멀티태스크 LLM 기반 정책 반영 여부 예측 성능 향상을 위한 랜덤포레스트 활용에 관한 연구

이동섭01, 김영재1
1서강대학교 AI·SW대학원
ldsstory@naver.com, youkim@sogang.ac.kr

Multi-task LLM-based Method for Policy Adoption Prediction with Random Forest

Dongseop Lee[01], Youngjae Kim[1]
[1]Graduate School of AI·SW, Sogang University

요 약

이 연구는 멀티태스크 LLM 기반 중소기업 정책 반영 여부 예측 모델의 성능을 향상하는 방법을 제안한다. LLM만 활용할 경우, 정책 건의별 토큰 수 불균형과 표현 다양성 및 데이터 부족에 의하여 예측 성능에 한계가 있었다. 이를 개선하기 위해 메타데이터를 랜덤포레스트로 학습시켜 예측값과 변수중요도를 도출하고, 데이터 특성으로써 프롬프트에 포함했다. 이후, 공유된 특성으로 분류와 회귀를 동시에 학습하는 Feature Transformation 기반 멀티태스크 LLM을 미세조정했다. A100 GPU(40.0GB VRAM) 환경에서 실험한 결과 Accuracy는 72.14%, F1-score는 71.13%로 나타났다. 이는 이전 실험 대비 각각 14.55%p, 26.54%p 향상된 수치로, 정량적인 성능 개선을 확인할 수 있었다.

1. 서론

중소기업은 경제 발전과 일자리 창출의 핵심이다. 그리고 중소기업을 지원하기 위한 정책이 있으며, 성공적으로 반영되거나, 입안에 어려움을 겪거나, 미반영된다. 이러한 차이는 정책 내용, 제안 방식과 시기 등 여러 요인이 복합적으로 작용한 결과이다. 따라서 정책 반영 여부를 사전에 예측하는 것은 정책 기획 과정에서 자원과 노력을 집중시켜 정책 실현 가능성을 높이는 데에 중요한 역할을 한다.

선행 연구들은 정책 반영 여부에 대하여 회귀 기반의 예측 모델 등 통계적 분석[1]과 머신러닝 모델[2]에 한정되었다. 이는 선형 관계와 정형 변수에만 기반하므로 정책의 내재된 의미 반영에는 한계가 있다.

최근 사회과학 실험 결과 예측[3] 등 다양한 분야에 LLM의 적용 가능성이 제시되었다. 따라서 정책 반영 여부 예측에 LLM 적용을 시도하였으며, 사전지식 활용과 비선형 구조 학습으로 회귀분석이 놓쳤던 정성적·복합적 분석이 가능했다. 그러나 토큰 수 불균형, 표현 차이, 데이터 부족 등으로 인해 예측 성능이 제한되는 한계가 있었다.

그래서 본 논문에서는 LLM 기반 정책 반영 예측의 한계를 극복하기 위해 메타데이터와 랜덤포레스트 기법을 결합한 모델을 제안한다. 우선 랜덤포레스트로 도출한 예측값과 변수중요도를 데이터 특성[4]으로써 LLM 학습을 위한 프롬프트에 포함한다. 그리고 분류 및 회귀에 대한 멀티태스크 LLM[5]을 QLoRA[6]를 통한 미세조정으로 학습한다. 분류 태스크는 Focal Loss[7]를 적용하여 데이터 불균형 완화 및 어려운 문제에 집중하고, 손실 가중치를 회귀 태스크보다 높게[8] 한다. 그 결과 모델의 예측 성능이 안정적으로 향상됨을 확인했다.

2. 연구배경 및 선행연구

2-1. 변수에 관한 조작적 정의

정책결정기간과 법률안 통과 여부[1]에 관한 연구에 따르면 독립변수로는 발의시기, 주요 선거, 상임위 구성, 발의 주체, 정책 유형 등이 있고, 종속변수로는 상임위 의결기간, 법률안 가결, 입법결과 등이 있다. [표1]은 이러한 변수들을 기반으로 정책건의백서에서 정의한 변수들이다.

표 1: 학습 데이터 변수 정의

분류		변수명	조작적 정의
독립	메타	구분	건의 방법
		분야	정책이 속한 영역
		건의횟수	건의한 횟수
		건의처(정부)	건의 대상 중 행정부
		건의처(국회)	건의 대상 중 입법부
		여당	건의 시점의 여당
		다수당	건의 시점의 다수당
		주요 선거	건의 시점의 주요 선거
		국회 기수	건의 시점의 국회 기수
		전후반기	건의 시점의 국회 전후반기
	text	현황 및 문제점	건의한 원인
		건의사항	건의 내용
종속	메타	건의결과	반영 여부

2-2. 랜덤포레스트와 멀티태스크 LLM

랜덤포레스트는 의사결정나무마다 훈련 샘플과 특성을 무작위로 선택하여 과적합을 방지한다. 그리고 변수중요도를 통해 결과에 주요 영향을 미친 요인을 파악할 수 있어, 안정적인 예측 도구로 활용되고 있다. 이미지 기반 CNN과 메타데이터 기반 랜덤포레스트를 결합한 멀티태스크 모델은 랜덤포레스트를 통해 예측 성능을 향상시켰다[4].

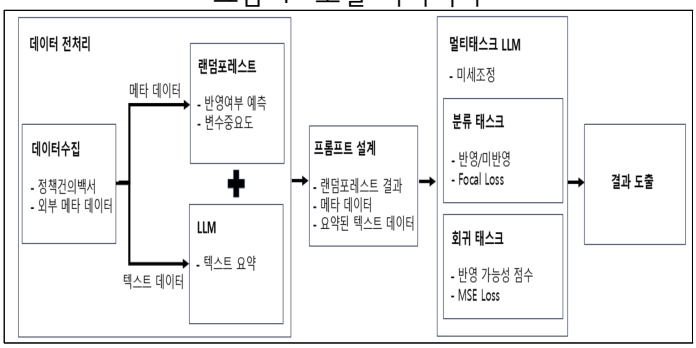
또한, LLM은 Attention 메커니즘을 기반으로 긴 문맥 정보를 이해하고, 사전지식과 프롬프트를 활용해 논리적

추론을 수행하는 데에 강점이 있다. GPT-4는 사회과학 실험 결과 예측에 대하여 높은 상관관계(r=0.85)를 보였으며, 이는 정책 반영 여부 예측에도 LLM이 활용될 수 있음을 시사한다[3]. 그리고 멀티태스크 학습은 태스크 간 지식공유와 손실 가중치 조절을 통해, 모델의 학습 효율을 높이고 일반화 성능을 강화하는 것으로 알려져 있다[5, 8].

3. 설계 및 구현

3-1. 모델 설계와 데이터 전처리

그림 1: 모델 아키텍처



[그림1]은 모델의 전체 구조이며, 세 단계로 구성된다. 먼저, 텍스트 데이터 및 메타데이터를 각각 LLM을 통한 전처리와 랜덤포레스트 기반 분석에 활용한다. 다음으로, 전처리된 텍스트 데이터와 랜덤포레스트 결과값을 프롬프트에 포함하여 학습 데이터로 활용한다. 마지막으로, 멀티태스크 LLM은 분류와 회귀를 동시에 미세조정한다.

데이터는 중소기업중앙회에서 발간한 정책건의백서(6개년)에서 수집한 총 2,152건의 정책 건의 사례를 사용했다. 건의결과는 반영 599건(27.8%), 일부반영 314건(14.6%), 검토중 661건(30.7%), 미반영 578건(26.9%)으로 구분된다. 분석의 효율을 위해 반영 및 일부반영은 '반영'으로, 검토중 및 미반영은 '미반영'으로 이분했다. 텍스트 길이의 편차가 커 토큰 수의 불균형이 심하므로, LLM을 활용한 텍스트 요약으로 토큰 수 편중을 완화하는 전처리를 했다.

3-2. 랜덤포레스트 및 멀티태스크 LLM 구조

랜덤포레스트는 RandomForestClassifier을 통해 [표1]의 메타데이터에 대한 이진 분류 모델로 학습했다. 랜덤포레스트 구성 및 파라미터 값은 [표2]와 같다. 변수중요도는 각 변수의 Gini Impurity 감소량을 평균하여 계산되었으며, 이를 통해 영향력이 높은 변수를 도출했다.

표 2: 랜덤포레스트 구성 및 주요 파라미터

구분	설정값
n_estimators	100
max_depth	None(제한 없음)
bootstrap	True
max_features	sqrt(제곱근)
criterion	gini(Gini Impurity)

멀티태스크 LLM은 beomi/KoAlpaca-Polyglot-12.8B를 활용했다. Polyglot-ko-12.8b을 기반으로 KoAlpaca 데이터셋으로 미세조정된 한국어 특화 모델로, 명령어 지시문 처리 및 자연어 추론에 강점을 보인다.

[표3]은 멀티태스크 LLM 구성 및 파라미터 값이다. 데이터 feature를 공유하는 멀티태스크 학습을 통해 분류(반영, 미반영)와 회귀(반영=1, 일부반영=0.75, 검토중=0.5, 미반영=0)를 동시에 미세조정했다. Focal Loss로 데이터 클래스 불균형 조정(α)과 학습 시 어려운 문제에 집중(γ)했으며, 손실 가중치를 조절하여 분류 학습(90%)에 집중했다. 그리고 QLoRA로 학습과 메모리 효율성을 높였다.

표 3: 멀티태스크 LLM 구성 및 주요 파라미터

구분	설정값
Feature 추출	출력의 마지막 히든 스테이트 평균값(mean pooling)
Dropout	과적합 방지(확률 0.1)
분류 헤드	Linear(hidden_size → 2)
회귀 헤드	Linear(hidden_size → 1) + Sigmoid
손실함수	분류: Focal Loss(α=[0.42, 0.58], γ=2), 회귀: MSE, 전체손실 = 0.9*분류손실 + 0.1*회귀손실
양자화	4-bit NF4, bfloat16, double quant enabled
최적화 전략	weight decay=0.01, cosine schedule, warmup ratio=0.15
배치 및 누적	batch_size=2, gradient_accumulation_steps=8
학습 에폭	13 epoch
학습률	2e-5

4. 실험 결과 및 분석

4-1. 랜덤포레스트 결과

실험은 Google Colaboratory Pro+의 NVIDIA A100 GPU (40.0GB VRAM) 환경에서 수행되었다. 랜덤포레스트 실험 결과 [표4]에서처럼 선행연구에서의 중요 변수인 주요 선거 등 정치적 요인보다 건의처 등 실무적 요인이 더 중요한 것으로 확인되었다. 도출한 결과는 [그림2]와 같이 프롬프트에 반영했다.

표 4: 변수중요도 결과

정부	분야	국회/정당	건의 횟수	구분	주요 선거	전후반기
0.3657	0.2520	0.1003	0.0848	0.0679	0.0407	0.0379

그림 2: 프롬프트 예시

```
[랜덤포레스트 예측 결과] 반영 (신뢰도: 0.64)
[변수중요도]
건의처(정부): 0.3657
[이 건의는 다음과 같은 상황입니다]
구분: 기술/벤처
[현황 및 문제점입니다]
- R&D 수요와 필요성은 높으나 인력, 자금, 역량 부족
[현황 및 문제점에 따라 제안된 건의사항입니다]
- 중소기업간 협업을 통한 업종 공통 R&D 지원
```

4-2. 멀티태스크 LLM 결과

멀티태스크 LLM 실험 결과 [표5]에서처럼, 미세조정 과정에서 에폭이 증가할수록 성능이 점차 개선됨을 확인했다. 학습손실과 검증손실이 안정적으로 감소하였으며, 과소/과대 적합 없이 잘 학습되었다. 분류 태스크의 성능은 [표7]에서처럼, Accuracy 72.14%, Precision 72.31%, Recall 72.14%, F1-score 71.13%로 전체적으로 준수한 결과를 보였다.

특히 [표9]를 보면 소수 클래스에 대한 과도한 성능 저하 없이 반영과 미반영 클래스 모두 양호한 성능을 보임을 알 수 있다. 이는 Focal Loss가 소수 클래스의 분류 성능 보완에 효과가 있었음을 뒷받침한다. 또한 [그림3]에서 확인할 수 있듯이 랜덤포레스트 결과값을 프롬프트에 포함한 모델의 분류 성능은 그렇지 않은 모델보다 Accuracy 14.55%p, Precision 21.75%p, Recall 14.55%p, F1-score 26.54%p 더 높았다.

다만, [표8]과 같이 회귀 태스크의 성능은 전반적으로 낮았다. 예측된 회귀 점수는 실제 반영 점수와의 통계적 상관성이 거의 없는 것으로 나타났다. 데이터 부족, 임의로 지정한 회귀 라벨의 불연속성 등 데이터 자체 및 외적 요인의 영향으로 보이며, 정확한 수치 예측보다는 분류 태스크의 보조 역할로 활용하는 것이 적절하였다.

표 5: 멀티태스크 LLM 미세조정 추이

에폭	학습 손실	검증 손실	Accuracy	F1-score	MSE	R^2
1	0.1039	0.0936	0.5776	0.4624	0.1508	-0.0522
5	0.0873	0.0787	0.6988	0.6471	0.1406	0.0188
8	0.0755	0.0747	0.7205	0.7149	0.1306	0.0886
11	0.0740	0.0744	0.7484	0.7460	0.1289	0.1005
12	0.0753	0.0743	0.7453	0.7426	0.1289	0.1002
13	0.0753	0.0742	0.7453	0.7430	0.1288	0.1008

표 6: 멀티태스크 LLM 혼동행렬

구분	예측 반영	예측 미반영
실제 반영	71개(0.5259)	64개(0.4741)
실제 미반영	26개(0.1383)	162개(0.8617)

표 7: 멀티태스크 LLM 분류 성능평가

Accuracy	Precision	Recall	F1-score
0.7214	0.7231	0.7214	0.7113

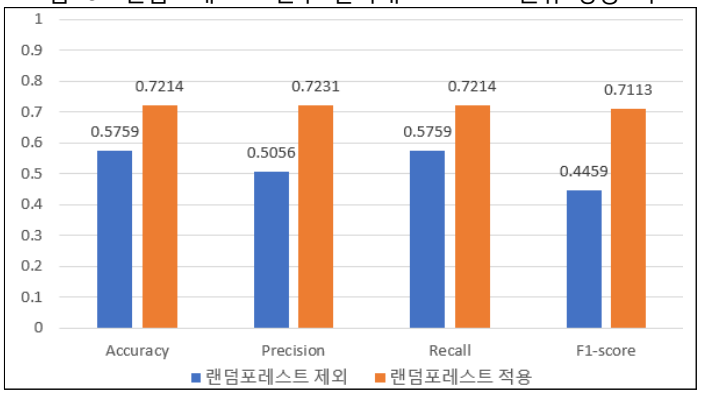
표 8: 멀티태스크 LLM 회귀 성능평가

RMSE	R2	Pearson 상관계수(p=4.4693e-01)
0.3758	-0.0227	0.0425

표 9: 멀티태스크 LLM 건의 결과별 성능평가

구분	Precision	Recall	F1-score	Support
반영	0.7320	0.5259	0.6121	135
미반영	0.7168	0.8617	0.7826	188

그림 3: 랜덤포레스트 전후 멀티태스크 LLM 분류 성능 비교



5. 결론

본 논문에서는 정책 반영 여부를 예측하기 위해 멀티태스크 LLM을 설계하고, 학습을 위한 입력 프롬프트에 랜덤포레스트 예측 결과와 변수중요도 정보를 포함하여 성능을 향상하는 방법을 제안하였다. 실험 결과, Accuracy는 57.59%에서 72.14%로, F1-score는 44.59%에서 71.13%로 성능이 크게 향상되었다. 이를 통해 랜덤포레스트가 멀티테스크 LLM 모델의 성능 향상에 기여함을 정량적으로 확인하였다.

참고문헌

[1] 서인석, 박형준, 권기헌, "정책유형과 정책대상집단에 따른 정책결정 소요시간: 발의 법률안의 통과기간의 영향요인 탐색연구", 한국행정학보, 47(2), 55-83, 2013.

[2] 송태민, "소셜빅데이터를 활용한 보건복지정책 동향 분석", 보건복지포럼, 213:101, 2014.

[3] Hewitt, L., Ashokkumar, A., Ghezae, I., & Willer, R., "Predicting results of social science experiments using large language models", arXiv preprint arXiv:2408.08477, 2024.

[4] Kasmanoff, N., Lee, M. D., Razavian, N., & Lui, Y. W., "Deep multi-task learning and random forest for series classification by pulse sequence type and orientation", Neuroradiology, 65, 77-87, 2023.

[5] Zhang, Y., & Yang, Q., "A survey on multi-task learning", IEEE Transactions on Knowledge and Data Engineering, 34(12), 5586-5609, 2021.

[6] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L., "QLoRA: Efficient Finetuning of Quantized LLMs", arXiv preprint arXiv:2305.14314, 2023.

[7] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P., "Focal loss for dense object detection", In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2999-3007, 2017.

[8] Kendall, A., Gal, Y., & Cipolla, R,. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7482-7491), 2018.