J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

BERT 기반 자연어처리 모델의 미세 조정을 통한 한국어 리뷰 감성 분석: 입력 시퀀스 길이 최적화☆

Fine-tuning BERT-based NLP Models for Sentiment Analysis of Korean Reviews: Optimizing the sequence length

황 성 아1 박 세 연1 장 백 철1*
Sunga Hwang Seyeon Park Beakcheol Jang

요 약

본 연구는 BERT 기반 자연어처리 모델들을 미세 조정하여 한국어 리뷰 데이터를 대상으로 감성 분석을 수행하는 방법을 제안한다. 이 과정에서 입력 시퀀스 길이에 변화를 주어 그 성능을 비교 분석함으로써 입력 시퀀스 길이에 따른 최적의 성능을 탐구하고자 한다. 이를 위해 의류 쇼핑 플랫폼 M사에서 수집한 텍스트 리뷰 데이터를 활용한다. 웹 스크래핑을 통해 리뷰 데이터를 수집하고, 데이터 전처리 단계에서는 긍정 및 부정 만족도 점수 라벨을 재조정하여 분석의 정확성을 높였다. 구체적으로, GPT-4 API를 활용하여 리뷰 텍스트의 실제 감성을 반영한 라벨을 재설정하고, 데이터 불균형 문제를 해결하기 위해 6:4 비율로 데이터를 조정하였다. 의류 쇼핑 플랫폼에 존재하는 리뷰들을 평균적으로 약 12 토큰의 길이를 띄었으며, 이에 적합한 최적의 모델을 제공하기 위해 모델링 단계에서는 BERT기반 사전학습 모델 5가지를 활용하여 입력 시퀀스 길이와 메모리 사용량에 집중하여 성능을 비교하였다. 실험 결과, 입력 시퀀스 길이가 64일 때 대체적으로 가장 적절한 성능 및 메모리 사용량을 나타내는 경향을 띄었다. 특히, KcELECTRA 모델이 입력 시퀀스 길이 64에서 가장 최적의 성능 및 메모리 사용량을 보였으며, 이를 통해 한국어 리뷰 데이터의 감성 분석에서 92%이상의 정확도와 신뢰성을 달성할 수 있었다. 더 나아가, BERTopic을 활용하여 새로 입력되는 리뷰 데이터를 카테고리별로 분류하고, 최종 구축한 모델로 각 카테고리에 대한 감성 점수를 추출하는 한국어 리뷰 감성 분석 프로세스를 제공한다.

☞ 주제어 : BERT, 하이퍼 파라미터 미세 조정, 입력 시퀀스 길이, 토픽 모델링, 감성 분석, 한국어 리뷰 분석

ABSTRACT

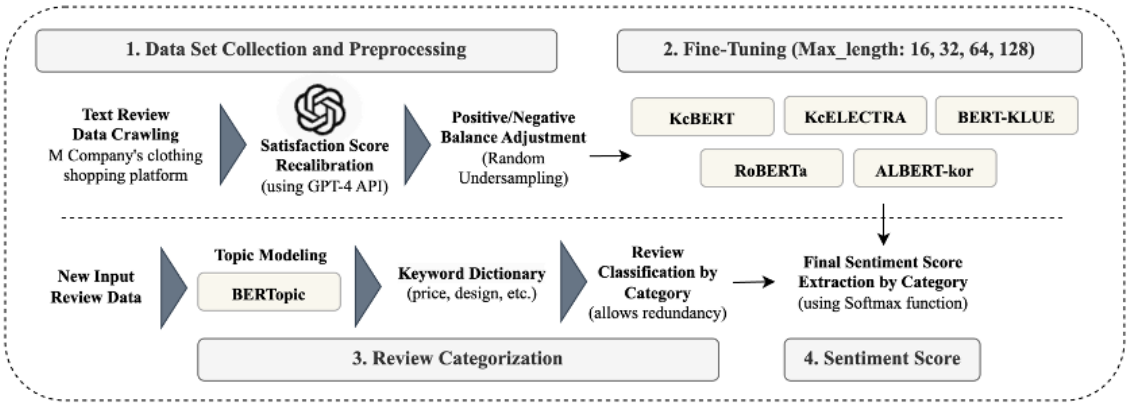
This paper proposes a method for fine-tuning BERT-based natural language processing models to perform sentiment analysis on Korean review data. By varying the input sequence length during this process and comparing the performance, we aim to explore the optimal performance according to the input sequence length. For this purpose, text review data collected from the clothing shopping platform M was utilized. Through web scraping, review data was collected. During the data preprocessing stage, positive and negative satisfaction scores were recalibrated to improve the accuracy of the analysis. Specifically, the GPT-4 API was used to reset the labels to reflect the actual sentiment of the review texts, and data imbalance issues were addressed by adjusting the data to 6:4 ratio. The reviews on the clothing shopping platform averaged about 12 tokens in length, and to provide the optimal model suitable for this, five BERT-based pre-trained models were used in the modeling stage, focusing on input sequence length and memory usage for performance comparison. The experimental results indicated that an input sequence length of 64 generally exhibited the most appropriate performance and memory usage. In particular, the KcELECTRA model showed optimal performance and memory usage at an input sequence length of 64, achieving higher than 92% accuracy and reliability in sentiment analysis of Korean review data. Furthermore, by utilizing BERTopic, we provide a Korean review sentiment analysis process that classifies new incoming review data by category and extracts sentiment scores for each category using the final constructed model.

☞ keyword : BERT, hyperparameter fine-tuning, input sequence length, topic modeling, sentiment analysis, Korean review analysis

1. 서 론

다양한 제품 리뷰는 소비자들이 구매 결정을 내리는 데 중요한 역할을 한다. 특히, 텍스트 리뷰는 소비자 경험을 전달하는 중요한 정보원으로써, 감성 분석을 통해 소비자 만족도를 파악하고, 제품 및 서비스 개선에 중요한

1 Yonsei Graduate School of Information Seoul, Korea.
* Corresponding author (bjang@yonsei.ac.kr)
[Received 05 July 2024, Reviewed 10 July 2024, Accepted 06 August 2024]
☆ This work was supported by the National Research Foundation of Korea (NRF) funded by Korean Government under Grant RS-2023-00273751.



(그림 1) 연구의 전체 프레임워크
(Figure 1) The overall framework of our research

인사이트를 제공한다[1-3]. 그러나 이러한 리뷰 데이터는 종종 부정확하거나 불균형을 이루는 만족도 점수 분포와 텍스트의 다양성으로 인해 분석에 어려움을 겪는다[4]. 본 연구는 이러한 문제를 해결하고자 의류 쇼핑 플랫폼 M사에서 수집한 리뷰 데이터를 기반으로 텍스트 감성 분석 프로세스를 제안한다.

텍스트 감성 분석은 자연어 처리 분야에서 중요한 연구 주제로, 최근에는 BERT(Bidirectional Encoder Representations from Transformers)[5]와 같은 사전학습된 언어 모델을 사용하여 높은 성능을 달성하고 있다. 본 연구에서는 한국어 BERT기반의 사전학습 모델들을 미세 조정하여 리뷰 데이터를 분석하고, 긍정 및 부정 감성 점수를 정확하게 예측하고자 한다. 이를 위해 데이터 전처리, 만족도 점수 라벨 값 재조정, 데이터 불균형 문제[6] 해결, 모델의 입력 시퀀스 길이 조절[7] 등의 과정을 거쳐 최적의 한국어 리뷰 감성 분석 모델을 구축한다.

데이터 전처리 단계에서는 웹 스크래핑을 통해 수집한 리뷰 데이터를 정제하고, 긍정 및 부정 라벨을 재조정하여 분석의 정확성을 높인다. 특히, 최근 GPT-4[8]의 API를 활용하여 리뷰 텍스트의 실제 감성을 반영한 라벨을 재설정[9,10]하고, 데이터 불균형 문제를 해결하기 위해 특정 비율로 데이터를 조정한다. 이후, BERT 기반 모델을 사용하여 감성 분석을 수행하고, 입력 시퀀스 길이에 따른 성능과 메모리 사용량을 비교 분석한다.

본 연구는 의류 플랫폼 산업 관련 리뷰 감성 분석을 기반으로 고객에게 가장 적합한 소비를 할 수 있도록 상품에 대한 세부 정보를 제공하는 것이 주된 목적이다. 구체적으로 리뷰 데이터의 평균 길이를 파악하여 최적의 입력 시퀀스 길이를 탐구하고, Python언어와 Pytorch, Tensorflow 프레임워크로 BERT 기반 딥러닝 사전학습 모델들을 활용하여 미세 조정을 거쳐 리뷰 분석 모델을 개발한다. 이때, 의류 쇼핑 플랫폼 M사에서 사용자들의 한국어 텍스트 형태의 리뷰 데이터를 중심으로 진행한다. 해당 데이터는 사용자의 반응을 직접적으로 보여주므로, 제품에 대한 사용자의 의견을 파악하는 데 유용하다[11-13]. 이는 소비자의 구매 실패율을 줄이고, 구매까지 도달하는 데에 소요되는 시간을 단축하여 소비자들의 만족도를 높이며 판매자에게도 매출 상승 등 긍정적인 영향을 미쳐 이커머스 시장 전반에서 유의미한 결과를 이끌어낼 것으로 기대한다. 본 연구의 주요 기여를 요약하자면 다음과 같다.

- 한국어 의류 쇼핑 플랫폼 리뷰 데이터를 수집하고 GPT-4 API로 구매 만족도 라벨 값을 재조정한다.
- 입력 시퀀스 길이에 따라 BERT 기반 다양한 사전학습 모델을 한국어 리뷰에 대해 미세 조정한다.
- 한국어 의류 쇼핑 플랫폼에서 최적의 BERT 기반 모델과 입력 시퀀스 길이를 탐구한다.
- 새로 입력되는 리뷰의 카테고리를 나누고, 리뷰의 감성 점수를 산출하는 프로세스를 제공한다.

이러한 과정을 종합하여 소비자가 제품 리뷰를 편리하고 효과적으로 파악하는 데 도움을 준다. 나아가, 본 연구는 의류 쇼핑 플랫폼에서 리뷰 데이터를 활용한 감성 분

석의 새로운 가능성을 제시하고, 향후 관련 연구 및 실무에 유용한 인사이트를 제공할 것이다.

2. 관련 연구

2.1 BERT 기반 사전학습 모델

자연어처리 분야에서 BERT기반 사전학습된 모델들은 다양한 도메인에서 효과적으로 활용되고 있다. BERT[5]는 Transformer[14]의 encoder 부분만 사용한 모델로 MLM(Masked Language Model)과 NSP(Next Sentence Prediction) 방법을 채택한다. RoBERTa[15]는 BERT 모델이 충분히 학습되지 않았다는 판단하에, 정적 마스킹 대신 동적 마스킹을 사용하고 학습 시 NSP 작업을 배제하며, 더 많은 데이터를 활용하여 성능을 향상시킨 모델이다. ALBERT[16]는, BERT 모델에서 factorized embedding layer parameterization, cross-layer parameter sharing의 방법으로 파라미터 수를 줄여 기존 BERT를 경량화시킨 모델이다. 또한, KLUE(Korean Language Understanding Evaluation) 벤치마크[17]는 2021년 Naver, Kakao, KIST등이 함께 만든 데이터 세트로, klue/bert*, klue/roberta**는 해당 데이터 세트로 학습시킨 모델이다. KcBERT는 기존의 한국어 BERT 모델이 주로 한국어 위키, 뉴스 기사, 책 등 잘 정제된 텍스트로 학습된 것과 달리, 정제되지 않은 댓글형 데이터에 적용할 수 있도록 온라인 뉴스의 댓글과 대댓글을 수집해 처음부터 학습한 모델이다. 비슷한 방법으로 KcELECTRA는 더 많은 데이터 세트, 그리고 더 큰 일반 단어를 통해 KcBERT 대비 대부분의 태스크에서 성능을 올린 모델이다.

2.2 BERTopic

BERTopic[18]은 먼저 BERT 모델을 사용하여 텍스트 데이터를 벡터화하고, UMAP(Uniform Manifold Approximation and Projection)[19]과 같은 기법을 사용하여 고차원 벡터를 저차원 공간으로 축소한다. 그리고 HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise)[20]을 사용하여 저차원 임베딩 벡터들을 클러스터링한다. 이를 통해 각 클러스터에서 자주 등장하는 단어들을 분석하여 주제를 대표하는 단어들을 추출하는 모델이다.

* https://huggingface.co/klue/bert-base

** https://huggingface.co/klue/roberta-base

2.3 한국어 리뷰 분석

한국어 리뷰 분석은 온라인 쇼핑 플랫폼에서 소비자들이 남기는 리뷰 데이터를 분석하여 유용한 인사이트를 도출하는 연구 분야로 다양한 산업 분야에서 활발히 연구되고 있다.

영화 리뷰 요약 분석[21]을 진행한 연구에서는 영화 리뷰에서 특징을 추출하고 이를 벡터 공간 모델 또는 특징 벡터로 표현한 후 나이브 베이즈 머신러닝 알고리즘을 사용하여 긍정과 부정으로 리뷰를 분류한다. 이후 가중 그래프 기반 알고리즘을 적용하여 각 리뷰 문장에 대한 순위 점수를 계산해 높은 순위 점수를 기준으로 선택하여 추출 요약을 진행한다. 음식점 리뷰 감정 분석[22]을 진행한 연구에서는 한국어로 작성된 음식점 리뷰를 대상으로, 감성분석을 수행하여 평가 항목별로 세분화된 평점을 제공하는 예측 방법론을 제안한다. 이를 위해, 음식점의 주요 평가항목으로 '음식', '가격', '서비스', '분위기'를 선정하고, 평가항목별 맞춤형 감성사전을 구축한다. 또한 평가항목별 리뷰 문장을 분류하고 감성분석을 통해 세분화된 평점을 예측하여 소비자가 의사결정에 활용 가능한 추가적인 정보를 제공한다.

본 연구에서는 딥러닝 기반 다양한 사전학습 모델들과 GPT-4[8] API을 사용한 라벨 값 재조정을 통하여 한국어 텍스트 리뷰 데이터를 분석한다. 해당 모델은 연구 시점인 2023년 10월 기준 API로 제공하는 생성형 AI 중 가장 높은 성능을 자랑했다. 이를 통해 딥러닝 기반 리뷰 분석 방법의 우수성을 입증하고, 한국어 리뷰 데이터 감성 분석의 정확성을 높이고자 한다. 특히, 딥러닝 사전학습 모델의 미세 조정 최적화와 딥러닝 기반 모델을 통한 리뷰 카테고리 분류를 통해 신뢰성 있는 정보를 제공하고자 한다. 그림1은 본 연구의 전체 분석 프로세스를 나타낸다.

3. 방법론

3.1 데이터 세트 수집 및 전처리

3.1.1 텍스트 리뷰 데이터 크롤링

본 연구에서는 의류 쇼핑 플랫폼 M사에서 구매자의 텍스트 리뷰를 크롤링하는 방식으로 데이터를 수집한다. 이 과정에서 Python 언어와 Selenium, BeautifulSoup 라이브러리 등을 활용하여 웹 스크래핑을 진행한다. 수집한 정보는 리뷰 텍스트 데이터, 기존 구매 만족도 점수 데이터이다. 구체적으로, 2023년 10월 5일부터 9일 기준으로

M사에 업로드된 회원 후기 중 '스타일 후기, '상품 후기', '일반 후기'를 모두 수집한다. 새로 입력될 데이터 세트는 학습 데이터에서 사용하지 않은 특정 상품을 정해 해당 상품의 모든 리뷰를 크롤링하여 사용한다. 각 의류 종류별 랜덤하게 10여 종 정도의 상품을 선택하여 수집한다. 이 때, 구매 만족도 점수가 1~3인 것은 부정(label:0)으로 판단하고 4~5인 것은 긍정(label:1)으로 판단하여 라벨 값을 설정한다.

3.1.2 데이터 라벨 재조정 및 불균형 완화

수집한 데이터는 기존 구매 만족도 점수가 긍정인 쪽에 몰려 분포되어 있다. 심지어 텍스트에는 부정적인 내용이 내포되어 있음에도 구매 만족도 값은 긍정으로 매겨져 있는 경우가 다수 존재한다. 이는 구매자들이 상품에 대한 만족스럽지 못한 부분이 있더라도 리뷰를 달아 혜택을 얻기 위해 깊은 고민 없이 합리적이지 않은 높은 점수를 매겼을 것이라 판단한다. 긍정적인 리뷰와 부정적인 리뷰의 큰 불균형은 추후의 분석 결과를 방해하기 때문에 불균형 완화가 필요하다. 따라서 본 연구에서는 기존 구매 만족도를 재조정하는 방안을 채택한다. 기존 구매 만족도 재조정은 GPT-4의 API를 사용하여 진행한다. 이때 사용한 프롬포트는 표1과 같으며 temperature 값은 0.5를 사용한다.

본 연구에서는 GPT-4가 제시해준 라벨 값과 사람 5명의 판단을 비교하여 GPT-4가 제시한 라벨의 정확성을 검증한다. GPT-4[8] API를 사용하여 라벨 값을 재조정하여 합리적이지 않은 사용자 만족도 점수를 처리한 후에도 긍정적인 리뷰가 부정적인 리뷰에 비해 개수가 훨씬 많이 존재한다. 따라서 구축할 예측 모델의 정확성을 향상하기 위해 데이터의 불균형을 다음과 같은 방법으로 완화한다. 긍정과 부정의 비율을 약 6:4로 설정하여 긍정 리뷰의 개수를 그 비율에 맞게 랜덤하게 삭제하는 작업을 진행한다. 이외에도 정규 표현식을 적용하여 한글 외의 문자 등을 제거하는 전처리를 통해 최종적으로 약 10,000개의 학습 데이터 세트를 구축한다.

3.2 미세 조정: 한국어 텍스트 감성 분류

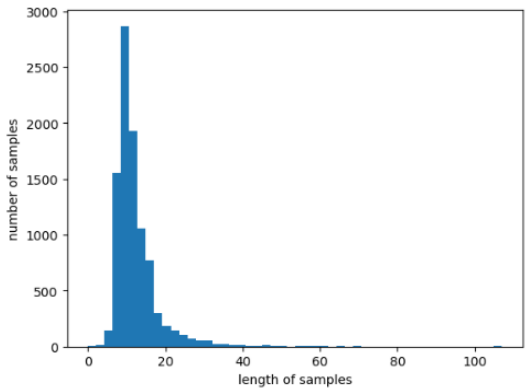
본 연구에서는 감성 분류를 위해 한국어 기반으로 사전학습된 BERT 기반의 다양한 모델을 사용하여 입력 시퀀스 길이에 따라 다양하게 미세 조정을 진행한다. 입력 시퀀스 길이는 모델이 한 번에 처리할 수 있는 텍스트의 최대 길이를 의미하며, 자연어 처리 모델에서 중요한 역할을 한다.

3.2.1 입력 시퀀스 설정

입력 시퀀스 길이를 설정하면 이보다 더 긴 텍스트는 잘리고, 더 짧은 텍스트는 패딩(padding)된다. 이는 모델이 고정된 길이의 입력을 필요로 하기 때문이다. 예를 들어, 입력 시퀀스 길이를 16로 설정한 경우, 16보다 긴 텍스트는 처음부터 16까지만 사용되고 나머지는 잘리게 된다. 반면, 16보다 짧은 텍스트는 전체 텍스트가 모두 입력되고 남은 부분은 패딩 토큰으로 채워 고정된 길이로 맞춘다. 패딩 토큰은 일반적으로 '[PAD]'와 같은 특별한 토큰으로, 실제로는 의미 없는 값이다. 이러한 과정은 모델이 일관된 입력 크기를 가지도록 하여 계산 효율성을 높이고, 패딩된 부분은 학습 과정에서 무시되도록 설계된다. 텍스트의 길이 특성과 모델의 메모리 사용량을 고려하여 입력 시퀀스 길이를 적절히 설정하는 것은 모델의 성능과 효율성에 큰 영향을 미친다. 적절한 입력 시퀀스 길이를 선택하면 중요한 정보 손실을 최소화하고, 불필요한 메모리 사용을 줄일 수 있기 때문이다. 본 연구에서는 다양한 입력 시퀀스 길이를 설정하여 각각의 경우에 대해 감성 분류 성능을 비교 분석한다. 해당 연구의 리뷰 길이 분포는 그림 2와 같이 평균 길이 약 12, 최대 길이 107이다. 따라서, 16, 32, 64, 128으로 설정하여 결과를 비교한다.

(표 1) 구매 만족도 점수 라벨 재조정을 위한 프롬포트
(Table 1) Prompt for recalibrating customer satisfaction scores

Role	Prompt
System	You are an AI language model that analyzes and detects customer sentiments in fashion clothing reviews.
User	Analyze the following clothing reviews and determine whether each customer's sentiment is positive or negative. Respond with either 'positive' or 'negative' without any additional explanation.



(그림 2) 수집된 리뷰 데이터의 길이 분포
(Figure 2) Distribution of review data lengths collected

3.2.2 BERT기반 모델의 미세 조정

미세 조정에 사용할 모델은 허깅페이스에서 제공하는 klue/bert, klue/roberta, kykim/albert-kor*, beomi/kcbert**, beomi/KcELECTRA*** 모델이며, 모두 base 모델을 사용한다. 이때 입력 시퀀스 길이에 따른 성능과 메모리 소모량 측정에 초점을 둔다.

3.3 리뷰 카테고리 분류

본 연구에서는 새로 입력될 리뷰 데이터에 대해 카테고리별로 감성 점수를 나타내고자 BERTopic[18]을 사용하여 토픽 모델링을 진행한다. 다양한 토픽 개수에 대해 모델을 학습시키고 각 모델의 적합도를 그래프로 그린 후, 그래프에서 굴곡이 나타나는 지점을 최적의 토픽 개수로 선택하는 엘보우 방법(Elbow Method)으로 토픽의 개수를 선정한다. 이를 통해, 리뷰 데이터 세트에 속해 있는 카테고리를 파악해 Okt 형태소 분석기****를 사용하여 분류된 카테고리에서 개수가 많이 나온 명사를 기준으로 카테고리별 키워드 사전을 구축한다. 테스트 세트를 해당되는 명사가 포함된 카테고리로 해당 리뷰를 분류한다. 이때, 한 리뷰 안에 여러 카테고리에 해당하는 명사가 존재할 때에는 중복을 허용하여 두 개 이상의 카테고리로 분류한다.

* https://huggingface.co/kykim/albert-kor-base
** https://huggingface.co/beomi/kcbert-base
*** https://huggingface.co/beomi/KcELECTRA-base
**** https://github.com/open-korean-text/open-korean-text

3.4 감성 점수 추출

마지막으로 미세 조정을 통해 얻은 최적의 모델 가중치와, 토픽 모델링과 형태소 분석기를 통해 구축한 카테고리별 키워드 사전을 활용하여 최종적으로 새로 입력될 리뷰 데이터에서 요약하고자 하는 제품에 대한 키워드별 감정 점수를 추출한다. 먼저, 재조정한 부정, 긍정 라벨 값을 각각 0과 1로 지정한 다음, 원 핫 인코딩을 한 후 부정인 경우와 긍정인 경우의 확률값을 sotfmax 함수를 활용하여 구한다. 다시 말해, 리뷰 한 개당 부정과 긍정의 정도가 각각 몇 퍼센트의 비중을 차지하는지를 나타내었으며 각 확률의 합은 1이 된다.

$$p_i = softmax(z_i) \tag{1}$$
$$= \left[\frac{e^{z_i[0]}}{e^{z_i[0]} + e^{z_i[1]}}, \frac{e^{z_i[1]}}{e^{z_i[0]} + e^{z_i[1]}} \right]$$

위의 수식(1)은 리뷰 i에 대한 모델의 출력 로짓 z_i에 Softmax 함수를 적용하여 확률값 p_i를 구하는 과정을 나타낸다.

$$Sentiment_i = \begin{cases} \mathrm{Neg}, & \mathrm{if}\, p_i[0] > p_i[1] \\ Pos, & \mathrm{if}\, p_i[1] > p_i[0] \end{cases} \tag{2}$$

$$Score_i = \max(p_i[0], p_i[1]) \tag{3}$$

다음으로, 수식(2)와 같이 Sentiment는 각 리뷰의 부정과 긍정의 정도를 비교하여 더 확률이 높은 값을 택하여 해당 리뷰가 부정인지 긍정인지 판단하고, 최종적인 감정 점수 Score는 수식(3)과 같이 더 높은 확률값을 부정과 긍정의 정도(퍼센트)로 활용한다. 하나의 제품마다 총 요약된 감정 점수를 제시하기 위하여 각 리뷰별로 앞에서 구한 확률값을 모두 더하고 총 리뷰 개수로 나누어 최종적인 감정 점수를 추출한다.

4. 실험 설정

4.1 실험 환경

본 연구에서는 사전학습된 모델의 가중치를 본 연구에서 사용하는 리뷰 데이터에 맞게 세밀하게 조정해 성능을 향상시키기 위하여 미세 조정을 진행한다. 미세 조정을 진행한 사전학습 모델은 ALBERT-kor, KcELECTRA,

KcBERT, KLUE-RoBERTa, KLUE-BERT 총 5가지의 base 모델이다. 훈련 설정은 Adam 옵티마이저, 학습률은 2e-5, 손실 함수는 이중 분류이므로 binary cross entropy를 사용한다. 또한, batch_size는 16, 검증 분할은 20%, epoch 수는 early stopping을 적용하여 patience를 10으로 설정한다. 입력 시퀀스 길이(max_length)는 각 모델에 따라 16, 32, 64, 128으로 변화를 주어 padding을 각각 진행한 후 성능과 메모리 사용량을 집중적으로 비교한다.

4.2 평가 지표

평가 지표로는 분류 평가 지표인 정확도(Accuracy), 재현율(Recall), 정밀도(Precision), F1-스코어(F1-Score)를 사용한다. 정확도는 실제 데이터가 예측 데이터와 얼마나 같은지를 판단하는 지표이고, 재현율은 실제 값이 사실인 대상 중 예측을 사실로 일치한 데이터의 비율을 나타낸다. 정밀도는 예측을 사실로 한 대상 중 실제로 사실인 데이터의 비율을 나타내며, 재현율이 높아지면 정밀도는 낮아지고 재현율이 낮아지면 정밀도는 높아지는 반비례 관계를 가지고 있다. 따라서 F1-스코어를 통하여 정밀도와 재현율의 관계 확인이 가능하며 둘 모두 어느 한 쪽으로 치우치지 않는 수치를 나타낼 경우 상대적으로 높은 값을 갖는다.

5. 실험 결과

5.1 BERT 기반 모델의 미세 조정

다양한 한국어 기반의 BERT 모델들이 서로 다른 최대 입력 길이(Max Length)에서 어떠한 성능과 메모리 변화를 보이는지 비교하였다. 표2는 성능에 대한 결과, 표3은 메모리 사용량에 대한 결과를 보여준다. 각 모델은 정확도, 재현율, 정밀도, 그리고 F1-스코어 값으로 평가되었으며, early stopping을 적용하여 학습된 epoch도 함께 비교하였다.

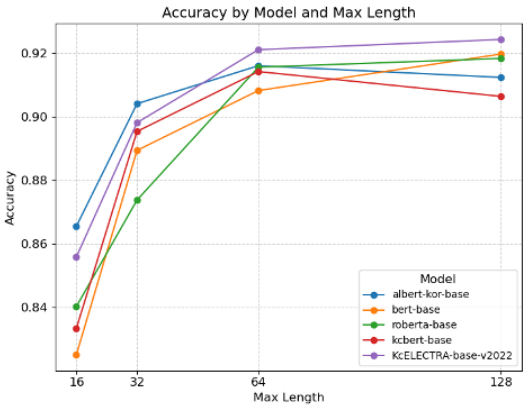
ALBERT-kor 모델의 경우, 최대 입력 시퀀스 길이 16보다 최대 길이 64에서 정확도 0.9160, 재현율 0.9160, 정밀도 0.9160, F1-스코어 0.9155로 성능이 상승하였다. 최대 길이 128에서는 정확도 0.9123, 재현율 0.9123, 정밀도 0.9123, F1-스코어 0.9140을 기록하였다. 또한, 최대 길이 16에서 165,893MB, 64에서 181,046MB, 128에서 183,342MB의 메모리를 사용하였다. 해당 모델의 가장 우수한 성능은 최대 길이 64에서 나타났다. KLUE-BERT 모델의 메모리 사용량은 최대 길이 16에서 1,343,479MB, 64에서 1,409,345MB, 128에서 1,434,165MB로 비교적 일정한 수준을 유지하였으며, 최대 길이 128에서 최고의 성능을 나타냈다. KLUE-RoBERTa 모델은 최대 길이 128에서 정확도 0.9183, F1-스코어 0.9173을 기록하여 가장 우수한 성능을 보였다. 메모리 사용량은 최대 길이 16에서 1,426,852MB, 64에서 1,388,385MB, 128에서 1,397,415MB였다. KcBERT 모델은 최대 길이 16보다 64에서 정확도 0.9141, F1-스코어 0.9130로 성능이 향상되었다. 메모리 사용량은 최대 길이 16에서 1,399,659MB, 64에서 1,392,244MB, 128에서 1,397,315MB였다. 해당 모델은 최대 길이 64에서 최고의 성능을 보였다.

KcELECTRA 모델은 모든 길이에서 우수한 성능을 보였으며, 최대 길이 16에서 정확도 0.8558, F1-스코어 0.8507를 기록하였고, 최대 길이 64에서는 정확도 0.9210, F1-스코어 0.9201로 성능이 향상되었다. 최대 길이 128에서는 정확도 0.9242, F1-스코어 0.9239로 최고의 성능을 보였다. 이 모델의 메모리 사용량은 최대 길이 16에서 1,590,798MB, 64에서 1,625,487MB, 128에서 1,662,332MB로 증가하였다. 해당 모델의 가장 우수한 성능은 최대 길이 128에서 나타났다.

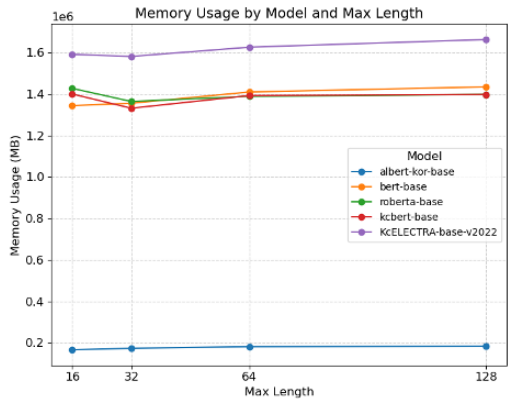
하지만, KcELECTRA 모델은 최대 길이 128에서 최고의 성능을 기록했으나 메모리 사용량 또한 가장 많았다. 최대 길이를 64에서 128로 증가시키는 과정에서 메모리 사용량은 대체적으로 증가하는 반면, 성능은 변화가 거의 없거나 오히려 감소하는 모델도 존재하였기에 메모리 효율을 고려할 때 64를 가장 최적의 길이로 판단하였다.

그림 3의 그래프는 각 모델의 최대 입력 길이 증가에 따른 정확도 변화를 보여준다. 모든 모델에서 입력 길이가 증가함에 따라 성능이 향상되는 경향을 보인다. 특히, KcELECTRA 모델은 모든 길이에서 높은 성능을 유지하며, 최대 길이 128에서 최고의 성능을 기록하였다. 그림 4의 그래프는 각 모델의 최대 입력 길이 증가에 따른 메모리 사용량 변화를 나타낸다. 대부분의 모델은 입력 길이에 따라 메모리 사용량이 크게 증가하지 않지만, ALBERT-kor 모델과 KcELECTRA 모델은 상대적으로 낮은 메모리 사용량에서 시작해 최대 길이 64에서 큰 폭으로 증가하였다. 특히 KcELECTRA 모델은 메모리 사용량이 전체적으로 높아지는 경향을 보였다.

모델의 최대 입력 시퀀스 길이가 증가함에 따라 성능이 전반적으로 향상되는 경향을 보인다. 따라서, 특정 애플리케이션에서 높은 성능이 요구되는 경우, 최대 입력 길이를 늘리는 것이 효과적일 수 있다. 하지만, 메모리 사용량은 입력 길이와 함께 증가하는 경향이 있다. 이는 메



(그림 3) BERT 기반 모델 리뷰 데이터 미세 조정의 입력 시퀀스 길이에 따른 성능 변화 그래프
(Figure 3) Performance variation graph according to input sequence lengths for fine-tuninig review data with BERT-based models



(그림 4) BERT 기반 모델 리뷰 데이터 미세 조정의 입력 시퀀스 길이에 따른 메모리 사용량 변화 그래프
(Figure 4) Memory usage variation graph according to iput sequence lengths for fine-tuning review data with BERT-based models

(표 2) BERT 기반 모델의 입력 시퀀스 길이에 따른 리뷰 데이터 미세 조정 성능표
(Table 2) Performance table of fine-tuning review data according to input sequence lengths of BERT-based Models

Model	Max Length	Accuracy	Recall	Precision	F1-Score	Epochs
ALBERT-kor	16	0.8655	0.8655	0.8663	0.8663	12
	32	0.9040	0.9040	0.9040	0.9038	13
	64	0.9160	0.9160	0.9160	0.9155	11
	128	0.9123	0.9123	0.9123	0.9140	12
KLUE-BERT	16	0.8251	0.8251	0.8310	0.8269	11
	32	0.8893	0.8893	0.8903	0.8870	11
	64	0.9082	0.9082	0.9090	0.9085	12
	128	0.9196	0.9196	0.9201	0.9185	12
KLUE-RoBERTa	16	0.8402	0.8402	0.8390	0.8363	12
	32	0.8737	0.8737	0.8732	0.8715	12
	64	0.9155	0.9155	0.9153	0.9146	12
	128	0.9183	0.9183	0.9184	0.9173	13
KcBERT	16	0.8334	0.8334	0.8393	0.8351	12
	32	0.8953	0.8953	0.8958	0.8934	12
	64	0.9141	0.9141	0.9144	0.9130	12
	128	0.9063	0.9063	0.9062	0.9062	11
KcELECTRA	16	0.8558	0.8558	0.8583	0.8507	12
	32	0.8981	0.8981	0.9020	0.8951	12
	64	0.9210	0.9210	0.9212	0.9201	12
	128	0.9242	0.9242	0.9239	0.9239	12

(표 3) BERT 기반 모델의 입력 시퀀스 길이에 따른 리뷰 데이터 미세 조정 메모리 사용량 측정표
(Table 3) Memory usage measurement table for fine-tuning review data according to input sequence lengths of BERT-based models

Max Length	Memory Usage(MB)				
	ALBERT-kor	KLUE-BERT	KLUE-RoBERTa	KcBERT	KcELECTRA
16	165,893	1,343,479	1,426,852	1,399,659	1,590,798
32	173,179	1,355,026	1,363,567	1,331,035	1,580,752
64	181,046	1,409,345	1,388,385	1,392,244	1,625,487
128	183,342	1,434,165	1,397,415	1,397,315	1,662,332

모리 자원이 제한된 환경에서는 중요한 고려 사항이 된다.

결론적으로, 모델의 최대 입력 시퀀스 길이가 증가함에 따라 성능이 전반적으로 향상되는 경향을 나타내며, 특히 KcELECTRA 모델이 모든 지표에서 우수한 성능을 보임을 알 수 있다. 그러나 메모리 사용량도 함께 증가하므로, 이를 고려하여 최적의 모델과 최대 입력 길이를 설정하는 것이 필요함을 시사한다.

5.2 카테고리별 감성 점수 추출

토픽 모델링을 통해 리뷰 데이터 세트에 어떤 토픽이 속해 있는지 확인한다. 토픽 1,2,7는 핏, 토픽 3은 색감, 토픽 4는 품질, 토픽 5는 배송, 토픽 6은 디자인으로 판단하여 총 6가지로 카테고리를 분류하였다. 한국어 형태소 분석기 Okt로 명사를 추출해 개수가 많이 나온 명사를 기준으로 카테고리별 키워드 사전을 구축하였다. 예를 들어, 핏에는 [‘사이즈’, ‘기장’, ’폼’] 색감에는 [‘색감’, ’색상’, ’명도’] 품질에는 [‘재질’, ‘두께’, ‘촉감’] 배송에는 [‘배송’, ‘포장’, ‘교환’] 디자인에는 [‘디자인’, ‘스타일’, ‘심플’] 등의 명사가 포함한다. 키워드 사전을 구축한 후, 카테고리별로 해당되는 명사가 포함된 리뷰를 각 카테고리로 분류하였다.

최종적으로 최대 시퀀스 길이를 64로 지정한 KcELECTRA를 미세 조정한 모델로 새로 입력될 리뷰의 제품을 임의로 선정하여 해당하는 제품의 카테고리별 감성 점수를 추출한 결과 예시는 표 4와 같다.

(표 4) 카테고리별 감성 점수 추출 예시
(Table 4) Example table of sentiment scores by category

Category	Result
Price	75.88% negative
Design	80.52% positive
Delivery	66.49% negative
Color	81.90% positive
Material	79.00% positive
Fit	76.10% positive

6. 결론 및 향후 연구

본 연구는 BERT 기반의 자연어처리 모델을 한국어 리뷰 데이터에 적용하여 감성 분석을 수행하고, 입력 시퀀스 길이를 최적화하는 방법을 탐구하였다. 이를 위해 의류 쇼핑 플랫폼 M사의 텍스트 리뷰 데이터를 수집하여 GPT-4 와 같은 생성형 AI를 통해 구매 만족도 값을 재조정하고 다양한 BERT 기반 모델을 미세 조정하였다. 실험 결과, 입력 시퀀스 길이가 64와 128 일 때 대부분의 모델이 가장 높은 성능을 보였으며, 특히 KcELECTRA 모델이 입력 길이 64와 128 모두 정확도 92%이상으로 가장 우수한 성능을 기록하였다. 메모리 사용량은 입력 길이가 128일 때가 가장 많았다. 이는 한국어 의류 쇼핑 플랫폼, 혹은 길이가 평균 약 12인 리뷰 데이터에 대해서 메모리 사용량과 정확도를 함께 고려한다면, KcELECTRA 모델의 입력 시퀀스 길이를 64로 설정하여 미세 조정 할 때 가장 적합한 결과를 제공할 것임을 시사한다.

나아가, 새로 입력될 리뷰 데이터에 대해 BERTopic를 활용하여 키워드 사전을 구축한 후 리뷰의 카테고리를 분류하고 Softmax 함수를 적용하여 확률값을 산출하여 정확한 감성 점수를 추출하는 방향성을 제시함으로써, 실제 구축된 최적의 모델로 감성 분석을 위한 실질적인 프레임워크를 제공하였다. 이를 통해 사용자들은 각 제품에 대한 세부적인 감성 정보를 얻을 수 있게 되며 결과적으로 소비자 만족도를 높이는 데 기여할 것이다.

마지막으로, 본 연구에서는 새로 입력될 리뷰 데이터

세트를 카테고리화하는 과정에서 한 리뷰가 여러 카테고리에 속할 경우 중복을 허용하여 여러 카테고리에 동시에 분류하는 방법을 취하였다. 하지만 이는 여러 정보를 포함할 수 있으므로 해당 리뷰 카테고리에 대한 점수만 추출하는 데에 한계점이 존재한다. 따라서, 텍스트 내 존재하는 여러 속성들에 대해 감성을 추출하는 속성 기반 감성 분석(ABSA)[23] 기법을 활용한다면 더욱 정확한 감성 분석을 할 수 있을 것이라 기대한다.

참고문헌(Reference)

[1] Fernandes, Semila et al., “Measuring the impact of online reviews on consumer purchase decisions – A scale development study,” Journal of Retailing and Consumer Services, Vol.68, 2022.
https://doi.org/10.1016/j.jretconser.2022.103066

[2] Chen, Tao et al., “The impact of online reviews on consumers’ purchasing decisions: Evidence from an eye-tracking study,” Frontiers in Psychology, Vol.13, 2022.
https://doi.org/10.3389/fpsyg.2022.865702

[3] Kutabish, Saleh, Ana Maria Soares, and Beatriz Casais, “The influence of online ratings and reviews in consumer buying behavior: a systematic literature review,” in Proc. of International Conference on Digital Economy, Springer, Cham, pp.113-136, 2023.
https://doi.org/10.1007/978-3-031-42788-6_8

[4] Wankhade, Mayur, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni, “A survey on sentiment analysis methods, applications, and challenges,” Artificial Intelligence Review, Vol.55, No.7, pp.5731-5780, 2022.
https://doi.org/10.1007/s10462-022-10144-1

[5] Devlin, Jacob, et al., “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
https://doi.org/10.48550/arXiv.1810.04805

[6] Kulkarni, Ajay, Deri Chong, and Feras A. Batarseh. “Foundations of data imbalance and solutions for a data democracy,” Data democracy, pp.83-106, Academic Press, 2020.
https://doi.org/10.1016/B978-0-12-818366-3.00005-8

[7] Conglong Li, Minjia Zhang, and Yuxiong He, “The stability-efficiency dilemma: Investigating sequence length warmup for training GPT models,” Advances in Neural Information Processing Systems, vol.35, pp.26736-26750, 2021.
https://doi.org/10.48550/arXiv.2108.06084

[8] Achiam, Josh et al., “Gpt-4 technical report,” arXiv preprint arXiv:2303.08774, 2023.
https://doi.org/10.48550/arXiv.2303.08774

[9] Hariri, Walid, “Sentiment Analysis of Citations in Scientific Articles Using ChatGPT: Identifying Potential Biases and Conflicts of Interest,” arXiv preprint arXiv:2404.01800, 2024.
https://doi.org/10.48550/arXiv.2404.01800

[10] Fatouros, Georgios et al., “Transforming sentiment analysis in the financial domain with ChatGPT,” Machine Learning with Applications, Vol.14, 2023.
https://doi.org/10.1016/j.mlwa.2023.100508

[11] Jin, Jian, Ping Ji, and Chun Kit Kwong, “What makes consumers unsatisfied with your products: Review analysis at a fine-grained level,” Engineering Applications of Artificial Intelligence, Vol.47, p.38-48, 2016.
https://doi.org/10.1016/j.engappai.2015.05.006

[12] Qi, Jiayin et al., “Mining customer requirements from online reviews: A product improvement perspective,” Information & Management, Vol.53, No.8, pp.951-963, 2016.
https://doi.org/10.1016/j.im.2016.06.002

[13] Kwark, Young, Jianqing Chen, and Srinivasan Raghunathan, “User-generated content and competing firms’ product design,” Management Science, Vol.64, No.10, pp.4608-4628, 2018.
https://doi.org/10.1287/mnsc.2017.2839

[14] Vaswani, Ashish et al., “Attention is all you need,” in Proc. of 31st Conference on neural information processing systems, Vol.30, 2017.
https://doi.org/10.48550/arXiv.1706.03762

[15] Liu, Yinhan et al., “Roberta: A robustly optimized bert pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.
https://doi.org/10.48550/arXiv.1907.11692

[16] Lan, Zhenzhong et al., “Albert: A lite bert for self-supervised learning of language representations,” arXiv:1909.11942, 2019. https://doi.org/10.48550/arXiv.1909.11942

[17] Park, Sungjoon et al., “Klue: Korean language understanding evaluation,” arXiv:2105.09680, 2021. https://doi.org/10.48550/arXiv.2105.09680

[18] Grootendorst, Maarten, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” arXiv:2203.05794, 2022. https://doi.org/10.48550/arXiv.2203.05794

[19] McInnes, Leland, John Healy, and James Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” arXiv:1802.03426, 2018. https://doi.org/10.48550/arXiv.1802.03426

[20] McInnes, Leland, John Healy, and Steve Astels, “hdbscan: Hierarchical density based clustering,” J. Open Source Softw., Vol.2, No.11, 2017. http://dx.doi.org/10.21105/joss.00205

[21] Khan, Atif et al., “Movie Review Summarization Using Supervised Learning and Graph-Based Ranking Algorithm,” Computational intelligence and neuroscience, Vol.2020, No.1, 2020. https://doi.org/10.1155/2020/7526580

[22] So, Jin-Soo, and Pan-Seop Shin, “Rating prediction by evaluation item through sentiment analysis of restaurant review,” Journal of the Korea Society of Computer and Information, Vol.25, No.6, pp.81-89, 2020. https://doi.org/10.9708/jksci.2020.25.06.081

[23] Pontiki, Maria et al., “Semeval-2016 task 5: Aspect based sentiment analysis,” in Proc. of Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, 2016. https://doi.org/10.18653/v1/S16-1002

◐ 저 자 소 개 ◑

황 성 아(Sunga Hwang)
2019년~2023년 연세대학교 미래캠퍼스 정보통계학과 학사
2023년~현재 연세대학교 정보대학원 비즈니스 빅데이터 분석 트랙 석사과정
관심분야 : Natural Language Processing, Deep Learning
E-mail : sungahwang@yonsei.ac.kr

박 세 연(Seyeon Park)
2018년~2023년 동국대학교 경영정보학과 학사
2023년~현재 연세대학교 정보대학원 비즈니스 빅데이터 분석 트랙 석사과정
관심분야 : Natural Language Processing, Deep Learning
E-mail : seyeon@yonsei.ac.kr

장 백 철(Beakcheol Jang)
2009년 North Carolina State University 컴퓨터공학과(공학박사)
2021년~현재 연세대학교 정보대학원 교수
관심분야 : Natural Language Processing, Artificial Intelligence, Bigdata Analytics
E-mail : bjang@yonsei.ac.kr