# 112 신고 데이터를 활용한 BERT 사전학습모델 특징 융합 파인튜닝 성능평가

한인규, *권은정, *변성원, *박현호, *정의석, **윤준호

한국외국어대학교, *한국전자통신연구원, **경찰대학

abcolligh@hufs.ac.kr, *{ejkwon, swbyon, hyunhopark, esjung}etri.re.kr, **20200057@police.ac.kr

# BERT pre-learning model feature convergence fine tuning performance evaluation using 112 reported data

InGyue Han, EunJung Kwon, SungWon Byon, HyunHo Park, EuiSuk Jung, JunHo Yoon

Hankuk University of Foreign Studies, Electronic and Telecommunications Research Institute (ETRI), Korean National Police University

요 약

본 논문은 112 신고내용의 분류 정확도를 높이기 위해 BERT 사전학습 모델과 112 신고 데이터를 활용하여 BERT 사전학습 모델을 재학습한 모델에 대하여 모델 성능 비교 결과를 제시한다. 각각의 모델을 잠재 의미 분석의 결과와 합성하여 머신러닝 분류기를 적용한 결과에 대한 성능을 제공함으로써 112신고 데이터를 활용하여 재학습된 모델의 성능이 치안 도메인의 특징을 잘 추출한다는 것을 알 수 있었다. 본 실험을 통해 112 신고내용에 따른 사건 종별에 대한 분류 정확도를 개선할 가능성을 제공하였다.

## Ⅰ. 서 론

Transformer[1]모델을 기반으로 한 BERT[2]는 언어의 구조, 표현, 형태 등을 모두 학습할 수 있고, 이를 토대로 텍스트 분류와 같은 다양한 자연어처리 태스크에서 우수한 성능으로 두각을 나타내고 있다. 하지만 BERT 다국어 사전모델은 위키백과 데이터로만 학습했기 때문에 데이터 양이 절대적으로 부족하다. 또한, 경제, 과학, 의학 등과 같이 사용되는 단어가 다르고 언어의 특성이 다른 분야에 사용되었을 때 성능이 현저하게 떨어지는 문제가 생긴다.

위에서 언급된 BERT의 한계를 극복하고 성능을 개선하기 위해 다양한 방법들이 연구되고 있는데, 대표적인 방법으로 사전학습(pre-training)과 파인튜닝(fine-tuning) 기법이 있다. Chi Sun[4]의 연구에서는 BERT 레이어 수정, 학습률 조정을 통한 파인튜닝과 특정 도메인에 대한 BERT 모델 사전학습 기법을 통해 문서 분류와 감성 분석 모델을 생성하였다. 그리고 기존 연구들과 성능 비교를 통해 사전학습과 파인튜닝 기법의 효용성을 입증하고 있다.

본 논문은 기본 BERT 모델, 112 신고 데이터를 학습시켜 구축한 사전학습 모델 그리고 사전학습 모델에 값을 합성한 모델의 비교를 통해 사전학습과 잠재 의미 분석 값 합성을 통한 파인튜닝이 텍스트 분류 정확도에 끼치는 영향에 대해 알아보고자 한다.

## Ⅱ. 본론

본 논문에서는 사전학습 그리고 비교를 위한 데이터 셋으로 112 경찰 데이터를 사용하였다. 112 신고 데이터는 1만 건의 신고내용 데이터가 44가지 사건종별로 구분되어있다. 본 실험에서는 44가지의 사건종별 중 7가지 사건종별(보호조치, 기타형사범, 시비, 교통사고, 소음, 행패소란, 주취자)에 해당하는 신고내용 데이터 5,686건을 추출하여 활용하였다. 7가지 사건종별 데이터 수는 그림 1에서 나타나는 것과 같은데, 사건종별마다 신고내용 데이터 수에 편차가 있었다.

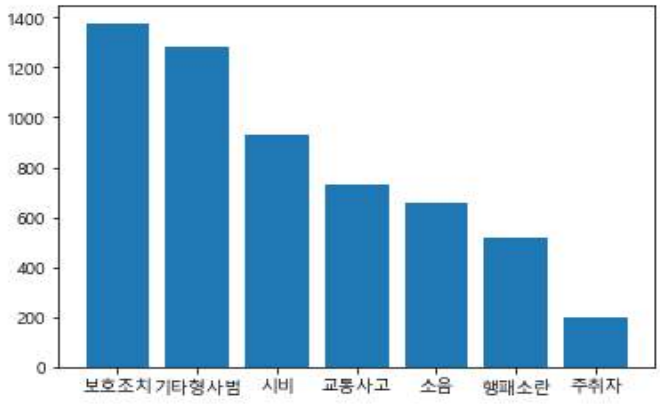


그림 1. 사건종별 데이터 수

그림 2는 본 실험의 모델 구성도를 나타낸다. 그림 2의 화살표 색에서 구분되는 것과 같이 본 모델은 세 가지 방법의 입력을 머신러닝 분류 모델에 제공하고 분류 결과를 얻는다.

첫 번째 방식은 기본으로 제공되는 BERT 모델 마지막 레이어의 출력을 머신러닝 모델의 입력으로 활용하는 방법이다. 하나의 문장이 BERT를 통과하게 되면 768개의 출력 유닛이 나온다. 이렇게 얻은 출력들이 머신러닝 분류 모델의 입력으로 들어가 학습을 통해 문장의 종류를 구분해준다.

두 번째 방식에서는 BERT 모델에 112 신고 데이터를 학습시켜 BERT 사전학습 모델을 생성하고 저장하여 언제든 모델을 호출할 수 있도록 한다. 사전학습 모델을 생성할 때 BERT 모델의 구조는 수정하지 않고 가중치만 업데이트한다. 따라서 구축된 사전학습 모델에 하나의 문장이 통과하게 되면 첫 번째 방식과 마찬가지로 768개의 출력 유닛을 얻게 된다. 이렇게 얻은 출력들이 머신러닝 분류 모델의 입력으로 들어가 학습을 통해 문장의 종류를 구분해준다.

마지막은 두 번째 방식에서 얻게 되는 768개의 출력과 112 신고 데이터에 잠재 의미 분석을 통해 얻어낸 7개의 토픽에 대한 값을 합성하는 파인튜닝 방식이다. 이렇게 합성된 값이 머신러닝 분류 모델의 입력으로 들어가 학습을 통해 문장의 종류를 구분해준다.

세 가지 방식 모두 BERT의 12개의 레이어 중 마지막 레이어를 머신러닝의 입력으로 사용하고 있는데, 그 이유는 Chi Sun [4]의 연구결과 BERT의 마지막 레이어를 가지고 파인튜닝을 진행했을 때 가장 좋은 성능을 보였기 때문이다.

머신러닝 분류 모델로는 대중적인 모델인 Logistic Regression, Support Vector Machine(SVM), Random Forest, 그리고 K-Nearest Neighbor(KNN) 총 네 가지를 사용하였다.

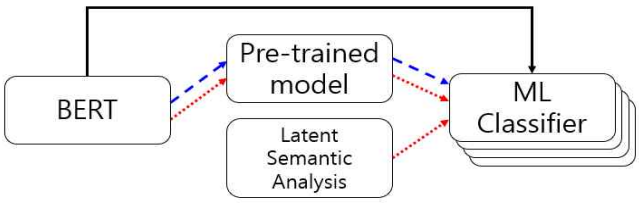


그림 2. 모델 구성도

| | BERT | | BERT-PT | | BERT-PTLSA | |
|---|---|---|---|---|---|---|
| | ACC | F1 | ACC | F1 | ACC | F1 |
| Logistic Regression | 0.627 | 0.571 | 0.777 | 0.778 | 0.777 | 0.778 |
| SVM | 0.646 | 0.570 | 0.774 | 0.775 | 0.776 | 0.777 |
| Random Forest | 0.545 | 0.451 | 0.774 | 0.776 | 0.775 | 0.778 |
| KNN | 0.518 | 0.466 | 0.771 | 0.774 | 0.772 | 0.774 |

표 1. 실험 결과

표 1에서 ACC는 Accuracy, F1은 f1-score, BERT-PT는 112 신고 데이터를 학습한 사전 학습한 것이고 BERT-PTLSA는 사전 학습 모델에 LSA 값을 합성한 것을 의미한다. 모든 수치는 소수점 셋째 자리에서 반올림하였다.

모든 머신러닝 분류 모델에서 112 신고 데이터로 만든 사전학습 모델을 활용한 BERT-PT가 BERT와 비교했을 때 더 높은 정확도와 f1-score를 기록하였다. 특히 K-Nearest Neighbor를 분류기로 사용했을 때 정확도는 0.253, Random Forest를 사용했을 때 f1-score는 0.325로 가장 큰 성능 개선이 이루어진 것을 확인하였다. 그러나 BERT-PT와 BERT-PTLSA를 비교해본 결과 Logistic Regression을 사용했을 경우 성능 개선이 전혀 이루어지지 않았고 나머지 Support Vector Machine, Random Forest, K-Nearest Neighbor에서도 성능 개선이 이루어지긴 했지만 유의미한 차이라고 보기엔 어려운 수준이었다. 가장 큰 차이를 보였던 분류기는 Support Vector Machine으로 0.002의 정확도와 f1-score 상승을 보였다.

그림 3에서 왼쪽은 BERT에 SVM을 합성한 confusion matrix이고 오른쪽은 BERT-PTLSA에 SVM을 합성한 confusion matrix이다. 세로축은 실제 문장의 사건 종별 분류결과이고 가로축은 SVM에서 예측한 문장의 사건 종별 분류결과이다. 두 matrix를 비교했을 때 BERT-PTLSA가 0~6까지 모든 사건 종별을 더 많이 맞춘 것을 확인하였다.

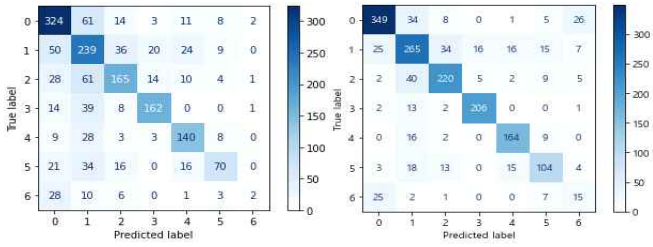


그림 3. Confusion matrix

## Ⅲ. 결론

본 논문에서는 문장의 분류 정확도를 높이기 위해 112 신고 데이터를 활용하여 BERT 사전학습 모델을 구축하고 문장의 잠재 의미 분석을 통해 도출한 값을 합성하여 다양한 머신러닝 분류기에 대입 후 정확도를 측정하였다.

특정한 분야의 언어적 특성을 갖는 문장을 분류할 때 BERT 사전학습 모델을 재학습하여 활용하는 것은 효과적이다. 즉, 도메인 언어로 학습된 모델을 기반으로 학습용 데이터가 충분하지 않은 경우 본 논문과 같이 BERT 모델 마지막 레이어의 출력을 머신러닝 모델의 입력으로 활용하는 방법이 문장의 분류 정확도를 높이는데 효과적임을 알 수 있었다. 차원을 축소하여 잠재 의미 분석 결과값 합성에 대한 파인튜닝은 정확도 상승에 유의미한 영향을 주지는 못한다는 것을 알 수 있었다. 머신러닝 분류 모델 종류마다 잠재 의미 분석값 합성의 영향이 다른 것을 확인할 수 있었는데 추후 머신러닝 분류 모델의 구조를 분석하여 그 원인을 알아보고자 한다. 그리고 잠재 의미 분석값 합성보다 효과적인 파인튜닝 기법에 관해 연구해볼 것이다.

## ACKNOWLEDGMENT

이 논문은 2021년도 정부(경찰청)의 재원으로 지원받아 수행된 연구결과임 [내역사업명: 112 긴급출동 의사결정 지원 시스템/ 연구개발과제번호: PR08-03-000-21]

## 참 고 문 헌

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. "Attention is all you need." Advances in Neural Information Processing Systems, arXiv:1706.03762, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova "Bert: Pre-training of deep bidirectional transformers for language understanding." Conference of the North American Chapter of the Association for Computational Linguistics. 2018.

[3] Lee, Chi Hoon, Lee, Yeon Ji, Lee, Dong Hee "A Study of Fine Tuning Pre-Trained Korean BERT for Question Answering Performance Development" Journal of Information Technology Service, pp. 83-91, October 2020.

[4] Chi Sun, Xipeng Qiu, Yige Xu, Xuanjing Huang "How To Fine-Tune BERT For Text Classification?" Chinese Computational Linguistics, pp. 194 - 206, 2019.