# EDA 기법을 적용한 BERT 기반의 감성 분류 모델 생성

이진상O, 임희석*
O고려대학교 인공지능융합학과,
*고려대학교 컴퓨터학과
e-mail: jinsang16@korea.ac.krO, limhseok@korea.ac.kr*

# Sentiment Classification Model Development Based On EDA-Applied BERT

Jin-Sang LeeO, Heui-Seok Lim*
ODept. of Applied Artificial Intelligence, Korea University,
*Dept. of Computer Engineering, Korea-Digital University

● 요 약 ●

본 논문에서는 데이터 증강 기법 중 하나인 EDA를 적용하여 BERT 기반의 감성 분류 언어 모델을 만들고, 성능 개선 방법을 제안한다. EDA(Easy Data Augmentation) 기법은 데이터가 한정되어 있는 환경에서 SR(Synonym Replacement), RI(Random Insertion), RS(Random Swap), RD(Random Deletion) 총 4가지 세부 기법을 통해서 학습 데이터를 증강 시킬 수 있다. 이렇게 증강된 데이터를 학습 데이터로 이용해 구글의 BERT를 기본 모델로 한 전이학습을 진행하게 되면 감성 분류 모델을 생성해 낼 수 있다. 데이터 증강 기법 적용 후 전이 학습을 통해 생성한 감성 분류 모델의 성능을 증강 이전의 전이 학습 모델과 비교해 보면 정확도 측면에서 향상을 기대해 볼 수 있다.

**키워드:** 감성 분류, 데이터 증강, 전이 학습

## I. Introduction

최근 BERT와 같은 PLM(Pretrained Language Model)을 베이스 모델로 이용하여 텍스트 감성을 분류하고 연구하는 활동이 활발하게 이루어지고 있다. 베이스 모델에 데이터만 준비하여 전이 학습만 수행하면 손쉽게 새로운 텍스트 감성 분류 모델을 생성해 낼 수 있기 때문이다. 하지만 이를 학습시키는 데이터는 매우 한정적이며, 이런 상황이 한계로 작용하는 경우가 많다. 그래서 본 연구에서는 이를 개선하기 위한 방법의 하나로 EDA 기법을 적용한 BERT 기반의 감성 분류 모델을 소개한다. 본 논문에서는 EDA 기법과 BERT에 대한 소개와 이를 적용한 감성 분류 모델의 생성 방법, 생성한 모델의 성능의 측정 결과를 차례로 소개하며, 마지막으로 개선 방향에 대해서도 기술하고 있다.

## II. Preliminaries

### 1. Related works

#### 1.1 EDA(Easy Data Augmentation)

학습 데이터셋이 매우 부족하여, 모델의 성능을 보장하기 어려운 상황에서 데이터 량을 증강시키기 위한 방법이다. EDA[1]는 SR, RI, RS, RD 총 4가지의 세부적인 방법을 가지고 있다. SR(Sysnonym Replacement)는 텍스트 기반 학습 데이터 내에서 임의로 단어를 동의어로 변경하여 새로운 데이터셋을 만드는 방법이다. RI(Random Insertion)은 학습데이터 내의 임의의 단어의 동의어를 임의의 위치에 삽입하여 새로운 데이터셋을 만드는 방법이다. RS(Random Swap)은 데이터셋 내의 2개의 단어의 위치를 임의로 교체하여 새로운 데이터셋을 생성하는 방법이다. RD(Random Deletion)은 데이터셋 내의 임의의 단어를 삭제하여 새로운 데이터셋을 생성하는 방법이다.

### 1.2 BERT

BERT(Bidirectional Encoder Representation from Transformer)[2]는 2018년 구글에서 공개한 언어 모델로써, 매우 높은 성능을 보이기 때문에 현재 자연어 처리 분야에서 가장 폭넓게 쓰이고 있는 텍스트 임베딩 모델이다. BERT가 높은 성능을 보이는 가장 주된 이유는 임베딩 시에 문맥을 고려하기 때문이다. BERT가 문맥을 고려할 수 있는 가장 큰 이유는 트랜스포머(Transformer)의 인코더를 사용하기 때문에 이 인코더에 문장을 입력하게 되면, 트랜스포머 인코더의 멀티 헤드 어텐션 매커니즘을 이용해 문장을 구성하는 단어의 관계를 이해하고, 이를 통해 문맥을 고려한 임베딩을 할 수 있게 된다.

## III. The Proposed Scheme

### 1. 연구 환경 구성

본 연구는 BERT 모델을 기반으로 감성 분석 언어 모델을 생성하는 것이므로 여러 버전의 BERT 모델 중 한국어를 지원하는 BERT-base-mulitilingual 모델을 베이스 모델로 사용하였다. 모델 학습의 텍스트 데이터는 네이버 영화 리뷰 코퍼스 NSMC(Naver Sentiment Movie Corpus)를 사용하였다. 이 코퍼스는 학습 데이터 15만개, 시험 데이터 5만개로 구성되어 있다. Tokenizer는 Hugging Face 라이브러리에서 제공하는 BertTokenizer를 활용하였다.

### 2. 연구 진행

EDA 기법을 적용하기 위한 구현을 1차적으로 진행하였다. 문장 형태의 학습 데이터를 토크나이징 하는 것으로 가정하고, EDA 4가지 기법인 SR, RI, RS, RD 총 4가지를 구현하였다. 증강 시키는 데이터의 양은 학습 데이터 1개당 EDA 4가지 기법을 모두 적용하는 것으로 고려하여 최대 4개가 새롭게 생성되도록 하였다. 이를 통해서 증강된 데이터 중 한국어가 아닌 경우는 제거 하였으며, 생성된 데이터 중 중복이 되는 데이터들은 학습 데이터에서 전부 삭제 하였다. 이 과정을 통해 기존 NSMC 학습 데이터 15만 건을 353060건으로 증강시켰다. 그리고 기존의 15만 건의 NSMC 데이터와 증강된 데이터를 이용하여 감성 분류 모델을 각각 생성하고, 성능을 비교하였다.

### 3. 연구 결과

NSMC 기본 데이터만을 사용하여 생성한 기본 모델과 EDA를 적용한 모델을 학습 시키면서 Training Loss와 Validation data의 accuracy를 측정하였다. EDA를 통해 data 증강을 하는 경우에 Training Loss는 0.07, Validation accuracy는 0.07 개선이 되는 것을 확인할 수 있었다.

Table 1. 모델 별 Training Loss, Validation accuracy

| | 기본 모델 | EDA 적용 모델 |
|---|---|---|
| Training Loss | 0.18 | 0.11 |
| Validation accuracy | 0.87 | 0.94 |

NSMC의 5만 건의 시험 데이터를 이용하여, 기본 모델과 EDA를 적용한 모델의 accuracy를 측정 하였으며, 0.001 정도 소폭가 개선되는 것을 확인할 수 있었다.

Table 2. 모델 별 Accuracy

| | 기본 모델 | EDA 적용 모델 |
|---|---|---|
| Accuracy | 0.872 | 0.873 |

## IV. Conclusions

BERT 모델을 기반으로 하여 텍스트 감성 분류 모델을 만들고 성능을 개선하는 방법은 다양하다. 그 중 본 연구에서 제안하는 EDA 방식을 적용하는 경우 연구 결과의 validation accuracy 성능이 보여주듯 학습 모델이 overfitting 되는 것을 방지할 수 있다. 또한 비록 작은 폭이긴 하지만 학습 모델의 accuracy도 향상도 기대할 수 있다. 본 연구에서 제안한 방법 이외에도 BERT 베이스 모델을 좀 더 한국어에 특화된 KoBERT와 같은 모델을 사용하거나, 한국어 Tokenizer를 사용한다면 모델 의 성능을 더 개선할 수 있을 것이라 생각한다.

## REFERENCES

[1] Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.