The Korean Journal of Applied Statistics
2023, Vol. 36, No. 6, 515–528

DOI: https://doi.org/10.5351/KJAS.2023.36.6.515
Print ISSN1225-066X / Online ISSN 2383-5818

A study on the aspect-based sentiment analysis of multilingual customer reviews

Sungyoung Ji[a], Siyoon Lee[a], Daewoo Choi[a], Kee-Hoon Kang[1,a]

[a]Department of Statistics, Hankuk University of Foreign Studies

Abstract

With the growth of the e-commerce market, consumers increasingly rely on user reviews to make purchasing decisions. Consequently, researchers are actively conducting studies to effectively analyze these reviews. Among the various methods of sentiment analysis, the aspect-based sentiment analysis approach, which examines user reviews from multiple angles rather than solely relying on simple positive or negative sentiments, is gaining widespread attention. Among the various methodologies for aspect-based sentiment analysis, there is an analysis method using a transformer-based model, which is the latest natural language processing technology. In this paper, we conduct an aspect-based sentiment analysis on multilingual user reviews using two real datasets from the latest natural language processing technology model. Specifically, we use restaurant data from the SemEval 2016 public dataset and multilingual user review data from the cosmetic domain. We compare the performance of transformer-based models for aspect-based sentiment analysis and apply various methodologies to improve their performance. Models using multilingual data are expected to be highly useful in that they can analyze multiple languages in one model without building separate models for each language.

Keywords: BERT, multilingual BERT, natural language process, transformer encoder, XLM-RoBERTa

1. 서론

자연어 처리(natural language process; NLP)는 우리가 일상생활에서 사용하는 다양한 언어를 분석하고 이해할 수 있도록 컴퓨터가 처리하는 모든 기술을 의미한다. NLP 기술은 크게 2가지로 나눌 수 있는데 자연어 이해(natural language understanding; NLU)와 자연어 생성(natural language generation; NLG) 분야가 있다. NLU 분야에서 적용할 수 있는 업무는 문서 분류, 기계 독해, 감성 분석, 유사도 측정 등이 있으며 NLG 분야에서는 문서 요약, 대화문 생성, 문서 번역 등이 있다. 이러한 NLP 솔루션은 은행, 보험사와 같은 금융권의 경우 고객과의 상담 내용 분류 및 요약 더 나아가 고객의 질문에 쉽게 대응하기 위한 챗봇 등에 활용되고 있으며 이외에도 사내 검색 시스템 구축, 검색어 자동 완성, 보고서 자동 요약 등 활용할 수 있는 분야가 다양하다.

한편, 최근 전자상거래 시장의 성장과 더불어 사용자가 작성한 후기에 대한 분석 니즈가 증가하고 있다. 대부분의 사람들은 상품 구매나 서비스 이용 시 다른 사람이 작성한 후기를 참고하여 의사결정을 한다. 대량으로 쌓이는 후기를 사람이 일일이 살펴보고 분석하는 것은 한계가 있기 때문에 효율적으로 분석하는 방법이 필요하다. 기존 분석 방법은 사용자가 작성한 후기에 대해서 단순한 감성분석(sentiment analysis)을 진행하는 것인데 예를 들면, “향은 잘 모르겠는데 향 지속력은 좋아서 마음에 들어요” 와 같은 사용자 후기의 경우 해당

This work was supported by Hankuk University of Foreign Studies Research Fund of 2023.
[1]Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Cheoin-koo, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: khkang@hufs.ac.kr

Published 31 December 2023 / journal homepage: http://kjas.or.kr

후기를 단순히 긍정으로 분류하는 것이다. 하지만 최근 들어 속성기반 감성분석의 방법이 주목받고 있는데 이 경우 위의 후기를 향은 중립, 향 지속력은 긍정의 속성별로 감성을 분류하는 방식이다. 이처럼 속성기반 감성분석의 경우 사용자의 후기를 다면적으로 분석할 수 있다는 점에서 최근 주목받고 있다.

이에 본 논문에서는 최신 자연어 처리 알고리즘인 트랜스포머 인코더(transformer encoder) (Vaswani 등, 2017) 계열의 모델을 활용하여 다국어 데이터에 대한 속성기반 감성분석 결과를 비교 분석하고자 한다. 한편, 트랜스포머 인코더 계열 모델 중 하나인 BERT (bidirectional encoder representations from transformers) (Devlin 등, 2019)를 활용하여 속성기반 감성분석을 진행하는 선행 연구들이 있었지만 대부분 영어만을 대상으로 하였다. 본 논문에서는 사용자 후기는 영어뿐만 아니라 다른 다양한 언어로 작성될 수 있다는 점을 고려하여 다양한 다국어 데이터를 활용하여 실험을 진행하였다. 즉, 속성기반 감성분석이 하나의 다국어 모델로 여러 언어에 적용될 수 있는지 확인하였다.

본 논문의 2장에서는 트랜스포머 인코더 계열에서 많이 사용되는 BERT 모델과 속성기반 감성분석 내용을 소개한다. 3장에서는 BERT를 활용한 속성기반 감성분석 모형을 기술하고 4장에서는 활용한 데이터와 모형의 평 지표에 대하여 설명하고 성능 향상을 위한 비교 실험과 다국어 데이터에 대한 실험 결과를 분석한다. 5장에서는 간단한 결론과 향후 연구에 대해 언급한다.

2. 선행 연구

이 장에서는 최신 자연어 처리 알고리즘인 BERT와 속성기반 감성분석이 무엇인지 또한 관련된 선행연구는 어떤 것이 있는지 설명한다.

2.1. BERT

자연어 처리 알고리즘은 트랜스포머의 등장 전과 후로 나눌 수 있다. 트랜스포머가 등장하기 이전의 NLP 연구는 RNN, CNN 계열의 딥러닝 모델을 중심으로 연구되어왔다. 하지만 입력 데이터가 커지면 병렬처리가 불가능하여 계산이 오래 걸린다는 단점이 있고 기존 모델들은 성능이 높지 않아 여러 산업 분야에서 활용되기는 어려웠다. 하지만 Vaswani 등 (2017)에서 트랜스포머 모델이 소개된 이후 NLU 분야에서는 BERT를 NLG 분야에서는 Brown 등 (2020)을 필두로 다양한 언어 모델 연구가 진행되어왔다. 특히 구글에서 발표한 트랜스포머 모델의 인코더 부분만 활용한 언어 모델인 BERT는 당시 대부분의 NLP 과업에서 최고 성능을 달성하며 NLP 역사의 한 획을 그었다고 평가받는다.

BERT는 사전학습(pre-training)과 미세조정(fine-tuning) 2가지 단계로 학습이 진행된다. BERT 모델의 사전학습과 미세조정 구조를 이해하기 위한 도식은 Devlin 등 (2018)을 참고하면 된다. 사전학습 단계에서는 MLM (masked language model)과 NSP (next sentence prediction) 2가지 과제를 통해 문맥을 이해하게 된다. MLM 과제는 문장에서 특정 토큰(token)을 감추고(masking) 해당 토큰을 맞추는 학습 방법이며 NSP 과제는 두 개의 문장이 이어지는 관계인지 아닌지를 학습하는 방법이다. 이 과정을 통해서 자연스럽게 BERT는 문맥을 이해할 수 있게 된다. 이렇게 언어 전반에 대한 이해를 마치고 나면 최종 적용할 과제에 맞게 미세조정을 진행해 주면 된다. 미세조정 과제의 경우 특정 토큰을 통해 분류를 진행하거나 특정 토큰의 시작과 끝을 예측하는 등의 방법을 통해 분류 문제, QA 문제, NER (name entity recognition) 등과 같은 문제를 해결할 수 있다. 본 논문에서는 속성기반 감성분석을 해결하기 위해 분류 문제의 접근 방식을 활용하였으며 다국어 데이터를 다루기 위해서 공개된 다국어 언어 모델을 활용하여 비교 분석하였다.

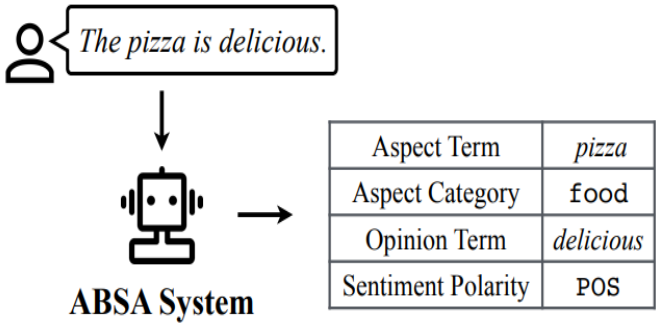

Figure 1: *Example of aspect-based sentiment analysis task.*

2.2. 속성기반 감성분석

속성기반 감성분석(aspect-based sentiment analysis; ABSA)은 기존 단순 감성분석의 방법에서 더 나아가 속성별로 감성분석을 진행하는 방법을 말한다. 단순 감성분석에서는 특정 문장 전체의 긍/부정에 대한 분류만을 하는 반면 속성기반 감성분석은 문장의 내용을 속성별로 세분화하여 각 속성에 대한 긍/부정 여부를 분석하는 것이다.

한편, Zhang 등 (2022)에 의하면 Figure 1에서 볼 수 있듯이 속성기반 감성분석 과제(task)도 1) 속성 용어 추출 2) 속성 카테고리 추출 3) 감성 용어 추출 4) 속성 감성분류로 나눌 수 있다. “The pizza is delicious”의 예문에서 pizza는 문장에서 직접적으로 드러난 속성 용어에 해당되고 food는 pizza라는 속성 용어가 포함되는 속성 카테고리를 의미한다. 즉, 속성 카테고리가 좀 더 상위개념에 해당된다. 한편 최근에는 이러한 개별 과업들을 동시에 진행하는 복합 과제가 최근 주목받고 있는데 특히 End-to-End ABSA의 경우 속성 카테고리를 추출하고 해당 속성 카테고리에 대한 감성분류까지 진행하는 방식을 의미한다. 본 논문에서는 위의 예시처럼 “향은 중립, 향 지속력은 긍정”과 같이 다면적으로 분석하는 방법인 End-to-End ABSA 과제를 중점적으로 다루었다.

이 방법은 두 가지 과제를 동시에 해결하는 방법이기 때문에 까다롭기는 하지만 분석 결과가 의미있기 때문에 많이 연구되고 있다. 해당 End-to-End ABSA 과제 수행을 위해 먼저 속성 카테고리를 추출하고 속성 카테고리에 대한 감성분석 수행의 순서로 진행할 수도 있지만 모델을 두 번 사용해야 하고 자연스럽게 모델 성능 문제와 관리의 어려움이라는 단점이 발생한다. 그래서 입력 데이터의 형태를 변형하여 한 번에 속성 카테고리 추출과 감성분석을 적용할 수 있는 방법인 Sun 등 (2019)에서 제안된 ABSA-BERT-pair 방식을 활용하여 진행하였다.

3. 모형

본 장에서는 속성기반 감성분석 과제 수행을 위해서 활용한 BERT의 미세조정(fine-tuning) 과제 중 하나인 문장-쌍(sentence pair) 분류 아키텍처에 대해 설명하고 본 논문에서 사용한 다국어 모델에 대해서 설명한다.

3.1. BERT의 미세조정

BERT 모델의 미세조정 방식은 최종 과제에 따라 달라지는데 본 연구에서는 2.2절에서 설명하였듯이 End-to-End ABSA 과제 수행을 위해 Devlin 등 (2019)에서 제안된 문장-쌍 분류 아키텍처를 활용하였다. 문장-쌍 분류 아키텍처는 문장 A와 B 두 개를 함께 넣어주고 두 개의 문장 사이의 관계가 어떤지 분류하는 형식의

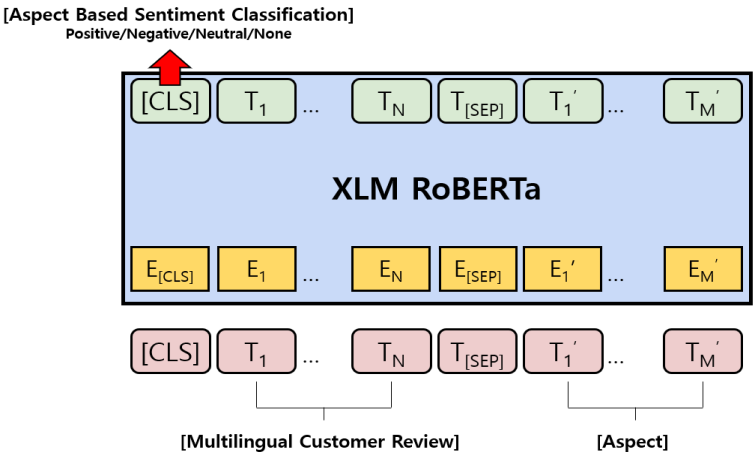

Figure 2: *Structure of aspect-based sentiment analysis using sentence-pair classification.*

아키텍처를 의미한다. 이 아키텍처를 활용하여 Figure 2에서 볼 수 있듯이 모델 입력값으로 사용자 후기에 해당되는 문장 A와 속성에 해당되는 B를 함께 넣어주고 두 개의 문장 사이의 관계를 감성분석으로 분류한다. 즉, 앞부분에는 사용자 후기에 해당되는 문장을, 뒷부분에는 분석하고자 하는 속성을 넣어주고 두 개의 문장 간의 관계를 '긍정(positive)', '부정(negative)', '중립(neutral)', '해당 없음(none)'으로 분류한다.

Figure 2에서 '해당 없음' 레이블을 추가하여 문장-쌍 분류 방식을 활용하는 이유는 바로 위 아키텍처를 활용하게 되면 속성 추출 모델을 따로 사용하지 않아도 된다는 장점이 있기 때문이다. 예를 들어 "The pizza was delicious but price was expensive"라는 문장이 존재하고 분석하고자 하는 대상 속성이 3가지인(가격, 맛, 분위기) 경우 모델의 입력으로는 [문장 – 속성 1(가격)], [문장 – 속성 2(맛)], [문장 – 속성 3(분위기)]이 들어가게 되고 이에 대한 출력은 각각 부정, 긍정, 해당 없음이 나오게 된다. 즉, 모든 속성에 대해서 다 각각 예측을 하기 때문에 별도로 속성 추출을 할 필요가 없게 되는 것이다. 반면, 기존 방법론 중에 속성을 먼저 추출하고 추출된 해당 속성에 대해서만 감성분석을 진행하는 방법도 있다. 하지만 모델을 두 번 사용해야 하기 때문에 성능 저하와 모델 관리의 어려움이라는 단점이 발생하게 된다. 따라서 본 논문에서는 위의 아키텍처를 활용하여 End-to-End ABSA 과제를 수행하였다.

3.2. 다국어 사전훈련 모델

분석해야 할 대상의 언어가 하나가 아닌 경우에는 언어별로 모델을 따로 구축하여 사용할 수 있지만 이는 모델 관리 측면에서 비효율적이다. 그렇기 때문에 본 논문에서는 하나의 모델로 여러 언어를 다룰 수 있도록 다국어 모델을 활용하였다. Pires 등 (2019)에서 제안된 다국어 모델인 Multilingual BERT는 기존 BERT를 학습시켰던 방식을 그대로 사용하되 학습된 코퍼스(corpus)가 영어가 아니라 다국어라는 점에서 차이가 있다. 위키피디아(wikipedia)에 있는 코퍼스를 활용하여 사전훈련(pre-training)을 진행하였고 104개 언어가 학습되었다. 코퍼스 간에 불균형이 있어서 성능이 언어별로 동일한 것은 아니지만 다양한 언어를 다룰 수 있다는 점에서 장점이 있다.

Devlin 등 (2019)과 Pires 등 (2019)을 참고하여 그린 Figure 3에서 볼 수 있듯이 사전훈련시 다국어 데이터를 활용하는 한편 토큰 정보가 들어갈 때 해당 토큰이 어떤 언어에 해당된다는 정보는 넣어주지 않는다. 또한 다국어 모델이기 때문에 다양한 텍스트들이 나타나는 공유 사전(shared vocabulary)을 활용하고 있다. 나머지 사전훈련 방식은 앞서 설명한 BERT와 동일하다.

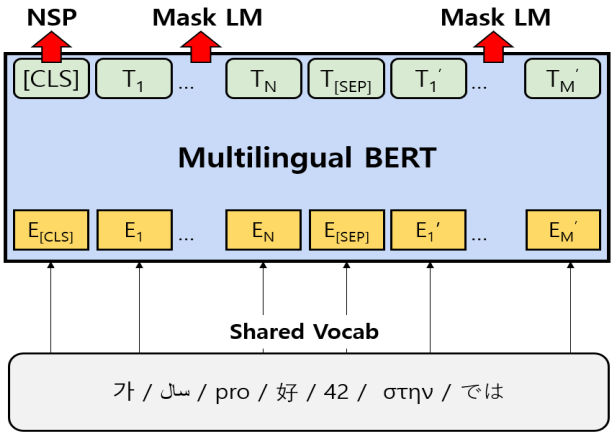

Figure 3: *Multilingual BERT pre-training architecture.*

Table 1: Restaurant 2016 data summary by language

	Total	English	Spanish	French	German	Russian	Turkish
Sentence (train/test)	10943/3344	2000/587	1625/677	1525/583	1722/451	2839/902	1232/144
Aspect per sentence	1.23	1.16	1.47	1.43	1.06	1.18	1.18
Length (mean/std)	76/59	67/43	83/83	83/64	73/50	80/56	61/41

다국어 모델은 단일어 모델에 비해 사전의 용량이 커서 계산하는데 시간이 다소 오래 걸린다는 단점이 있지만 하나의 모델로 다국어 데이터를 다룰 수 있다는 장점이 존재한다. 한편 Multilingual BERT 외에도 Conneau 등 (2020)에서 제안된 XLM-RoBERTa와 같은 다국어 모델이 있으며 실험을 통해 다국어 모델 간 성능 비교를 추가로 진행하였다.

4. 실제자료 분석

본 장에서는 제안된 방법의 성능을 확인하기 위한 실험에 사용된 데이터인 “Restaurant 2016”과 “화장품 도메인 다국어 사용자 후기”에 대한 설명과 모형 평가지표 소개 및 분석 결과를 제시한다.

4.1. 데이터 정보

AI 및 NLP 관련 대표적인 국제 워크숍 중 하나인 International Workshop on Semantic Evaluation 2016 (SemEval, 2016)에서 “Restaurant 2016” 데이터가 공개되었다. 참고문헌의 인터넷 주소에서 다운로드 받을 수 있는 해당 데이터는 SemEval-2016의 Task 5에서 공개된 데이터로 특정 상품이나 서비스에 대한 사용자 후기에 대해 속성기반 감성 분석을 진행하는 과제이다. 공개된 데이터의 도메인은 노트북, 식당, 호텔이 있으며 이 중 다국어가 많이 존재하는 도메인은 식당 도메인이었다. 따라서 식당 도메인의 Restaurant 2016 데이터로 실험을 진행하였다.

Restaurant 2016의 언어별 학습/훈련 데이터 건수와 문장별 속성의 수는 Table 1과 같다. 문장 당 몇 개의 속성이 나타났는지를 살펴보면 평균 약 1.23개로 생각보다 하나의 후기에 많은 속성이 담겨 있지 않다는 것을 알 수 있다. 그리고, 문장 길이를 살펴보는 이유는 이후 실험에서 최대 입력(input) 길이를 설정해 주어야 하기 때문이다. 데이터에는 ‘긍정’, ‘부정’, ‘중립’, ‘충돌’ 레이블이 있었으나 ‘충돌’ 레이블은 문장 개수가 적어

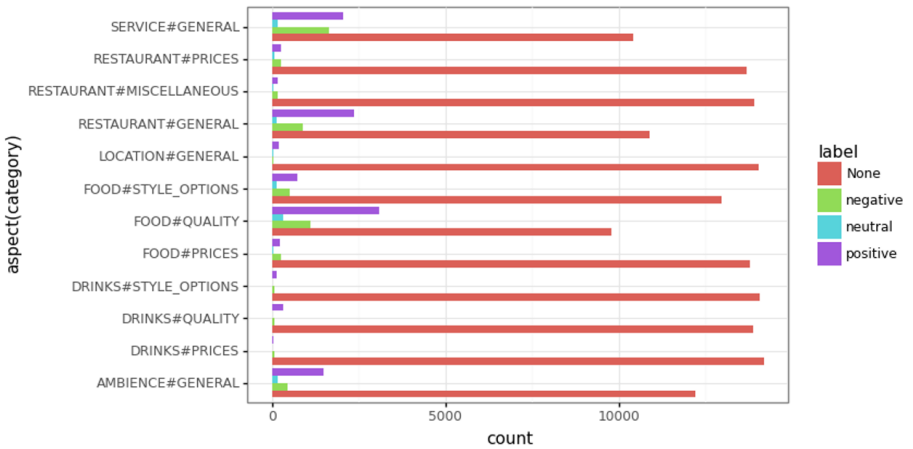

Figure 4: *Label distribution by aspect of Restaurant 2016 data.*

Table 2: Example of restrant 2016 data after preprocessing

Example	Aspect(category)	Label
The food was well prepared and the service impecable.	FOOD#QUALITY	positive
The food was well prepared and the service impecable.	SERVICE#GENERAL	positive
The food was well prepared and the service impecable.	RESTAURANT#QUALITY	none
⋮	⋮	⋮

Table 3: Cosmetic domain multilingual user review data summary

	Total	English	Korean	Japanese
Sentence (train/test)	39565/9893	12580/3145	13337/3335	13648/3413
Aspect per sentence	1.57	1.74	1.4	1.56
Length (mean/std)	74/59	111/61	51/44	61/49

제외하였다. 속성별 레이블 분포는 Figure 4와 같으며 '해당 없음', '긍정', '부정', '중립'의 순서로 레이블이 많음을 알 수 있다.

Figure 4에서 속성의 경우 총 12가지가 존재하며 속성별 문장 개수는 불균형한 분포를 보인다. 그리고 End-to-End ABSA 과제 수행을 위해 해당 데이터의 전처리가 Table 2와 같이 사람이 직접 마킹하여 진행되었으며 데이터를 보면 aspect와 label이 주어져 있다. 문장 하나에 속성별로 감성분석을 진행해야 하기 때문에 기존 문장에서 12배로 늘어나는 모습을 확인할 수 있다.

한편, "화장품 도메인 사용자 후기" 데이터의 경우 화장품 사용 후기에 대한 데이터로 한국어, 영어, 일본어 데이터를 각각 약 15,000건 씩 크롤링을 통해 직접 수집하였고 대상 속성은 12가지(고정력, 기타 효과, 두피, 모발, 배송, 사용감(제형), 세정력, 탈모 완화, 포장, 피부, 향, 향 지속력)를 선정하였다. 한국어의 경우 "Xation"이라는 화장품 업체로부터 자체 보유 데이터를 받았으며 영어의 경우 "Shopee" 홈페이지에서 화장품 카테고리에 있는 상품 후기들을 크롤링하였고, 일본어의 경우는 "Qoo10" 홈페이지에서 화장품 카테고리에 있는 후기들을 크롤링하였다. 데이터 레이블링은 해당 원어를 사용할 수 있는 사람이 주석을 다는 작업(annotation)을 진행했으며 다국어 모델을 활용했기 때문에 굳이 원어를 한국어로 번역하는 작업은 필요하지

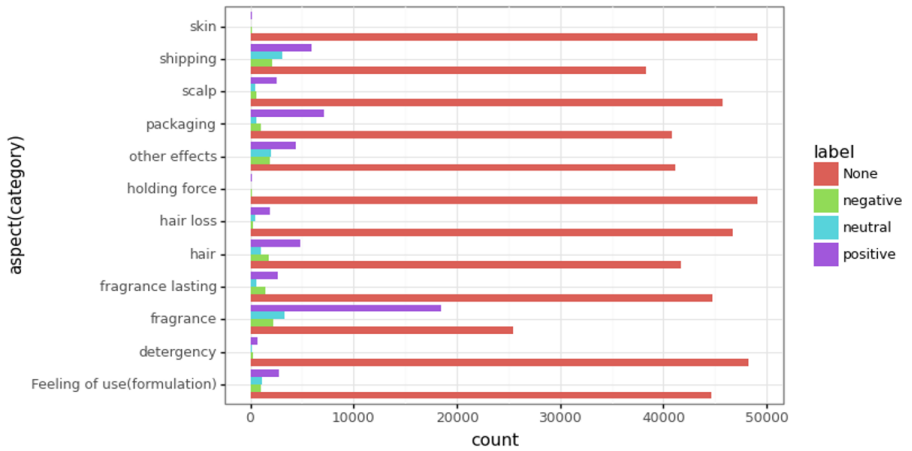

Figure 5: *Label distribution by aspect of cosmetic domain multilingual user review data.*

Table 4: Example of cosmetic domain multilingual user review data after preprocessing

Example	Aspect(category)	Label
향은 좋은데 향이 오래가지는 않아요.	Fragrance	Positive
향은 좋은데 향이 오래가지는 않아요.	Persistence of fragrance	Negative
향은 좋은데 향이 오래가지는 않아요.	Hair	None
⋮	⋮	⋮

않았다. Table 3에 언어별로 사용된 학습/훈련 데이터 건수와 문장별 속성의 수 등을 정리하였고, 속성별 레이블 분포는 Figure 5와 같다. 앞에서와 마찬가지로 End-to-End ABSA 과제 수행을 위해서 데이터 전처리를 진행하였으며 데이터 예시는 Table 4와 같다.

4.2. 모형 평가지표 및 실험 진행

모형 평가를 위한 지표로는 정확도(accuracy; ACC), 정밀도(precision), 재현율(recall), F1-score와 Macro F1-score 등이 있다. 정확도는 전체 데이터 중 모델이 맞게 예측한 비율이고, 정밀도는 모델 예측값 기준으로 관측값이 맞는 비율, 재현율은 관측값 기준으로 모델이 맞게 예측한 비율이며, F1-score는 정밀도와 재현율의 조화평균이며, Macro F1-score는 모든 클래스 라벨들의 F1-score의 평균을 의미한다. 이들 평가지표와 관련된 계산식 사례는 Figure 6에 제시하였고, 이 지표들 중에 본 논문에서는 ACC와 Macro F1-score를 사용하였다. ACC의 경우 가장 직관적으로 모델의 성능을 파악할 수 있는 지표이기에 활용하였고 한편 과제 특성상 '해당 없음' 레이블이 많이 나타나는 것을 고려하여 F1-score 중 모든 레이블의 F1-score 평균을 나타내는 Macro F1-score도 추가로 선정하였다.

실험은 크게 세 부분으로 나뉜다. 먼저 Restaurant 2016 데이터를 활용하여 다국어 모델 성능 비교를 위한 실험을 진행하였다. 이때 사용된 다국어 모델은 공개된 모델인 Multilingual Bert와 XLM-RoBERTa이다. 이 중 최고의 성능을 보인 모델을 모델인 XLM-RoBERTa를 기본 모델로 정하고 이후 실험들을 진행하였다. 추가로 다국어 모델이 단일어 모델에 비해 성능이 떨어지는지 확인을 위한 분석도 진행하였다.

두 번째 실험에서는 속성기반 감성분석 진행 시 성능 향상을 위한 여러 가지 방법론을 비교하였다. Sun 등 (2019)에 의하면 기본적으로 End-to-End ABSA에서 속성을 나타내는 문장 B의 형태에 따라 주어진 질문에

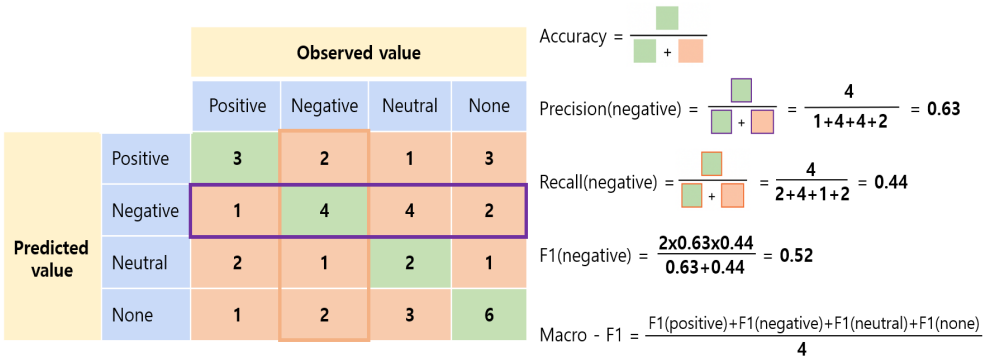

Figure 6: *Calculation examples of model evaluation indicators.*

Table 5: Comparison of multilingual model performance

Multilingual model	Sentence format	Learning method	Performance	
			ACC	Macro F1
Multilingual BERT	QA-M	CLS token	0.9380	0.6340
XLM-RoBERTa	QA-M	CLS token	0.9488	0.7001

Table 6: Performance comparison of multilingual and monolingual models based on English data

Model	Sentence format	Learning method	Performance	
			ACC	Macro F1
BERT	QA-M	CLS token	0.9536	0.6892
RoBERTa	QA-M	CLS token	0.9540	0.7177
XLM-RoBERTa	QA-M	CLS token	0.9561	0.7411

대한 해답을 찾는 QA (question answering)나 두 문장의 관계를 추론하는 NLI (natural language inference) 방식으로 나눠진다. 구체적인 예시를 들자면 QA 방식의 경우 문장 B를 “향 속성은 이 문장에서 어떤 감성으로 나타납니까?”와 같이 문장 형식으로 구성하는 것을 의미하고 NLI 방식의 경우 ‘향’과 같이 단어 형식으로 구성하는 것을 의미한다. 그래서 두 번째 실험은 문장 B 구성 방식에 따른 성능 비교를 진행하였다. 더 나아가 QA 방식도 세부적으로 보면 QA-M 방식과 QA-B 방식으로 나눠질 수 있다. 여기서 M과 B는 각각 정답의 형태가 다중분류와 이진분류라는 것을 의미하며 예를들면, QA-M 방식의 경우 문장 B는 ”향 속성은 이 문장에서 어떤 감성으로 나타납니까?” 그리고 대답은 ‘긍정’, ‘부정’, ‘중립’ 등에서 하나로 나타나는 다중분류 방식을 의미한다. 반면 QA-B 방식의 경우 구상은 “이 문장에서 향 속성은 긍정입니다”, 대답은 ‘맞음’, ‘틀림’과 같이 둘 중 하나로 나타나는 이진분류 방식을 의미한다. 본 논문에서는 이전 Sun 등 (2019)의 결과를 참고하여 성능이 높았던 NLI-M과 QA-M 방식을 적용하였다. 한편, BERT 모델 방식 중 기본이 되는 CLS 토큰(special classification token)을 통한 분류를 진행하였는데 이때 CLS 토큰은 분류 문제와 같은 과제의 해결을 위해서 문장 맨 앞에 붙는 토큰이며 대부분의 분류 문제에서 이와 같은 방식으로 모델링을 진행한다. 이외에도 모델 아키텍처를 일부 바꾸는 실험과 손실함수(loss function)를 바꿔 결과의 차이를 확인하는 실험을 추가로 진행하여 성능 향상을 위한 방법론들을 비교하였다.

마지막 실험에서는 화장품 도메인 다국어 사용자 후기 데이터를 활용하여 데이터 셋 크기에 따른 성능 변화를 파악하였다. 언어별로 약 6,000건의 데이터를 확보하면 충분한 성능을 얻을 수 있고 속성별 성능 차이는

Table 7: Additional experiments to improve the performance of aspect-based sentiment analysis

Multilingual model	Sentence format	Learning method	Performance	
			ACC	Macro F1
XLM-RoBERTa	NLI-M	CLS token + CE loss	0.9497	0.6973
	QA-M	CLS token + CE loss	0.9488	0.7001
	QA-M	CLS token + aspect token + CE loss	0.9505	0.6947
		CLS token + weighted loss	0.9412	0.6925
		CLS token + focal loss	0.9498	0.6954
		CLS token + aspect token + weighted loss	0.9487	0.6889
		CLS token + aspect token + focal loss	0.9494	0.7065

해당 속성의 자료 수에 따라 큰 차이가 있음을 확인하였다.

4.3. 성능 비교

첫 번째 실험의 경우 문장-쌍 분류 과제를 활용하여 속성기반 감성분석을 진행 시 다국어 모델별로 비교하여 추후 실험에서 활용될 기본(base) 모델을 선택하고자 하였다. 비교에 사용된 다국어 모델은 Multilingual BERT와 XLM-RoBERTa이며 추가로 다국어 모델이 단일어 모델에 비해 성능이 떨어지지 않고 오히려 더 나음을 확인하기 위해 단일어 모델과의 성능 비교도 진행하였다.

실험에서 Restaurant 2016 데이터의 모든 언어를 활용하였고 Table 5를 보면 XLM-RoBERTa 모델이 Multilingual BERT에 비해서 ACC 기준으로는 약 1%, Macro F1 기준으로는 약 6% 이상 성능이 높은 것을 확인할 수 있었다. 이를 토대로 살펴보면 다국어 모델은 XLM-RoBERTa를 사용하는 것이 성능이 더 우수함을 알 수 있다. XLM-RoBERTa 모델이 사전학습 단계에서 Multilingual BERT보다 더 많은 코퍼스로 학습하고 사전 크기도 더 크기 때문에 연산속도는 오래 걸리지만 성능은 더 높은 것으로 보인다.

XLM-RoBERTa의 경우 6개 언어 데이터를 모두 활용하여 학습하였고 단일어 모델은 영어 데이터만 활용하여 학습을 진행하였다. Table 6의 결과를 보면 다국어 모델인 XLM-RoBERTa와 단일어 모델인 BERT, RoBERT를 비교하였을 때 다국어 모델의 성능이 단일어 모델에 비해 오히려 더 우수한 것을 볼 수 있다. 물론, 학습 데이터 수의 차이는 있지만 다국어 데이터가 있는 경우에는 언어별로 모델을 별도로 구축하는 것보다 하나의 다국어 모델을 사용하는 것이 성능이 제일 우수하다는 것을 알 수 있다.

두 번째 실험의 목적은 속성기반 감성분석 수행 시 성능 향상을 위한 여러 가지 방법론을 비교하는 것이다. 먼저, 문장-쌍 분류에서 두 번째인 문장 B의 형태에 따라 QA, NLI 방식으로 나눠지는데 이에 대한 비교 실험을 진행하였다. 앞에서 서술한대로 QA 방식과 NLI 방식의 차이점은 문장 B를 입력으로 사용 시 QA는 질문의 형식으로 넣어주고 NLI는 별다른 변형 없이 속성 자체만 활용한다는 점에서 차이점이 있다. 즉, 속성이 “FOOD#QUALITY” 일 때 QA 방식은 “What do you think of the FOOD#QUALITY of it”과 같이 문장으로 넣어주고 NLI 방식은 “FOOD#QUALITY”만 문장 B에 넣어준다. 이외에도 4.2절에서 언급한대로 추가로 분류 방법에 따라 QA 방식은 QA-M, QA-B 등으로 세분화될 수 있는데 QA-M 방식의 경우 정답 데이터가 ‘긍정’, ‘부정’, ‘중립’, ‘해당없음’ 중 하나의 다중분류로 구성되도록 하는 반면 QA-B 방식은 정답 데이터가 ‘맞다/틀리다’의 이진분류로 구성되도록 하는 경우를 말한다.

기존 BERT와 같은 트랜스포머 인코더 계열을 활용한 분류 과제는 CLS 토큰만 활용하여 분류를 진행하는데 Park과 Shin (2020)처럼 CLS 토큰 뿐만 아니라 문장 B에 등장하는 특정 속성도 분류에 활용하는 경우 실제 성능 향상에 도움이 되는지 추가로 확인하였다. 그리고, 손실함수(loss function)의 경우 일반적으로 CrossEntropy loss를 활용하지만 레이블이 불균형한 데이터가 있을 경우 이를 보완하기 위한 weighted loss, focal loss 등을 고려해 볼 수 있다.

Table 8: Detailed performance of the best model based on Restaurant 2016 data

	Precision	Recall	F1-score	support
None	0.9772	0.9767	0.9770	35783
Negative	0.6856	0.6275	0.6552	1369
Neutral	0.4640	0.3692	0.4112	279
Positive	0.7570	0.8098	0.7825	2697
ACC			0.9494	
Macro F1			0.7065	

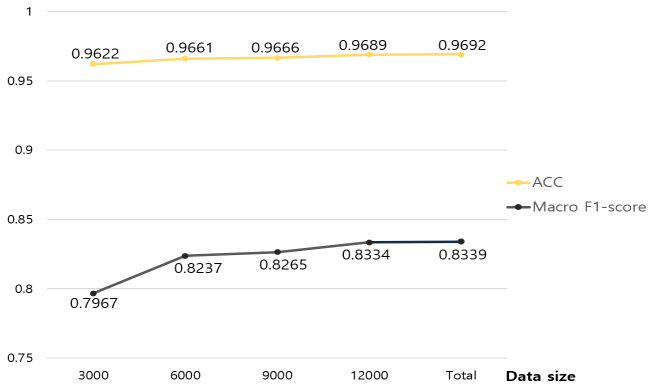

Figure 7: *Performance comparison by data size.*

Weighted loss의 경우 기존 CrossEntropy에 가중치(weight)를 추가로 곱해주는데 가중치를 부여할 때 표본 비율의 역수만큼 곱해주는 방식이다. 이렇게 계산하면 충분한 수의 클래스 데이터는 loss가 작게 반영되고 부족한 클래스의 데이터에는 loss가 상대적으로 크게 반영되어 데이터 개수가 적은 레이블을 더 잘 학습할 수 있게 된다. 한편, focal loss의 경우 분류 성능이 높은 클래스에 대해서는 하향 가중치(down weighting)을 부여하는 방식으로 마찬가지로 불균형 데이터의 경우 활용되는 loss이다.

Table 7의 추가실험 결과에서 QA 방식과 NLI 방식의 ACC를 보면 NLI 방식의 성능이 조금 높지만 데이터 특성상 '해당 없음'이 많이 나타나기 때문에 레이블 불균형을 고려한 성능 평가가 더 중요하다. 그렇기 때문에 Macro F1-score가 더 높은 QA 방식을 활용하여 나머지 실험을 진행하였다. 전체 성능 비교표를 보면 속성 토큰(aspect token) 임베딩 값을 함께 활용해 주면서 손실함수는 focal loss를 활용하는 방식이 가장 성능이 높은 것을 알 수 있다.

한편, 앞에서 제시한 가장 성능이 좋은 최적 모델에 대한 세부 성능을 비교한 결과는 Table 8과 같다. 클래스 자료의 개수가 적은 중립과 부정의 경우 성능이 상대적으로 떨어지는 것을 확인할 수 있으며 ACC가 높은 것은 '해당 없음' 레이블의 자료 개수가 많아서 그런 것으로 생각된다. 추가 성능 향상을 위해서는 자료의 개수가 적은 클래스 데이터에 대해 추가 학습 데이터가 있어야 할 것으로 판단된다.

마지막 실험에서는 좀 더 많은 데이터를 갖고 데이터 개수에 따른 성능 변화와 단일어 모델과의 성능 비교를 진행하였다. 이를 위해 데이터 개수가 많은 "화장품 도메인 다국어 사용자 후기" 데이터를 활용하였다. 해당 데이터는 영어, 일본어 및 한국어 3개 언어와 12개의 속성에 대해 레이블링된 데이터이다. 평가 데이터는 같게 하고 학습 데이터만 3,000건, 6,000건, 9,000건, 12,000건, 전체 데이터로 바꿔가면 학습을 진행했다. 이 때, 예를 들어 6,000건은 기존 3,000건에 3,000건을 새롭게 추가하는 방식과 같이 계속해서 누적되게끔

Table 9: Language-specific performance of the best model using cosmetics domain multilingual user review data

	English				Japanese				Korean			
	Precision	Recall	F1-Score	support	Precision	Recall	F1-Score	support	Precision	Recall	F1-Score	support
None	0.9867	0.9895	0.9881	33726	0.9870	0.9867	0.9869	35458	0.9904	0.9920	0.9912	32879
Negative	0.7890	0.7377	0.7625	446	0.7592	0.7837	0.7713	1082	0.7746	0.7942	0.7843	831
Neutral	0.7197	0.7434	0.7314	943	0.6640	0.5641	0.6100	897	0.7513	0.5885	0.6600	729
Positive	0.9201	0.9013	0.9106	4905	0.8608	0.8878	0.8741	3519	0.9072	0.9303	0.9186	3301
ACC			0.9701				0.9636				0.9744	
Macro F1			0.8481				0.8106				0.8385	

Table 10: Aspect-specific performance of the best model using cosmetics domain multilingual user review data

	Fragrance				Persistence of fragrance				Delivery			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
None	0.9430	0.9553	0.9491	3844	0.9889	0.9912	0.9901	85618	0.9863	0.9887	0.9875	8149
Negative	0.7822	0.7341	0.7574	455	0.8341	0.8638	0.8487	448	0.8357	0.8417	0.8387	278
Neutral	0.6757	0.5829	0.6259	772	0.6729	0.4737	0.5560	152	0.7709	0.8194	0.7944	382
Positive	0.9295	0.9457	0.9375	4822	0.8853	0.8962	0.8907	732	0.9297	0.8902	0.9095	1084
ACC			0.9114				0.9705				0.9672	
Macro F1			0.8175				0.8214				0.8825	
	Packaging				Sense of use				Other effects			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
None	0.9878	0.9926	0.9902	8374	0.9872	0.9841	0.9857	8704	0.9612	0.9639	0.9626	7988
Negative	0.7710	0.7710	0.7710	131	0.6303	0.7376	0.6797	141	0.7260	0.7139	0.7199	423
Neutral	0.7143	0.7778	0.7447	90	0.6996	0.6346	0.6655	312	0.7066	0.6156	0.658	450
Positive	0.9592	0.9230	0.9407	1298	0.8073	0.8424	0.8245	736	0.7793	0.8110	0.7949	1032
ACC			0.9786				0.9591				0.9215	
Macro F1			0.8616				0.7889				0.7838	
	Holding force				Scalp				Hair			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
None	0.9997	0.9998	0.9997	9874	0.9967	0.9954	0.9960	9484	0.9839	0.9818	0.9828	8194
Negative	0.8571	0.8571	0.8571	7	0.6471	0.6875	0.6667	48	0.7686	0.8095	0.7885	357
Neutral	0	0	0	1	0.6341	0.5532	0.5909	47	0.7053	0.5678	0.6291	236
Positive	0.9000	0.8182	0.8571	11	0.8667	0.9108	0.8882	314	0.8678	0.9024	0.8848	1106
ACC			0.9994				0.9891				0.9568	
Macro F1			0.6785				0.7855				0.8213	
	Hair loss relief				Cleaning power				Skin			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
None	0.9977	0.9978	0.9977	9463	0.9927	0.9969	0.9948	9612	0.9983	0.9985	0.9984	9816
Negative	0.7333	0.7097	0.7213	31	0.7143	0.6522	0.6818	23	0.5556	0.5882	0.5714	17
Neutral	0.8030	0.8154	0.8092	65	0.7778	0.7368	0.7568	57	0.3333	0.2000	0.2500	5
Positive	0.9189	0.9162	0.9175	334	0.8253	0.6816	0.7466	201	0.8148	0.8000	0.8073	55
ACC			0.9929				0.9882				0.9963	
Macro F1			0.8614				0.7950				0.6568	

구성하였다.

Figure 7의 결과를 살펴보면 '해당 없음'이 많은 불균형 데이터 특성상 데이터 수가 적더라도 ACC는 높게

나오고 있다. 그렇기 때문에 데이터 불균형을 고려한 성능 지표인 Macro F1-score를 참고해야 한다. 언어별로 데이터가 최소 6,000건은 있어야 Macro F1-score도 9,000건, 12,000건, 전체(15,000건) 대비 성능 차이가 크지 않음을 알 수 있다. 실제 모델 구축 시 데이터가 많으면 많을수록 좋겠지만 데이터를 구축하는 것 역시 많은 비용과 시간이 필요하게 된다. 이 때, 위 실험의 결과를 참고하면 대략적으로 일정 성능을 내기 위해 필요한 데이터 양을 파악할 수 있을 것으로 판단된다. 한편, 최고 성능 모델의 성능을 언어별로 살펴본 Table 9의 결과를 보면 Macro F1-score 기준으로 모든 언어에서 80% 이상의 성능을 나타내고 있음을 알 수 있으며 언어별 성능 차이가 크지 않음을 확인할 수 있다.

두 번째로 속성별 세부 성능은 Table 10에 제시하였다. 사용자 후기에서 많이 나타나는 속성인 향(fragrance), 향 지속력(persistence of fragrance), 배송(delivery), 포장(packaging), 사용감(sense of use) 등의 경우 자료의 개수가 충분하여 성능이 잘 나오는 반면 그렇지 않은 속성인 두피(scalp), 모발(hair), 고정력(holding force), 피부(skin) 등의 경우 성능이 떨어지는 것을 확인할 수 있다. 아무래도 사용자 후기에 많이 나타나지 않는 속성의 경우 학습 데이터에도 잘 나타나지 않기 때문에 학습이 잘되지 않는 것으로 추측할 수 있다. 또한 부정과 중립에 해당되는 표현이 데이터에 많이 나타나지 않기 때문에 성능 역시 긍정에 비해서 다소 떨어짐을 확인할 수 있다.

5. 결론 및 시사점

본 논문에서는 다국어 데이터를 활용하여 속성기반 감성분석을 진행하였다. 단순 감성분석이 아닌 속성기반 감성분석의 결과는 다음과 같이 효율적으로 활용할 수 있다. 기간별로 분석 결과를 집계하여 해당 기간 동안 어느 속성에 대한 평가가 어떻게 달라졌는지 비교할 수 있으며 더 나아가 타제품과의 비교에도 활용할 수 있다. 특히, 속성별로 소비자에 대한 반응을 집계할 수 있어서 소비자들의 니즈를 파악하는 데 도움이 될 수 있다. 또한, 결국 분류 과제이기 때문에 감성분석뿐만 아니라 다른 분류 과제에서도 활용할 수 있어 적용 범위가 넓다는 장점이 있으며, 본 논문의 결과를 활용하여 다양한 비즈니스 인사이트 도출이 가능할 것으로 기대된다. 한편, 기존 영어를 위주로 진행되던 연구에서 더 나아가 본 논문에서는 다국어 데이터를 활용하여 실험을 진행하였다. 특히 해외로 상품이나 서비스를 판매하는 경우 사용자 후기가 한 개의 언어가 아니라 다양한 언어로 나타날 수 있기 때문에 다국어 모델을 활용하여 분석하는 것이 절대적으로 필요할 것으로 생각된다.

본 논문은 기존 감성분석에서 더 나아가 다면적 감성분석 방법론을 시도해볼 수 있다는 점에서 의미가 있고 더욱이 본 논문에서 활용된 문장-쌍 분류 방식은 감성분석뿐만 아니라 다른 분류 과제에서도 적용할 수 있다는 점에서 의의가 크다. 또한 기존 한 가지 언어로만 진행된 선행 연구들과 달리 실제 다국어 데이터를 통해서 다국어 모델링이 가능하다는 점을 살펴봤다는 점에서 활용가치가 높다고 판단된다.

다만, BERT와 같은 언어 모델은 모델이 워낙 크기 때문에 실시간 예측이 필요한 경우 실제 비즈니스 상에서는 활용하기 까다롭고 GPU가 필요하기 때문에 운영하기 어려울 수 있다는 단점이 있다. 그래서 이를 위해 다양한 경량화 모델 연구가 진행되고 있다. 앞에서 실험한 다국어 데이터에 대한 속성기반 감성분석 모델도 경량화 방법을 적용한다면 모델 예측 시간을 더욱 줄일 수 있을 것으로 기대된다. 더 나아가 기존 학습된 모델에서 추가로 분류 레이블이 더 필요한 경우 학습 데이터를 다시 구축하고 재학습해야 하는 문제가 존재한다. 이와 관련하여 전체 학습 데이터를 재구축하지 않고 효율적으로 모델을 학습할 수 있는 방법론에 대한 연구가 필요할 것으로 보인다.

감사의 글

본 연구가 성공적으로 수행될 수 있도록 다각도로 지원해 주신 정보통신산업진흥원(NIPA)에 감사드립니다.

References

Brown T, Mann B, Ryder N *et al*. (2020). Language models are few-shot learners, *Advances in Neural Information Processing Systems*, **33**, 1–25, Available from: arXiv:2005.14165.

Conneau A, Khandelwal K, Goyal N *et al*. (2020). Unsupervised cross-lingual representation learning at scale, In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, 8440–8451, Available from: arXiv:1911.02116.

Devlin J, Chang MW, Lee L, and Toutanova K (2019). 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*, **1**, 4171–4186.

Park HJ and Shin KS (2020). Aspect-based sentiment analysis using BERT: Developing aspect category sentiment classification models, *Journal of Intelligence and Information Systems*, **26**, 1–25.

Pires T, Schlinger E, and Garrette D (2019). How multilingual is multilingual BERT?, In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 4996–5001, Available from: arXiv:1906.01502.

SemEval (2016). Restaurant 2016 data, Available from: https://alt.qcri.org/semeval2016/task5/index.php?id=data-and-tools

Sun C, Huang L, and Qiu X (2019). Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **1**, 380–385, Available from: arXiv:1903.09588.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, and Polosukhin I (2017). Attention is all you need, *Advances in Neural Information Processing Systems*, **30**, 5998–6008.

Zhang W, Li X, Deng Y, Bing L, and Lam W (2022). A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges, Available from: arXiv:2203.01054.

Received April 2, 2023; Revised May 7, 2023; Accepted May 30, 2023

다국어 사용자 후기에 대한 속성기반 감성분석 연구

지성영[a], 이시윤[a], 최대우[a], 강기훈[1,a]

[a]한국외국어대학교 통계학과

요 약

전자상거래 시장의 성장과 더불어 소비자들은 상품 및 서비스 구매 시 다른 사용자가 작성한 후기 정보에 기반하여 구매 의사를 결정하게 되며 이러한 후기를 효과적으로 분석하기 위한 연구가 활발히 이루어지고 있다. 특히, 사용자 후기에 대해 단순 긍/부정으로 감성분석하는 것이 아니라 다면적으로 분석하는 속성기반 감성분석 방법이 주목받고 있다. 속성기반 감성분석을 위한 다양한 방법론 중 최신 자연어 처리 기술인 트랜스포머 계열 모델을 활용한 분석 방법이 있다. 본 논문에서는 최신 자연어 처리 기술 모델에 두 가지 실제 데이터를 활용하여 다국어 사용자 후기에 대한 속성기반 감성분석을 진행하였다. 공개된 데이터 셋인 SemEval 2016의 Restaurant 데이터와 실제 화장품 도메인에서 작성된 다국어 사용자 후기 데이터를 활용하여 속성기반 감성분석을 위한 트랜스포머 계열 모델의 성능을 비교하였고 성능 향상을 위한 다양한 방법론도 적용하였다. 다국어 데이터를 활용한 모델을 통해 언어별로 별도의 모델을 구축하지 않고 한가지 모델로 다국어를 분석할 수 있다는 점에서 효용 가치가 클 것으로 예상된다.

주요용어: BERT, 다국어 BERT, 자연어 처리, 트랜스포머 인코더, XLM-RoBERTa

이 연구는 2023학년도 한국외국어대학교 교원연구지원사업 지원에 의하여 이루어진 것임.

[1]교신저자: (17035) 경기도 용인시 처인구 외대로 81, 한국외국어대학교 통계학과. E-mail: khkang@hufs.ac.kr