 6648970, 16 pages, 2023.

[5] M.Salman, M.Ikram, N.Basta and D.Kaafar, "SpaLLM-Guard: Pairing SMS Spam Detection Using Open-source and Commercial LLMs," arXiv:2501.04985, 2025.

[6] Q.V.Dang, "Fine-Tuning LLMs to Detect SMS Spam," in Proceedings of the International Conference on Information and Communication Technology for Competitive Strategies, Vol.1320, pp.457-466, Springer, Singapore, 2025.

[7] J.Devlin, M.W.Chang, K.Lee and K.Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1, pp.4171-4186, 2019.

[8] M.Lee and E.Park, "Real-time Korean Voice Phishing Detection Based on Machine Learning Approaches," Journal of Ambient Intelligence and Humanized Computing, Vol.14, No.7, pp.8173-8184, 2023.

[9] SKTBrain, KoBERT: Korean BERT Pre-trained Cased, [Internet], https://github.com/SKTBrain/KoBERT

[10] J.H.Kim and Y.S.Choi, "Lightweight PreTrained Korean Language Model Based on Knowledge Distillation and Low-Rank Factorization," Entropy, Vol.27, No.4, Article 379, 2025.

[11] R.Wang,Z.Gao,L.Zhang,S.Yue and Z.Gao, "Empowering Large Language Models to Edge Intelligence: A Survey of Edge Efficient LLMs and Techniques," Com- puter Science Review, Vol.57, Article 100755, 2025.

[12] E.Frantar,S.Ashkboos,T.Hoefler and D.Alistarh, "GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers," Proceedings of the 11th International Conference on Learning Representations (ICLR), 2023.

[13] P.Zhang, G.Zeng, T.Wang and W.Lu, "TinyLlama: An Open-Source Small Language Model," arXiv:2401.02385, 2024.

[14] Gemma Team: T.Mesnard et al, "Gemma: Open Models Based on Gemini Research and Technology," arXiv:2403.08295, 2024.

[15] DeepSeek-AI: X.Bi et al, "DeepSeek LLM: Scaling Open-Source Language Models with Longtermism," arXiv:2401.02954, 2024.

[16] M.Arya and Y.Simmhan, "Understanding the Performance and Power of LLM Inferencing on Edge Accelerators," arXiv:2506.09554, 2025.

[17] L.Seymour, B.Kutukcu and S.Baidya, "Large Language Models on Small Resource-Constrained Systems: Performance Characterization, Analysis and Trade-offs," arXiv:2412.15352, 2024.

[18] A.Yousefiramandi and C.Cooney, "Fine-Tuning Causal LLMs for Text Classification: Embedding-Based vs. Instruction-Based Approaches," arXiv:2512.12677, 2024.

김 동 주(Dongju Kim) [정회원]

- 1999년 2 : 경북대학교 물리학과 이학사
- 2011년 2월 : 대구가톨릭대학교 컴퓨터정보통신공학과 공학석사
- 2021년 3월 ~ 현재 : 경북대학교 컴퓨터학부 공학박사수료
- 2020년 3월 ~ 현재 : 대구가톨릭대학교 컴퓨터소프트웨어학부 교수

〈관심분야〉
Game, AI, Mobile, IoT, Edge computing

고 석 주(Seok-Joo Koh) [정회원]

- 1992년 2월 : KAIST 경영과학과 공학사
- 1994년 2월 : KAIST 경영과학과 공학석사
- 1998년 8월 : KAIST 산업공학과 공학박사
- 1998년 9월 ~ 2004년 2월 : ETRI 표준연구센터 선임연구원
- 2004년 3월 ~ 현재 : 경북대학교 컴퓨터학부 교수

〈관심분야〉
IoT, multicast, SCTP, QUIC, Visible Light Comm.