BERT를 활용한 부산교통공사 민원 자동분류

박준형, 황상현, 서정원, 김영찬*, 홍준표

국립부경대학교, *부산교통공사

jason990820@gmail.com, noul77@naver.com, sjw3570@naver.com, *channy@humetro.busan.kr, jphong29@gmail.com

Automatic Classification of Busan Transportation Corporation Civil Complaints Using BERT

Jun-Hyeong Park, Sang-Hyun Hwang, Jung-Won Seo, Young-Chan Kim*, Jun-Pyo Hong

Pukyong National Univ., *Busan Transportation Corporation

요 약

본 연구에서는 한국어 데이터로 사전 학습된 BERT(Bidirectional Encoder Representations from Transformers) 모델을 활용해 부산교통공사 민원 데이터의 처리부서를 자동으로 분류하는 시스템 구현하고 성능을 분석하였다. 사전학습된 대표적인 4가지 한국어 BERT 모델을 시스템에 적용해 분류 성능을 분석하고, 추가적인 성능 개선을 위한 아이디어를 도출하였다. 해당 연구 결과는 부산교통공사 민원분류에 직접 적용되어 지역산업 문제해결 및 산업 고도화에 기여할 수 있다.

Ⅰ. 서론

부산 지하철 사용인구 증가와 시설 노후화로 인해 민원 수가 꾸준히 증가하는 추세이지만, 민원을 분류해 할당하는 전담 인원이 배정되어있지 않아 신속한 민원처리에 어려운 문제가 있다. 실제로 부산교통공사는 연평균 5,500건의 민원이 들어오며, 현재 직원 한 명이 수동으로 처리부서를 분류하고 있다, 따라서 인력 부족 문제 완화 및 신속한 민원처리를 위해 자연어 처리(Natural Language Processing, NLP) 기술을 활용한 민원 자동분류 시스템의 필요성이 점차 증가하고 있다.

본 연구에서는 대규모 언어 모델(Large Language Model, LLM) 중에서도 BERT 모델을 기반으로 하여 자동분류 시스템을 만들고 성능을 분석하였다. 최근 특허상담 자동분류, 학술 문헌 자동분류에서 BERT 모델을 이용한 유사 연구가 있었으나, 하나의 BERT 모델을 사용한 기존의 연구들과 달리 본 연구에서는 서로 다른 타입의 한국어 데이터셋으로 사전 훈련된 여러 BERT 모델들을 적용해 성능을 비교, 분석했다는 점에서 차별화된다 [1-3]. 또한, 기존 연구에서는 다루지 않았던 부산교통공사의 실제 민원 데이터를 활용해 모델의 실제 적용 가능성과 효과를 확인한다는 점에서 의의가 있다.

Ⅱ. 본론

2.1 데이터 분석 및 전처리

민원 자동분류 모델을 구현하기 위해 본 연구팀은 부산교통공사에서 제공한 2015년부터 2023년까지의 민원 데이터셋 3,089건을 사용하였다. 데이터 전처리를 통해 파인튜닝에 필요한 학습 데이터셋 2,502개, 검증 데이터셋 278개, 테스트 데이터셋 309개를 구축하였다.

민원 자동분류 모델을 구현하기 위한 민원 텍스트 데이터 전처리
① 숫자, 영어, 한국어를 제외한 그 외 언어 제거(한자 등) ② 특수문자 및 이모티콘 제거 ③ URL 제거

표 1 민원 텍스트 전처리

분류에 대한 학습과 테스트를 실행하기 위해서 전체 데이터셋을 문장과 레이블(label, 총 26개) 컬럼으로 구분하였다. 문장은 민원내용이며 레이블은 민원처리 부서명이다. 부산교통공사 기회예산실의 자문을 통해 전체 민원내용 문장상에 분류 코드를 지정하였으며 이를 기준으로 자동분류를 수행하였다.

분류 코드	원 레이블
① 영업 및 고객 (37.5%)	영업처, 운영사업소
② 승무 (18.2%)	승무처, 승무사업소, 종합관제소
③ 차량 (13.9%)	차량처, 차량사업소
④ 홍보 및 디자인 (4.9%)	홍보문화실
⑤ 시설 및 건설 (8.7%)	시설처, 시설사업소, 건설계획처, 건설공사처
⑥ 통신 및 신호 (3.5%)	신호통신처, 신호통신사업소
⑦ 전기 및 기계환경 (5.9%)	기계환경사업소, 전기기계환경처, 전기사업소
⑧ 기타 (7.6%)	전략사업처, 경영지원처, 미래성장연구원, 안전관리처, 회계처, 노사협력처, BTC아카데미, 기획예산실, 감사실

표 2 원 레이블과 분류 코드 간의 상관관계

기본적으로 BERT의 최대 입력 토큰(token) 수는 512이다. 총 3,089건의 민원 문장 중 그 길이가 512토큰 이하인 문장의 개수는 3,024건으로 전체의 97.89%를 차지한다.

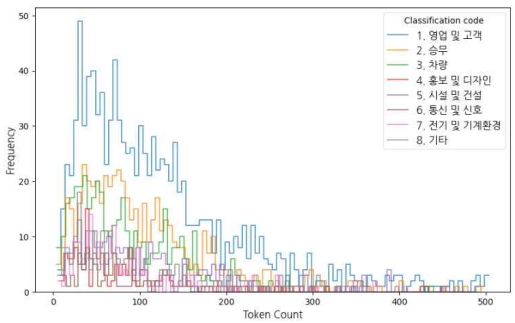

그림 1 토큰 수 512 이하에 대한 분류 코드별 문장 개수

2.2 모델 선택 및 이유

트랜스포머 모델의 인코더에 기반하여 2018년 구글에 의해 발표된 BERT 모델은 대규모 데이터를 활용한 사전 훈련을 통해 풍부한 문맥 정보를 습득하고, 이를 바탕으로 파인 튜닝을 통해 특정 작업에 최적화하여 사용할 수 있다.

이에 본 연구팀은 Hugging Face에서 한국어로 사전 훈련된 잘 알려진 네 가지 BERT 모델을 이용하여 민원 자동분류 모델을 구축하고 테스트하였다. 아래는 네 가지 모델과 주요 속성을 나타낸 표이다.

속성 / 모델명	hidden size	layers	max token	vocab size	pre-train data size
kykim/bert-kor-base	768	12	512	42000	70GB
klue/bert-base	768	12	512	32000	62GB
snunlp/KR-BERT-char16424	768	12	512	16424	2.47GB
beomi/kcbert-base	768	12	300	30000	15.4GB

표 3 네 가지 BERT 모델과 주요 속성

2.3 파인튜닝 방법론

한국어로 사전 훈련된 BERT 모델을 기반으로 하여, 모델의 최종 레이어에 8개의 출력 클래스를 갖는 분류기를 추가하였다. 이는 각 민원의 분류 코드를 예측하도록 설계되었다.

학습 환경은 Google Colab Pro+를 사용하였으며 pytorch와 transformers 라이브러리를 사용하였다. 아래 표는 학습 과정에서 사용한 하이퍼파라미터 값이다.

하이퍼파라미터	값
epoch	15
learning rate	0.00003
batch size	64
hidden dropout probability	0.1%
attention dropout probability	0.1%
L2 regularization coefficient	0.001

표 4 하이퍼파라미터 값

2.4 성능 평가

본 연구에서는 정확도(accuracy)를 기준으로 민원 자동분류의 성능을 평가하였다. 검증 데이터셋에서 가장 높은 정확도를 보인 모델을 최종적으로 선택하여 테스트 데이터셋에 적용하였다. 각 모델 별 테스트 정확도는 아래 표와 같다.

모델명	정확도(%)
kykim/bert-kor-base	82.85
klue/bert-base	76.05
snunlp/KR-BERT-char16424	77.99
beomi/kcbert-base	74.28

표 5 각 모델별 테스트 정확도

테스트 결과, kykim/bert-kor-base 모델의 테스트 정확도는 82.85%로 다른 모델들에 비해 가장 높은 정확도를 보였다. 이러한 결과는 해당 모델이 상대적으로 더 많은 한국어 데이터로 사전 훈련을 받았으며, 더 큰 단어 집합을 보유하고 있기 때문인 것으로 추정된다.

가장 성능이 좋았던 kykim/bert-kor-base 모델을 사용하였을 경우 테스트 데이터셋의 혼동 행렬(confusion matrix)은 다음과 같다.

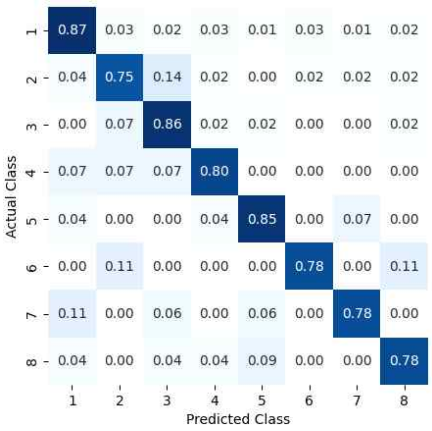

그림 2 혼동행렬

위 혼동행렬은 테스트 데이터셋에 대해 실제 클래스들 중 예측한 클래스의 비율을 나타낸 것이다. 실제로 분류 코드가 ② 승무이지만 ③ 차량으로 오분류한 비율이 0.14로 전체 오분류 비율 중 가장 많았다. 이는 승무관련 부서와 차량관련 부서의 역할과 민원의 내용이 다소 겹쳐 모델이 혼동한 것으로 추정된다. 추후 더 많은 민원 데이터와 복잡한 민원 유형을 포함시켜 모델의 일반화 능력을 강화하고 오분류를 최소화할 예정이다.

Ⅲ. 결론

본 연구는 BERT를 기반으로 민원 자동분류 모델을 만들고 성능을 분석하였다. 부산교통공사의 실제 민원 데이터를 활용하여, 한국어로 사전 훈련된 여러 BERT 모델의 성능을 비교 분석함으로써, 실제 환경에서의 모델 적용 가능성과 효과를 검증하였다.

앞으로의 연구에서는 훈련 과정에서 더 많은 언어 데이터와 복잡한 민원 유형을 포함시켜 모델의 일반화 능력을 강화하고, 실시간 민원 처리 시스템 구축을 위한 모델 최적화가 필요하다.

참 고 문 헌

[1] Dong-Hun Noh, Jae-Ok Min, So-Youn Woo, “A Study on the Performance Improvement of Automatic
Intellectual Property Counseling Classification: Using
the Transformer-based AI Model BERT,” The Journal of Intellectual Property, vol. 19, no. 1, pp. 159-177, Mar. 2024.

[2] In-hu Kim, Seong-hee Kim, “Automatic Classification of Academic Articles Using BERT Model Based
on Deep Learning,” Journal of the Korean Society for Information Mangerment, vol. 39, no. 3, pp. 293-310, 2022.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and KristinaToutanova, “BERT: Pretraining of deep bidirectional transformers for language understanding,” arXiv, preprintarXiv:1810.04805, 2018.