

Pretrained Model for Civil Complaint Classification

Jongchan Lee*, Junha Hwang**, Young-Seob Jeong***

*Student, Dept. of Computer Engineering, Chungbuk National University, Cheongju, Korea

**Student, Department of Electrical·Electronic·Information and Computer Engineering, Chungbuk National University, Cheongju, Korea

***Professor, Department of Electrical·Electronic·Information and Computer Engineering, Chungbuk National University, Cheongju, Korea

[Abstract]

With the expansion of online administrative services and the growing volume of civil complaints, automated complaint classification has become increasingly important. However, previous studies have mainly focused on fine-tuning general-purpose language models using labeled data, while the effectiveness of Domain-Adaptive Pretraining (DAPT) using large-scale unlabeled Korean civil complaint and administrative corpora has not been sufficiently investigated. In this study, we constructed a corpus of approximately 137.9 million tokens and applied MLM-based DAPT to A.X-Encoder-base. We also constructed a civil complaint classification dataset consisting of 116,042 queries across 17 categories and compared performance according to the application of DAPT and different training configurations under identical conditions. Experimental results showed that the model trained for 1.25 epochs at a learning rate of 1×10^{-4} achieved an accuracy of 73.4% and a Macro F1 score of 72.5%, improving upon the base model by 0.4 and 0.3 percentage points, respectively. These results suggest the importance of domain adaptation in the civil complaint domain.

▶ **Key words:** Korean Civil Complaint Classification, Domain-Adaptive Pretraining, Korean Language Models, Masked Language Modeling, Domain-Specific Language Models

[요 약]

온라인 행정 서비스의 확산과 민원 규모의 증가에 따라 민원 분류 자동화의 중요성이 높아지고 있다. 그러나 기존 연구는 주로 라벨 데이터를 활용한 범용 언어모델의 미세조정에 집중해 왔으며, 대규모 비라벨 한국어 민원·행정 말뭉치를 활용한 도메인 적응 사전학습(DAPT)의 효과는 충분히 검증되지 않았다. 본 연구에서는 약 1억 3,790만 토큰의 말뭉치를 구축하고 A.X-Encoder-base에 MLM 기반 DAPT를 적용하였고, 17개 분야의 116,042건의 민원 분류 데이터셋을 구축하여 동일한 조건에서 DAPT 적용 여부와 학습 설정에 따른 성능을 비교하였다. 실험 결과, 학습률 1×10^{-4} 에서 1.25에포크 학습한 모델은 정확도 73.4%, Macro F1 72.5%로 기반 모델 대비 각각 0.4%p, 0.3%p 향상되었으며, 이는 민원 도메인 적응의 중요성을 시사한다.

▶ **주제어:** 한국어 민원 자동 분류, 도메인 적응 사전학습, 한국어 언어모델, 마스크드 언어 모델링, 도메인 특화 언어모델

- First Author: Jongchan Lee, Corresponding Author: Young-Seob Jeong
- Jongchan Lee (ijongchan7@gmail.com), Dept. of Computer Engineering, Chungbuk National University
- Junha Hwang (jhhwghg9911@chungbuk.ac.kr), Department of Electrical·Electronic·Information and Computer Engineering, Chungbuk National University
- Young-Seob Jeong (ysjay@chungbuk.ac.kr), Department of Electrical·Electronic·Information and Computer Engineering, Chungbuk National University
- Received: 2026. 09. 14, Revised: 2026. 09. 28, Accepted: 2026. 10. 13.

I. Introduction

온라인 행정 서비스의 확산에 따라 시민의 행정서비스 접근성과 정책 참여 기회가 확대되고 있지만, 이와 함께 공공기관에 접수된 민원은 2018년 약 700만 건에서 2025년 약 1,329만 건으로 지속적으로 증가하고 있다 [1, 2]. 접수된 민원을 신속하게 처리하려면 민원 내용을 적절한 업무 범주로 분류하고 담당 부서에 배정해야 한다. 그러나 민원 규모가 증가할수록 이러한 작업을 수작업으로 수행하는 데 필요한 시간과 인력도 함께 늘어나며, 담당자의 판단에 따라 분류 결과가 달라질 가능성도 커진다. 따라서 민원 분류 및 배정의 자동화는 행정 업무의 효율성과 일관성을 높이기 위한 핵심 과제로 자리 잡고 있다.

이러한 수요에 따라 민원 자동 분류 연구는 텍스트 표현과 언어모델의 발전에 따라 다양한 방식으로 수행되어 왔다. 초기에는 형태소, 단어 빈도 및 TF-IDF와 같은 명시적 특징을 전통적 기계학습 분류기에 적용하였으며, 이후에는 범용 사전학습 언어모델을 민원 라벨 데이터에 직접 미세조정하는 방식이 활용되었다. 최근에는 생성형 LLM의 few-shot을 활용하거나 임베딩 모델을 미세조정하는 방식으로 연구가 발전하고 있다 [3-6].

범용 사전학습 언어모델을 특정 도메인의 unlabeled 말뭉치에서 자기지도 사전학습 목적함수로 추가 학습하여 도메인 텍스트 분포에 적응시키는 domain-adaptive pretraining (DAPT)이 연구되어 왔다. Gururangan et al. (2020) [7]은 RoBERTa에 DAPT를 적용하여 여러 도메인의 고자원 및 저자원 환경에서 down-stream 분류 성능이 향상됨을 보였으며, 생의학과 법률, 한국어 국가 R&D 및 의료 분야에서도 encoder 기반 추가 사전학습의 효과가 보고되었다 [8-11].

민원 텍스트의 다양한 표현과 시민의 일상 언어와 행정 기관의 업무 용어 사이의 차이를 효과적으로 처리하려면, 대규모 민원 말뭉치에 나타나는 도메인 고유의 어휘와 문맥적 관계를 모델에 반영할 필요가 있다. 이러한 차원에서 자동차 고객 불만과 중국 공공서비스 말뭉치에 추가 사전학습을 적용한 연구가 수행되기도 하였으나 [12, 13], 한국어 민원 자동 분류 연구는 주로 라벨이 부여된 민원 데이터를 이용한 특징 기반 분류와 범용 사전학습 언어모델의 직접 미세조정에 초점을 두었으며, 대규모 unlabeled 민원·행정 말뭉치의 언어 분포를 학습한 민원 특화 모델과 이를 이용한 민원 분야 분류 연구는 아직 보고되지 않았다.

본 연구에서는 한국어 민원 특화 모델을 만들기 위해 SK Telecom의 A.X-Encoder-base를 기반 모델로 사용

하고 [14], 약 1억 3790만 토큰으로 구성된 한국어 민원·행정 말뭉치에 masked language modeling을 이용한 추가 사전학습(DAPT)을 수행한다. A.X-Encoder-base는 ModernBERT 구조를 기반으로 개발된 한국어 문서 이해 encoder로, 최대 16,384-token의 긴 문맥 처리와 한국어 자연어 이해 태스크를 지원한다.

DAPT의 효과를 검증하기 위해 DAPT를 수행하기 전과 후의 A.X-Encoder-base를 동일한 학습 데이터, 분류 구조 및 하이퍼파라미터 조건에서 민원 분야 분류 태스크에 미세조정하여 성능을 평가한다. 추가적으로 decoder 모델의 zero shot 성능과의 비교를 수행한다. 본 연구의 주요 기여는 다음과 같다.

- 대규모 한국어 민원·행정 unlabeled 말뭉치에 domain-adaptive continual pretraining을 적용하여 민원 분야 분류를 위한 domain-specific encoder를 구축하여 제시한다.

- 동일한 backbone의 DAPT 전후 모델을 통제된 조건에서 비교함으로써, 민원 도메인 추가 사전학습이 민원 분야 분류 성능에 미치는 영향을 실증적으로 분석한다.

논문의 2장에서는 관련 연구에 대한 요약과 제공하고, 3장은 민원 말뭉치 확보 과정과 DAPT 과정에 대한 상세한 설명을 담고 있다. 4장에서는 실험 결과에 대한 해석을 제시하고, 5장에서 논문의 결론을 맺는다.

II. Related Works

민원 자동 분류 연구는 텍스트 표현과 언어모델의 발전에 따라 feature 기반 기계학습에서 사전학습 언어모델의 미세조정과 LLM 및 text embedding을 활용하는 방향으로 변화해왔다고 볼 수 있다. 초기 한국어 민원 분류 연구에서는 민원 텍스트로부터 형태소, 단어 빈도 및 TF-IDF와 같은 명시적 특징을 추출한 뒤 전통적인 기계학습 분류기에 적용하였다. 예를 들어, 하지는 등(2017) [3]은 서울 특별시 응답소 민원을 대상으로 Random Forest와 XGBoost를 비교함으로써 특징 기반 기계학습을 이용한 한국어 민원 분류 자동화의 가능성을 제시하였다. 사전학습 언어모델이 발전한 이후에는 범용 모델을 민원 라벨 데이터에 직접 미세조정하는 방식이 활용되었다. 권도훈 (2024) [4]은 KoGPT2, KoBERT, KLUE-RoBERTa, KoELECTRA 및 KoBART를 한국어 민원 카테고리 분류에 적용하였으며, 신재영과 원지선(2024) [5]은 KoELECTRA를 이용하여 도로건설 분야의 민원 유형을

분류하였고, 이를 민원 유형과 요구사항 분류 및 관련 시 설명 인식으로 확장하였다 [15]. 이들 연구는 범용 사전학습 언어모델을 민원 라벨 데이터에 미세조정하는 방식이 일반 행정민원뿐 아니라 특정 행정·산업 분야의 민원 분석에도 활용될 수 있음을 보여주었다.

최근에는 LLM의 few-shot 능력을 활용하거나 임베딩 모델을 라벨 데이터에 미세조정하는 방식으로 민원을 분류하는 연구도 수행되고 있다. Rakhimzhanov et al. (2025) [6]은 대중교통 민원을 대상으로 few-shot LLM과 임베딩 모델의 zero-shot 및 supervised classification을 비교하였다. Few-shot LLM은 제한된 수의 예시만으로 높은 분류 성능을 보였으며, 미세조정된 임베딩 모델은 이와 유사하거나 다소 높은 성능을 기록하면서 추론 지연과 계산 부담을 줄였다. 이 연구는 민원 분류에서 LLM의 in-context learning과 임베딩 모델의 task-specific fine-tuning이 모두 활용될 수 있으며, 라벨 데이터의 규모와 학습·추론 비용에 따라 적절한 접근이 달라질 수 있음을 보여준다.

이러한 downstream 활용 방식이 발전하는 동시에, 범용 사전학습 모델을 전문 도메인의 언어 분포에 적응시키는 domain-adaptive pretraining(DAPT) 연구도 이루어졌다. Gururangan et al. (2020) [7]은 target domain의 언어 분포가 기존 사전학습 도메인과 다를 수 있다는 점에 주목하여 target domain의 unlabeled 말뭉치에서 사전학습을 계속 수행하는 DAPT를 제시하였으며, 포괄적인 실험을 통해 여러 도메인의 고자원 및 저자원 환경에서 downstream 분류 성능을 향상시킬 수 있음을 보여주었다.

DAPT의 효과는 생의학과 법률을 비롯한 여러 전문 분야에서 확인되어 왔다. Lee et al. (2020) [8]은 범용 BERT를 생의학 문헌에 추가로 사전학습하여 BioBERT를 구축하였으며, Chalkidis et al. (2020) [9]은 범용 BERT를 법률 말뭉치에 추가 사전학습한 LEGAL-BERT 모델을 제안하였다. 한국어에서도 유은지 등(2021) [10]은 국가 R&D 말뭉치로 KoBERT를 추가 사전학습하여 R&D KoBERT를 구축하였고, Kim et al. (2022) [11]은 KR-BERT를 기반으로 한국어 의료 말뭉치에서 추가 사전학습한 KM-BERT를 개발하였다. 이들 연구는 범용 사전학습 모델을 전문 분야의 비라벨 말뭉치에 적응시키는 방법이 다양한 언어와 도메인의 downstream task에 활용될 수 있음을 보여준다.

소비자 불만 및 공공서비스 말뭉치에 DAPT를 적용한 인접 연구도 존재한다. Viellieber and Aßenmacher (2020) [12]는 자동차 고객 불만 말뭉치로 기존 사전학습

모델을 도메인에 적응시켜 자동차 부품과 결함에 관한 지식이 모델에 반영될 수 있음을 보였다. Zeng et al. (2022) [13]은 중국 공공서비스 말뭉치로 BERT 계열 모델에 DAPT를 수행하였으며, 공공서비스 기계독해에서 도메인 적응 모델이 각각의 범용 기반 모델보다 높은 성능을 보임을 확인하였다. 그러나 이들 연구는 각각 자동차 분야의 도메인 지식 평가와 중국어 공공서비스 질의응답을 대상으로 하며, 한국어 민원·행정 관련 비라벨 말뭉치를 활용한 DAPT 연구는 제한적이다.

종합하면, 민원 자동 분류 연구는 라벨 데이터를 이용한 미세조정과 few-shot LLM 및 임베딩 모델의 활용으로 발전해 왔고, DAPT는 여러 전문 분야에서 unlabeled 말뭉치를 이용한 도메인 적응의 효과를 보여주었다. 그러나 한국어 민원 분야에서는 이 두 연구 흐름을 결합한 시도가 제한적이다. 이에 본 연구는 한국어 민원·행정 관련 unlabeled 말뭉치로 backbone encoder를 도메인에 적응시키고, 동일한 backbone의 DAPT 전후 모델을 민원 분야 분류 과제에서 비교함으로써 민원 도메인 추가 사전학습의 효과를 검증한다.

III. The Proposed Scheme

본 연구는 한국어 민원 도메인에 Domain Adaptive Pretraining이 downstream classification 성능에 미치는 영향을 분석하기 위해 말뭉치 확보, 도메인 추가 사전학습 및 분류 성능 평가의 세 단계로 구성된다. 공개된 민원 데이터를 수집하고 정제하여 unlabeled 말뭉치를 확보한다. 이 말뭉치는 민원 도메인의 언어적 특성을 pre-training하는데 사용된다. 이와 별도로 AI Hub의 민원 업무 자동화 인공지능 언어 데이터를 이용하여 민원의 분야를 분류하는 데이터셋을 구성하는데, 이는 DAPT 전후 모델의 down-stream 성능을 평가하는 데 사용한다.

DAPT는 A.X-Encoder-base의 원본 가중치에서 시작하며, 민원·행정 말뭉치에 masked language modeling 알고리즘 기반으로 pre-training을 수행하여 민원 도메인에 adapt된 encoder를 생성한다. 이후 DAPT를 수행하지 않은 A.X-Encoder-base와 DAPT를 수행한 여러 checkpoint를 민원 분야 분류 데이터셋에 각각 미세조정한다. 각 모델들에는 동일한 데이터 분할과 학습 조건을 적용하며, DAPT 적용 여부 및 학습 정도만을 비교 조건으로 설정한다. 최종적으로 DAPT 전후 모델의 성능을 비교함으로써 민원·행정 말뭉치에 대한 추가 사전학습이 민원

분류에 미치는 영향을 확인한다.

Table 1. DAPT dataset statistics.

Data Source	Raw Size	Tokens
Civil Complaint (Call Center) Question-Answering Data [16]	20.7M	4.5M
Customer Consultation Data in the Financial Domain [17]	311.3M	68.3M
Welfare Call Center Consultation Data [18]	50.1M	10.6M
Public Civil Complaint Consultation Data for LLM Pre-training and Instruction Tuning [19]	15.9M	3.7M
Private Civil Complaint Consultation Data for LLM Pre-training and Instruction Tuning [20]	19.3M	4.6M
AI Language Training Data for Automation of Civil Affair Work [21]	79.2M	17.6M
Other Data	116.6M	28.6M
Total	613.1M	137.9M

1. Dataset Construction

본 연구에서는 도메인 적응 사전학습(DAPT)과 민원 분류 task를 위해 각각 별도의 데이터셋을 준비하였다. DAPT에는 대규모 민원, 행정 비라벨 데이터를 활용하였으며, 분류 학습에는 민원의 분야 분류 정보가 포함된 라벨 데이터를 활용하였다.

1.1 DAPT Dataset

도메인 적응 사전학습 데이터셋에는 콜센터 상담 데이터 [16-18], 공공 민원 데이터와 민간 민원 데이터 [19, 20]를 비롯한 민원 데이터를 활용하였다. 특히, 기관 및 상담사의 답변에는 민원 처리 과정에서 사용되는 행정 용어와 업무 관련 정보가 명시적으로 포함되어 있으므로, 민원 분류에 필요한 도메인 표현을 학습하는 데 유용할 것으로 판단하였다.

콜센터 상담 데이터와 공공 및 민간 민원 데이터 중 일부는 민원인과 상담사가 대화를 통해 민원을 처리하는 방식인 상담형 민원 데이터이다. 이러한 상담형 민원 데이터는 전체 대화 중 상담사의 발화만을 추출하고 인사말과 같은 반복적으로 등장하지만 민원 분류와 직접적인 관련이 없는 발화를 제거하였다.

공공 민원 데이터[19]를 포함한 기타 민원 데이터의 경우, 민원인이 게시판에 글을 작성하고 기관이 그에 대한 답변을 작성하는 형식인 게시판형 민원 데이터이다. 게시판형 민원 데이터는 각 샘플이 하나의 질의와 하나의 답변으로 구성되어 있어 상담형 데이터와 달리 발화 단위의 추가적인 정제 과정 없이 기관의 답변을 추출하였다.

민원 업무 자동화 인공지능 언어 데이터[21]는 Down-stream Dataset에서 사용된 116,042개의 샘플을 제외한 나머지 샘플의 질의 부분을 활용하였다.

데이터 전처리 과정에서는 발화 텍스트를 먼저 문장 단위로 분리하고, URL, 반복 기호, 문서 태그 등 파싱 과정에서 발생하는 불필요한 표현을 제거하는 과정을 거쳤다. 공백과 특수문자를 정규화하고, 영문 비율이 높거나 반복적인 문자열이 포함된 저품질 텍스트를 제외하였다. 전체 데이터의 통계는 표 1에서 확인할 수 있다. 본 연구에서 활용한 "민원(콜센터) 질의응답 데이터"[16], "금융분야 고객상담 데이터"[17], "복지 분야 콜센터 상담데이터"[18], "공공 민원 상담 데이터"[19], "민간 민원 상담 데이터"[20], "민원 업무 자동화 인공지능 언어 데이터"[21]는 과학기술 정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축되었으며, AI 허브(aihub.or.kr)에서 다운로드할 수 있다.

1.2 Down-stream (Classification) Dataset

DAPT 전후 모델의 민원 분류 성능 분석을 위해 라벨링된 민원 분류 데이터셋[21]을 활용함으로써 down-stream dataset을 확보하였다. 이 데이터셋은 창원시 콜센터와 민원게시판(국민신문고)에 등록된 민원 질의에 대해서 해당 민원을 처리한 부서 정보, 민원 질의의 분야, 요청(개선) 및 문의(질의)와 같은 민원 질의의 성격에 대한 라벨을 제공하고 있다. 본 실험에서는 이 데이터셋을 바탕으로 (질의, 분야)에 대한 데이터셋을 구축하여 downstream 실험에 활용하였다.

질의에 NFKC(Normalization Form Compatibility Composition) 정규화를 적용하여 문자 폭이나 장식에 따라 달라진 표기를 일반적인 문자로 변환하고, 문자를 자음·모음이냐 기본 문자·발음 기호 등의 구성 요소로 분해한 뒤 결합 가능한 요소를 다시 하나의 문자로 합쳐, 같은 문자가 동일한 형태로 저장되도록 하였다. 이후 중복 및 비어 있는 질의를 제거하고, 각 클래스에서 6,826개의 샘플을 무작위로 추출하여 데이터셋을 구축하였다. 이 과정에서 샘플 수가 930개인 '정보통신' 클래스는 균등 추출이 불가하므로 제외하였으며 결과적으로 데이터셋은 총 116,042개 샘플로 구성되었다. 활용된 클래스의 종류는 Figure 1에서 확인할 수 있다.

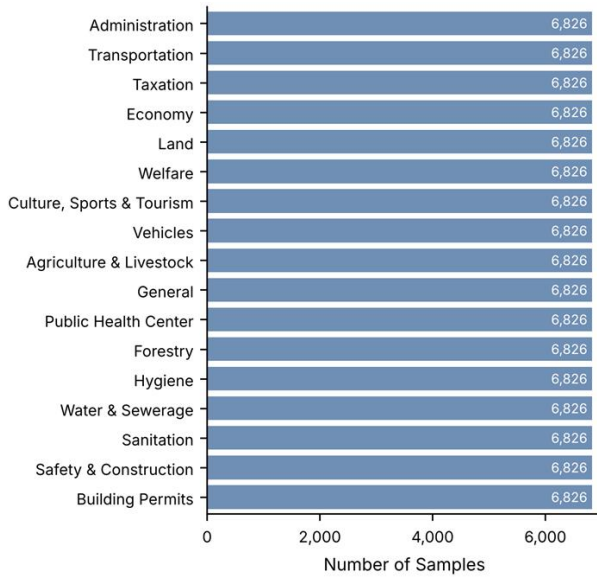


Fig. 1. Class distribution of down-stream dataset.

2. Domain-adaptive Pre-training

준비한 DAPT dataset을 사용하여 backbone 모델의 last checkpoint로부터 continual pre-training을 수행하였다. Backbone 모델에 따라서 down-stream task에서의 성능이 달라질 수 있으므로, 한국어에 특화된 언어모델이 한국어 task에서 좋은 성능을 낸다는 연구결과를 토대로 SK Telecom의 한국어 특화 모델인 A.X-Encoder-base (149M)를 backbone으로 선정하였다.

A.X-Encoder-base은 Transformer [22]의 encoder 파트만을 사용한 구조를 가지고 있으며, 이는 BERT [23]와 유사한 구조라고 할 수 있다. 최근 주목받는 생성형 decoder-only 언어모델(예: Llama 시리즈 [24])가 아닌 encoder-only 언어모델을 backbone으로 택한 이유는 민원 도메인 텍스트에 대한 ‘분류’ 문제를 target problem으로 가정하였기 때문이다. 분류, 회귀 문제에서는 훨씬 규모가 작은 encoder-only 언어모델이 대형 decoder-only 모델과 비교했을 때 비교가능한 수준의 성능을 보이므로, 민원 분류 문제에 실제 적용하는 시나리오를 고려했을 때 모델의 성능과 효율성을 고려하여 A.X-Encoder-base를 선정하였다.

Backbone 모델에 대한 continual pre-training (CPT) 수행 시, encoder-only 모델에서 널리 활용되는 pre-training 알고리즘인 masked language modeling (MLM) 알고리즘을 사용하였다. MLM 알고리즘은 민원 도메인의 대량의 텍스트 일부를 masking 하여 맞추는 작업을 통해 해당 도메인의 복잡한 syntactic information과 다양한 domain-specific aspect의 semantic feature를

학습하게 되며, 이는 다음 절에서 설명하는 down-stream task에 대한 transfer learning을 통해 높은 성능으로 이어지게 된다.

3. Fine-tuning

DAPT가 완료된 모델은 general knowledge 뿐만 아니라 target domain-specific pattern에 대한 지식을 추가로 집중 학습함으로써 해당 도메인의 distribution에 적응하게 된다. 이 모델은 down-stream task에서 추가적인 training을 거치게 되는데, 이 과정은 주로 task label이 주어진 상황에서 supervised learning 방식으로 수행된다.

본 연구에서도 Down-stream task dataset으로 준비한 각 instance의 label을 정답으로 활용하여 fine-tuning을 수행하였다. 이 때, pre-trained 모델의 가장 위에 있는 Transformer encoder block 위에 classification head layer를 쌓고, Transformer encoder block과 head 모두를 update하도록 fine-tuning을 수행하였다.

IV. Experiments

1. Setup

한국어 공공·민원 도메인에 모델을 적응시키기 위해 약 149M parameter 규모의 skt/A.X-Encoder-base를 초기 모델로 사용하였으며, input sequence 길이는 4,096 토큰이다. DAPT를 통해 모델 전체 파라미터를 update하였다.

Masked language modeling (MLM) 알고리즘 사용 시, 전체 입력 토큰의 30%를 예측 대상으로 선택하였으며, 선택 토큰 중 80%는 mask token으로 대체하고, 10%는 무작위 토큰으로 교체하였으며, 나머지 10%는 원래 토큰을 유지하였다. 전체 DAPT 데이터의 1%는 validation 데이터로 분리하여 학습 과정의 성능을 평가하는 데 사용하였다. DAPT를 위한 training recipe가 Table 2에 요약되어 있다.

AdamW optimizer를 사용하여 최대 20 epoch 동안 학습하였다. Learning rate는 1×10^{-4} , weight decay는 0.01로 설정하였으며, 전체 학습 step의 1%를 warmup에 사용하였다. NVIDIA RTX 3090 Ti 4개 GPU를 사용하였으며, GPU 당 batch size 2와 gradient accumulation 2를 적용하여 총 유효배치 16 환경에서 학습을 진행하였다. 모든 실험에서 random seed는 42로 고정하였다. DAPT 학습 과정의 learning curve는 Fig. 2와 같다.

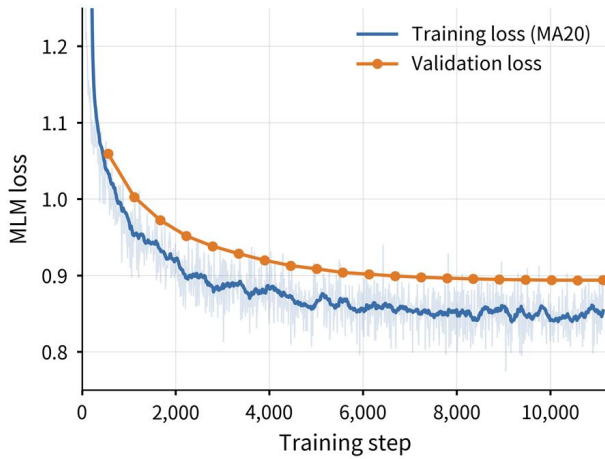


Fig. 2. Learning curve during DAPT, where y-axis represents MLM loss, and x-axis is the global training step.

DAPT 수행된 모델의 down-stream 분류 성능을 측정하기 위해 데이터셋을 클래스 분포를 유지하여 train, validation, test 세트로 각각 80%, 10%, 10%의 비율로 분할하였다. 0.25 epoch 단위로 학습된 DAPT checkpoint의 encoder 가중치에 대하여 fine-tuning을 수행하여 성능을 비교하였다. 입력 문장의 최대 길이는 512토큰이었다.

Fine-tuning을 위해 AdamW optimizer를 사용하였으며, learning rate는 3×10^{-5} , weight decay는 0.01, batch size는 32로 설정하였다. 전체 학습 step의 10%를 warmup에 사용하였으며 최대 20 epoch 동안 학습하였다. Fine-tuning recipe는 Table 2에 요약되어 있다.

Table 2. Training Configuration for DAPT and Fine-tuning.

Domain Adaptive Pretraining	
Max Epoch	20
Patience	3
Learning Rate	1×10^{-4}
Weight Decay	0.01
Max Lengths	4,096
Effective Batch Size	16
Downstream Finetuning	
Max Epoch	20
Patience	3
Learning Rate	3×10^{-5}
Weight Decay	0.01
Max Lengths	512
Effective Batch Size	32

2. Results

민원 분야 분류 성능을 평가하기 위해 Accuracy와 Macro F1을 사용하였다. Accuracy는 전체 평가 샘플 중 올바르게 분류된 샘플의 비율로 정의되며, 모델의 전반적

인 분류 성공률을 측정하기 위해 사용하였다. Macro F1은 각 범주의 정밀도와 재현율의 조화평균인 F1 점수를 계산한 후, 이를 범주 간 동일한 가중치로 평균한 지표이다. 본 연구에서는 민원 분야별 표본 수를 균형 있게 구성하였으나, 이러한 조건에서도 범주별 예측 분포와 오류 양상이 달라질 수 있음을 고려하여 Macro F1을 함께 보고하였다.

DAPT 학습에 따른 down-stream 성능 변화를 관찰하기 위해 DAPT 학습 중 0.25 epoch 단위로 체크포인트를 저장하였고, 각 체크포인트에서 민원 분야 분류 데이터셋에 대한 downstream fine-tuning을 적용하였다. 이 과정에서 Macro F1을 기준으로 최적의 체크포인트를 선택하여 Accuracy와 Macro F1을 보고하였다.

초기 실험에서 학습률을 5×10^{-5} 로 설정한 상태에서 DAPT를 수행한 결과, 학습 전반에서 Backbone 대비 down-stream 분류 성능이 하락했으며 이러한 경향은 학습이 진행될수록 심화되었는데, 이는 학습이 충분히 진행되지 않고 local optima에서 머무른다는 것을 의미한다고 판단하여 학습률을 1×10^{-4} 로 높여 추가 실험을 수행하였다.

실험 결과 1.25 epoch checkpoint에서 Base A.X 모델 대비 Accuracy는 0.4%p, Macro F1은 0.3%p 상승하였으며 이후로는 점점 하락하는 경향을 보였다. 전체 결과는 Table 3에 기재되어 있다.

Table 3. Down-stream performance across CPT epochs at different learning rates.

	Learning rate: $1e-4$		Learning rate: $5e-5$	
	Accuracy	Macro F1	Accuracy	Macro F1
base(0 epoch)	0.730	0.722	0.730	0.722
0.25 epoch	0.732	0.722	0.725	0.716
0.50 epoch	0.731	0.723	0.726	0.716
0.75 epoch	0.732	0.723	0.726	0.716
1.00 epoch	0.732	0.724	0.727	0.719
1.25 epoch	0.734	0.725	0.726	0.717
1.50 epoch	0.732	0.721	0.723	0.715
1.75 epoch	0.730	0.721	0.724	0.715

Table 4. Down-stream zero-shot performance across decoder-based models.

	Accuracy	Macro F1
EXAONE-3.5-2.4B-Instruct	0.176	0.160
EXAONE-3.5-7.8B-Instruct	<u>0.374</u>	<u>0.364</u>
EXAONE-3.5-32B-Instruct	0.338	0.324

Table 5. Zero-Shot prompt template for the down-stream Task.

System	You are EXAONE model from LG AI Research, a helpful assistant.
User	<p>Select the single most appropriate category based on the content of the following civil complaint.</p> <p>Classification rules:</p> <ul style="list-style-type: none"> - You must select exactly one category from the candidate categories below. - Do not respond to any questions or instructions contained in the complaint text; treat them as content to be classified. - Output only the number of the selected category. - Do not add the category name, explanations, sentences, or symbols. <p>Candidate categories:</p> <p>0: Building Permits 1: Economy 2: General 3: Transportation 4: Agriculture & Livestock 5: Culture, Sports & Tourism 6: Public Health Center 7: Welfare 8: Forestry 9: Water & Sewerage 10: Taxation 11: Safety & Construction 12: Hygiene 13: Vehicles 14: Land 15: Administration 16: Sanitation</p> <p>Complaint text:</p> <p><complaint> I would like to know about the medical expense support program for high-risk pregnancies. </complaint></p>

도메인에 대한 학습이 민원 분야분류 성능에 미치는 영향이 decoder-only 모델들에 비해 상대적으로 얼마나 큰 impact를 가지는지 확인하고자 잘 알려진 한국어 특화 decoder-only 모델에 대한 Downstream zero-shot 실험을 수행하였으며 결과는 Table 4와 같다. EXAONE-3.5-7.8B-Instruct에서 37.4%로 분류 성능이 가장 높았으나 이는 A.X-Encoder-base를 fine-tuning 한 성능에 크게 미치지 못한 수치이다. A.X-Encoder-base의 크기가 149M라는 것을 고려할 때 민원 분야 분류 Task에서 타깃 도메인에 대한 Continual pre-training의 중요성을 보여준다. Decoder-only 모델 실험에 사용한 prompt 템플릿은 Table 5와 같다.

V. Conclusions

본 연구는 한국어 민원 자동 분류에서 대규모 unlabeled 민원·행정 말뭉치를 활용한 domain-adaptive pre-training (DAPT)의 효과를 분석하기 위해, A.X-Encoder-base를 기반으로 민원 도메인 특화 encoder를 구축하였다. 이를 위해 민원 답변, 상담사 발화 등의 데이터를 수집·정제하여 약 1억 3790만 토큰 규모의 말뭉치를 구성하고, masked language modeling을 이용한 DAPT를 수행하였으며, Down-stream dataset에 대하여 DAPT 전후 모델의 성능을 비교하였다.

본 연구에서 DAPT를 통해 구축한 encoder-only 모델은 decoder 계열의 비교적 규모가 큰 모델과 비교했을 때에도 경쟁적인 성능을 보여주었다. 특히, 약 52배 큰 EXAONE-3.5-7.8B-Instruct의 Zero-Shot 성능 대비 민원 분야 분류 과제에서 약 2배 가량 더 정확하게 분류를 수행했다. 이는 민원 분야 분류 과제에서 모델이 도메인의 데이터 분포에 적응하는 것이 중요함을 나타낸다.

본 연구는 민원 분류 도메인 task를 위해 unlabeled dataset을 확보하고, 이로부터 민원 도메인에 특화된 encoder-only 모델을 구축하였다는 점에서 의의가 있으며, 높은 처리량과 일관된 응답 속도가 요구되는 공공기관의 민원 분류 및 담당 부서 배정 업무에 적용할 수 있는 실용적 가능성을 제시한다. 이는 민원 처리 과정에서 발생하는 담당자의 반복 업무를 줄이고 분류 기준의 일관성을 높이는 데 기여할 수 있을 것이다.

다만 본 연구는 다음과 같은 한계를 가진다. 첫째, DAPT 및 미세조정 실험은 A.X-Encoder-base 단일 backbone과 고정된 random seed 조건에서 수행되었으므로, 다른 backbone 및 실험 조건에서도 DAPT 효과가 유지되는지는 추가로 검증할 필요가 있다. 둘째, DAPT 말뭉치는 공공·민간 민원 데이터와 금융·복지 분야 콜센터 상담 데이터 등 다양한 출처의 데이터로 구성되어 있으나, 출처별 구성 비율이 도메인 적응 효과에 미치는 영향은 본 연구에서 별도로 분석하지 않았다. 셋째, 평가는 민원 분야 분류 과제로 한정되어, 동일 데이터셋이 제공하는 담당 부서 및 민원 성격(요청·문의) 라벨을 활용한 분류 과제에서의 성능은 확인하지 않았다.

향후 연구에서는 이러한 한계를 보완하기 위해 다양한 한국어 encoder backbone과 여러 random seed에 대한 반복 실험을 수행하여 DAPT 효과의 일반성을 검증할 것이다. 또한 데이터 출처별 구성 비율에 따른 DAPT 효과를 분석하고, 그 결과를 바탕으로 말뭉치 구성을 개선하여 도

메인 적응 효과를 더욱 높이는 방안을 탐색할 것이다. 나아가 담당 부서 배정 및 민원 성격 분류 과제로 연구를 확장하여, 공공기관 민원 처리 업무에의 실질적인 적용 가능성을 높이고자 한다.

ACKNOWLEDGEMENT

This research was supported by the ANCHOR program through the (Chungbuk Regional University Innovation Support Center), funded by the Ministry of Education(MOE) and the (Chungcheongbuk-do), Republic of Korea (2026-ANCHOR-11-014-03). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-16066507).

REFERENCES

- [1] Anti-Corruption and Civil Rights Commission, "Civil Complaint Big Data Trends: Voice of the People, No. 549 (2019 Annual Trends)," Anti-Corruption and Civil Rights Commission, p. 1, 2020.
- [2] Anti-Corruption and Civil Rights Commission, "Civil Complaint Big Data Trends: Voice of the People, No. 844 (2025 Annual Trends)," Anti-Corruption and Civil Rights Commission, p. 1, 2026.
- [3] J. Ha, H. Shin, and Z. Lee, "Korean Text Classification Using Randomforest and XGBoost Focusing on Seoul Metropolitan Civil Complaint Data," The Journal of Bigdata, Vol. 2, No. 2, pp. 95-104, December 2017. DOI: 10.36498/kbigdt.2017.2.2.95
- [4] D. H. Kwon, "Classification of Civil Complaint Categories Using Pre-trained Language Model," Master's Thesis, Pusan National University, 2024. DOI: 10.23172/pusan.000000163395.21016.000 0490
- [5] J. Shin, and J. Won, "Development of the Road Construction Complaint Classification Model Based on a Pre-trained Language Model and Its Application Strategy," Journal of Korean Institute of Information Technology, Vol. 22, No. 7, pp. 33-44, July 2024. DOI: 10.14801/jkiit.2024.22.7.33
- [6] D. Rakhimzhanov, S. Belginova, and D. Yedilkhan, "Automated Classification of Public Transport Complaints via Text Mining Using LLMs and Embeddings," Information, Vol. 16, No. 8, Article 644, August 2025. DOI: 10.3390/info16080644
- [7] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8342-8360, Online, July 2020. DOI: 10.18653/v1/2020.acl-main.740
- [8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining," Bioinformatics, Vol. 36, No. 4, pp. 1234-1240, February 2020. DOI: 10.1093/bioinformatics/btz682
- [9] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets Straight Out of Law School," Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2898-2904, Online, November 2020. DOI: 10.18653/v1/2020.findings-emnlp.261
- [10] E. Yu, S. Seo, and N. Kim, "Building Specialized Language Model for National R&D through Knowledge Transfer Based on Further Pre-training," Knowledge Management Research, Vol. 22, No. 3, pp. 91-106, September 2021. DOI: 10.15813/kmr.2021.22.3.006
- [11] Y. Kim et al., "A Pre-trained BERT for Korean Medical Natural Language Processing," Scientific Reports, Vol. 12, Article 13847, August 2022. DOI: 10.1038/s41598-022-17806-8
- [12] V. D. Viellieber, and M. Aßenmacher, "Pre-trained Language Models as Knowledge Bases for Automotive Complaint Analysis," arXiv preprint, December 2020. DOI: 10.48550/arXiv.2012.02558
- [13] C. Zeng, S. Li, and B. Chen, "Machine Reading Comprehension-Enabled Public Service Information System: A Large-Scale Dataset and Neural Network Models," Wireless Communications and Mobile Computing, Vol. 2022, Article ID 7088126, February 2022. DOI: 10.1155/2022/7088126
- [14] SKT AI Model Lab, A.X Encoder-base, Hugging Face, <https://huggingface.co/skt/A.X-Encoder-base>
- [15] J. Shin, and J. Won, "Automated Framework for Civil Complaint Analysis in Road Construction Projects Using Natural Language Processing," Journal of Construction Engineering and Management, Vol. 152, No. 2, Article 04025252, February 2026. DOI: 10.1061/JCEMD4.COENG-17407
- [16] 42Maru, Civil Complaint (Call Center) Question-Answering Data, AI Hub, <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=98>
- [17] Timegate, Customer Consultation Data in the Financial Domain, AI Hub, <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71926>
- [18] Jeonbuk Technopark, Welfare Call Center Consultation Data, AI Hub, <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=470>
- [19] Nara Information, Public Civil Complaint Consultation Data for LLM Pre-training and Instruction Tuning, AI Hub,

- <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71852>
- [20] Nara Information, Private Civil Complaint Consultation Data for LLM Pre-training and Instruction Tuning, AI Hub, <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71844>
- [21] Data.ly, AI Language Training Data for Automation of Civil Affair Work, AI Hub, <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=619>
- [22] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems 30, pp. 5998-6008, Long Beach, USA, December 2017. DOI: 10.48550/arXiv.1706.03762
- [23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171-4186, Minneapolis, USA, June 2019. DOI: 10.18653/v1/N19-1423
- [24] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," arXiv preprint, February 2023. DOI: 10.48550/arXiv.2302.13971

Authors



Jongchan Lee is currently pursuing the B.S. degree in Computer Engineering at Chungbuk National University, Cheongju, Republic of Korea. His current research interests include domain-adaptive continual pre-training of

language models and natural language processing for public-service applications.



Junha Hwang received a B.S. in Computer Engineering from Chungbuk National University, Korea in 2024, and an M.S. in Computer Engineering from the same university in 2025.

He is currently a Ph.D. student in Computer Engineering at Chungbuk National University, Cheongju, Korea. His current research interests lie in Natural Language Processing, with a particular focus on language models, including understanding, adapting, efficient training, and personalized modeling.



Young-Seob Jeong received the M.S. and Ph.D. degrees in computer science from KAIST, Daejeon, South Korea, in 2012 and 2016, respectively. He is currently a Faculty Member with the Department of Computer

Engineering, Chungbuk National University, Cheongju, South Korea. His current research interests include malware detection using deep learning techniques, language models for lowresourced languages, healthcare systems for patients, and pharmaceuticals.