JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 2, pp. 43-50, February 2024
https://doi.org/10.9708/jksci.2024.29.02.043

Zero-shot Korean Sentiment Analysis with Large Language Models: Comparison with Pre-trained Language Models

Soon-Chan Kwon*, Dong-Hee Lee*, Beak-Cheol Jang*

*Student, Graduate School of Information, Yonsei University, Seoul, Korea
*Student, Graduate School of Information, Yonsei University, Seoul, Korea
*Professor, Graduate School of Information, Yonsei University, Seoul, Korea

[Abstract]

This paper evaluates the Korean sentiment analysis performance of large language models like GPT-3.5 and GPT-4 using a zero-shot approach facilitated by the ChatGPT API, comparing them to pre-trained Korean models such as KoBERT. Through experiments utilizing various Korean sentiment analysis datasets in fields like movies, gaming, and shopping, the efficiency of these models is validated. The results reveal that the LMKor-ELECTRA model displayed the highest performance based on F1-score, while GPT-4 particularly achieved high accuracy and F1-scores in movie and shopping datasets. This indicates that large language models can perform effectively in Korean sentiment analysis without prior training on specific datasets, suggesting their potential in zero-shot learning. However, relatively lower performance in some datasets highlights the limitations of the zero-shot based methodology. This study explores the feasibility of using large language models for Korean sentiment analysis, providing significant implications for future research in this area.

▸ **Key words**: Language Model, Sentiment analysis, AI, Natural Language Processing, Deep Learning

[요 약]

본 논문은 GPT-3.5 및 GPT-4와 같은 대규모 언어 모델의 한국어 감성 분석 성능을 ChatGPT API를 활용한 zero-shot 방법으로 평가하고, 이를 KoBERT와 같은 사전 학습된 한국어 모델들과 비교한다. 실험을 통해 영화, 게임, 쇼핑 등 다양한 분야의 한국어 감성 분석 데이터셋을 사용하여 모델들의 효율성을 검증한다. 실험 결과, LMKor-ELECTRA 모델이 F1-score 기준으로 가장 높은 성능을 보여주었으며, GPT-4는 특히 영화 및 쇼핑 데이터셋에서 높은 정확도와 F1-score를 기록하였다. 이는 zero-shot 학습 방식의 대규모 언어 모델이 특정 데이터셋에 대한 사전 학습 없이도 한국어 감성 분석에서 높은 성능을 발휘할 수 있음을 시사한다. 그러나 일부 데이터셋에서의 상대적으로 낮은 성능은 zero-shot 기반 방법론의 한계점으로 지적될 수 있다. 본 연구는 대규모 언어 모델의 한국어 감성 분석 활용 가능성을 탐구하며, 이 분야의 향후 연구 방향에 중요한 시사점을 제공한다.

▸ **주제어**: 언어 모델, 감성 분석, 인공 지능, 자연어처리, 딥러닝

• First Author: Soon-Chan Kwon, Corresponding Author: Beak-Cheol Jang
*Soon-Chan Kwon (elrnjs13@yonsei.ac.kr), Graduate School of Information, Yonsei University
*Dong-Hee Lee (dlehjdgml1031@yonsei.ac.kr), Graduate School of Information, Yonsei University
*Beak-Cheol Jang (bjang@yonsei.ac.kr), Graduate School of Information, Yonsei University
• Received: 2024. 01. 16, Revised: 2024. 02. 05, Accepted: 2024. 02. 06.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

정보의 바다 속 트랜스포머(Transformer) 아키텍처의 등장은 자연어 처리(Natural Language Processing) 분야에서 혁명적인 변화를 가져왔다[1]. 이는 어텐션 메커니즘을 기반으로 하여 텍스트의 깊은 의미를 이해하는 데 큰 성공을 거두었다. 이후 BERT[2]와 같은 트랜스포머 기반의 모델들이 등장하면서, 사전 학습된 모델들이 다양한 자연어 처리 작업에서 뛰어난 성능을 보여주었다[3].

최근 인공 지능 분야의 급속한 발전은 ChatGPT[4]와 Bard[5]와 같은 대화형 AI 플랫폼의 등장으로 이어졌다. 이러한 플랫폼은 GPT-4[6] 및 PaLM 2[7]와 같은 대규모 언어 모델(Large Language Model)에 기반하고 있으며, 이 모델들은 광범위한 데이터 세트로 학습되어 감성 분석을 비롯한 다양한 자연어 처리 작업에서 뛰어난 성능을 보여주었다.

이러한 배경 속에서, 대규모 언어 모델(Large Language Models, LLM)에 대한 연구는 최근 몇 년간 눈에 띄게 증가하고 있다. 특히, 감성 분석과 같은 자연어 처리의 하위 태스크는 대규모 언어 모델의 능력을 평가하는 데 핵심적인 역할을 수행하고 있다. 자연어 처리의 핵심 태스크중 하나인 감성 분석은 제품 리뷰, 소셜 미디어 게시물, 뉴스 기사 등 특정 대상에 대한 사람들의 의견, 태도, 감정을 파악하는 과정이다[8].

다양한 언어의 감성 분석 문제를 해결하기 위해 규칙 기반, 사전 기반, 머신러닝 알고리즘 등 다양한 방법이 연구됐다. 대규모 언어 모델은 감성 분석 작업에서 뛰어난 성능을 보이며, 기존 방법들에 비해 상당한 발전을 이루었다.

그러나 기존 연구들은 주로 영어와 같은 대중적인 언어에 집중되어 있기에 한국어를 대상으로 대규모 언어 모델의 성능을 비교하는 연구는 수행된 적 없었다. 특히, 제로샷(Zero-shot) 기반의 한국어 감성 분석에 대한 효과는 광범위하게 연구되지 않았다. ChatGPT와 같은 대규모 언어 모델의 아랍어 감성 분석 능력을 평가한 연구는 있었지만[9], 한국어에 초점을 맞춘 연구는 아직 부족한 실정이다.

본 연구는 이러한 연구 공백을 메우기 위해 대규모 언어 모델 한국어 감성 분석 성능을 기존 지도 학습 모델과 비교 평가하여 효용성을 입증하고자 한다. GPT-3.5, GPT-4, Bard AI와 같은 생성 모델의 한국어 감성 분석 성능을 평가하고, 이를 KoBERT와 같은 미세 조정(Fine-tuning)된 기존의 지도 학습 모델과 비교함으로써 이러한 격차를 해소하고자 한다.

본 연구의 주요 목적은 이러한 대규모 언어 모델이 한국어 감성 분석에서 어떤 성능을 보이는지를 평가하고, 이를 통해 해당 분야에 새로운 통찰력을 제공하고자 한다. 구체적으로 한국어 감성 분석을 위한 대규모 언어 모델의 제로샷 학습 능력을 평가하고, 기존의 미세 조정된 모델과 성능을 비교하는 것이다.

본 논문은 다음과 같은 연구 질문을 제시한다.

√ 연구질문 1 : 대규모 언어 모델의 제로 샷 기반 한국어 감성 분석 성능은 어떠한가?

√ 연구질문 2 : 대규모 언어 모델의 제로 샷 기반 한국어 감성 분석 성능은 사전 학습된 언어 모델을 미세 조정한 방법과 성능이 얼마나 차이가 나는가?

본 논문에서는 주요 연구 질문들을 해결하기 위해 대표적인 세 가지 한국어 감성 분석 데이터셋을 사용하여 다양한 모델의 성능을 비교한다. 논문의 구성은 다음과 같다. 2장에서는 한국어 감성 분석에 대한 이전 연구와 대규모 언어 모델에 대한 연구를 소개한다. 3장에서는 다양한 모델과 실험에 사용한 데이터 및 평가지표 등 실험 설계에 대해 설명한다. 4장에서는 실험 및 분석 결과를 제시한다. 5장에서는 결론과 함께 논문을 매듭짓는다.

II. Related Works

1. Large Language Model

자연어 처리 분야에서 언어 모델은 여러 단계를 거쳐 발전했다. 초기의 정적 언어 모델(Static Language Models)은 고정된 단어 집합을 사용하여 단어 간 관계를 모델링했다. 이후, 신경 언어 모델(Neural Language Models)은 인공 신경망(Artificial Neural Network)을 사용해 언어의 순차적 특성을 포착하는 방식으로 발전했다[10].

사전 학습된 언어 모델(Pre-trained Language Models, PLMs)의 등장은 자연어 처리 분야에 혁신적인 변화를 가져왔다. BERT와 GPT[11]는 이러한 변화의 중심에 있다. BERT는 양방향 트랜스포머를 사용하여 텍스트의 깊은 언어적 이해를 가능하게 했으며, 특히 텍스트 분류, 개체명 인식 등 다양한 자연어 처리 작업에서 좋은 성능을 보여주었다. 한편, GPT는 생성적 언어 모델로서, 텍스트를 생성하고 이해하는 능력을 갖추었다. 이 모델들은 대규모 데이터셋으로 사전 학습된 후 특정 작업에 맞게 미세 조정하는 방식으로 활용되었다[12]. BERT, GPT 외에도 T5[13], BART[14], RoBERTa[15]와 같이 성능이 개선된 다양한 언어 모델이 등

장하였다. 또한, SciBERT[16], BioGPT[17]와 같은 다양한 도메인에 특화된 모델도 등장했다.

대규모 언어 모델의 시대로 접어들면서, 모델 크기와 성능은 기하급수적으로 증가했다. GPT-3와 GPT-4와 같은 모델들은 수십억 개의 파라미터를 가지고 있으며, 이를 통해 더욱 정교하고 다양한 언어 작업을 수행할 수 있게 되었다[18]. 이러한 대규모 언어 모델들은 감성 분석, 텍스트 생성, 질문 응답 등의 자연어 처리 작업에서 뛰어난 성과를 보여주었다.

대규모 언어 모델의 발전은 또한 생성형 언어 모델의 시대를 열었다. ChatGPT는 사용자의 질문을 이해하고 사람이 작성한 것과 같은 대답을 생성한다. 이를 통해 기본적인 작업과 복잡한 작업을 포함하는 다양한 과제를 수행할 수 있으며, 자연어 처리 및 인공 지능 분야에서 혁신적인 능력을 선보였다[19]. 이외에도 PaLM 2[7], LlaMa[20]와 같은 다양한 대규모 언어 모델이 등장했다. 이러한 모델들은 자연어 처리의 응용 범위를 확장하고, 언어에 대한 이해를 바탕으로 산업 전반에서 인공 지능의 활용 가능성을 증대시켰다.

이러한 발전의 흐름은 한국어를 포함한 다양한 언어에서의 감성 분석에 큰 영향을 미쳤다. 대규모 언어 모델을 활용한 최근의 연구는 이전의 사전 학습된 언어 모델 기반 접근 방식보다 더욱 정교하고 포괄적인 언어 이해를 가능하게 한다. 이러한 흐름 대로 본 연구에서도 한국어 기반 대규모 언어 모델을 활용하여 다양한 한국어 데이터셋에 대한 깊이있는 분석을 이어간다.

2. Korean Sentiment Analysis

사용자들이 자신의 의견을 자주 표현하는 소셜 미디어와 온라인 플랫폼의 증가로 인해 한국어 감성 분석에 대한 관심이 높아지고 있다. 한국어 감성 분석에 대한 기존의 접근 방식은 대부분 다른 언어의 방식을 모방했으며, 초기에는 규칙 기반 및 어휘 기반 방식에 중점을 두었다. 머신러닝 알고리즘이 발전하면서 한국어 특유의 형태적, 구문적 특징을 활용하는 나이브 베이즈(Naive Bayes), 서포트 벡터 머신(Support Vector Machine), 신경망(Nerual Network)[21]과 같은 기법이 도입되었다.

뉴스와 주가 사이의 연관성을 찾기 위해 뉴스 속 단어들을 감성 사전을 사용하여 태깅하는 통계적인 방법의 감성 분석 연구가 진행됐다[22]. 더불어, 변수 간 유사성을 표현하는 포인트별 상호 정보(Point-wise Mutual Information)값을 사용하여 영화 장르별 사용되는 단어에 따라 감성 사전을 구축한 연구가 진행됐다[23].

신경망을 사용한 다양한 감성 분석 연구가 진행됐다. 양향향 장단기 메모리(Bidirectional Long Short-Term Memory, Bi-LSTM)를 사용하여 표준 국어 대사전을 감성 분류하고 감성 어휘를 확장한 연구[24] 및 여러 형태소 벡터 도출 대안에 따른 감성 분석의 정확도를 합성곱 신경망(Convolutional Neural Network, CNN)을 사용하여 비교 검증한 연구가 진행됐다[25].

이외에도 감성 분석의 결과를 다른 연구의 변수로 활용한 연구가 진행됐으며[26], 추천 시스템(Recommendation System)의 성능을 향상시키기 위해 감성 분석 결과인 감성 수치를 입력 변수로 활용한 연구가 진행됐다[27].

특히, BERT(Bidirectional Encoder Representations from Transformers)와 그 변형 모델인 트랜스포머 기반 모델의 도입은 감성 분석에 큰 변화를 가져왔다. 대규모 한국어 말뭉치에 대해 사전 학습된 KoBERT[28], LMkor-BERT[29]와 같은 모델은 한국어 감성 분석 분야에서 뛰어난 성능을 보여주었다. 이 모델들은 특히 대규모 한국어 말뭉치로 학습되어 한국어에 대한 이해가 필요한 작업에서 기존 머신러닝 방법보다 우수한 성능을 보여주었다. 또한, 순환 신경망(Recurrent Neural Network) 기반 모델과 트랜스포머 기반 모델의 감성 분석 정확도를 비교 분석한 연구도 진행됐다. 본 연구에서는 이러한 한국어에 특화 트랜스포머 기반 모델의 사용과 더불어 다양한 한국어 감성분석 데이터셋을 활용하였다.

3. Sentiment Analysis with Large Language Models

감성 분석에 GPT-3, GPT-4, Bard AI와 같은 대규모 언어 모델을 적용하는 것은 최근 많은 연구의 주제로 자리잡고 있다. 이러한 연구는 영어 및 특정 언어를 포함한 다양한 언어에서 대규모 언어 모델의 감성 분석 성능을 탐구한다.

자연어 처리의 대표적인 작업의 다양한 데이터셋을 사용하여 ChatGPT의 성능을 평가한 연구[31], 감성 분석을 포함한 다양한 자연어 처리 작업에서 GPT-4의 제로 샷 기반 성능을 평가한 연구[32], ChatGPT의 제로 샷 감성 분석 성능을 미세 조정된 사전 학습된 언어 모델의 감성 분석 성능과 비교한 연구가 진행됐다[33]. 흥미로운 점은 기존 연구에서 다른 작업과 달리 감성 분석은 대규모 언어 모델이 미세 조정된 사전 학습된 언어 모델보다 같거나 오히려 더 높은 성능을 보여주고 있다는 것이다[34].

영어가 아닌 다른 언어를 대상으로 대규모 언어 모델의 감성 분석 성능을 비교한 연구가 진행됐다. 8가지 일반적

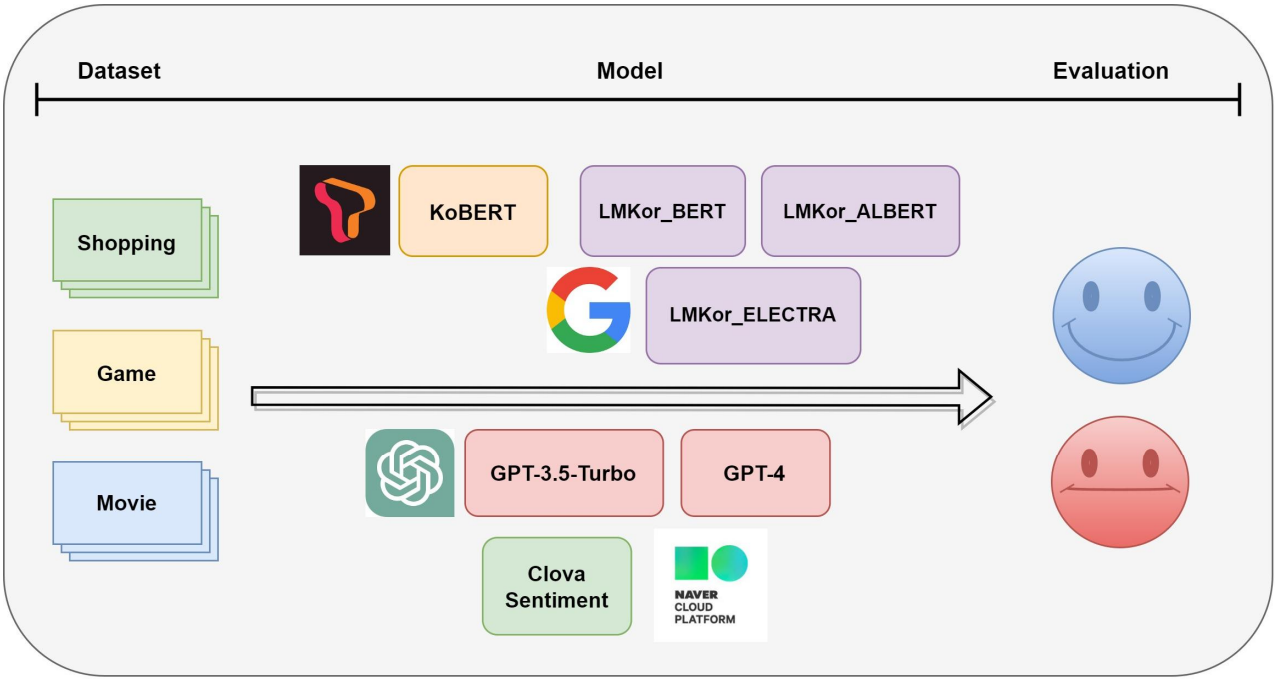

Fig. 1. Sentiment analysis comparison experiment framework

인 자연어 처리 응용 작업의 23개 데이터셋을 사용하여 ChatGPT의 다국어 성능을 평가한 연구[35], ChatGPT와 BARD의 아랍어 감성 분석 및 문장 생성 기능의 성능을 비교한 연구가 진행됐다[36]. 또한, 한국어 영화 리뷰 데이터를 사용하여 GPT-4의 한국어 감성 분석 능력을 다른 사전 학습된 모델의 감성 분석 성능과 비교한 연구가 진행됐다. 본 연구에서는 위와 같은 다양한 유사 연구들에게 영감을 받아 다양한 한국어 데이터셋에 대한 대규모 언어 모델의 성능을 비교 분석한다.

III. Experiment setting

본 연구에서는 한국어 감성 분석에서 대표적으로 사용되는 세 가지 데이터셋을 사용하였다[1)2)].

- Naver Shopping: 이 데이터셋은 20만 개의 네이버 쇼핑 리뷰로 구성되어 있으며, 각 리뷰는 별점과 함께 수집되었다. 중립적인 감성을 나타내는 3점 리뷰를 제외하였으며, 긍정(4~5점), 부정(1~2점) 리뷰의 비율이 1:1에 가깝게 구성되어 있다.

- Steam: 이 데이터셋은 10만 개의 Steam 게임 리뷰로 구성되어 있으며, 게임 커뮤니티의 특성상 비속어 및 은어가 다수 포함된 것이 특징이다. 이 데이터 또한 긍정과 부정 리뷰의 비율이 1:1에 가깝게 구성되어있다.

- NSMC: 이 데이터셋은 네이버 영화에서 수집한 한국어 영화 리뷰 20만 개로 구성되어 있으며 중립적인 감성을 나타내는 5~8점의 리뷰는 제외하였다.

세 가지 데이터셋에서 각각 랜덤 샘플링을 통해 미세 조정을 위한 학습용 데이터셋 10,000개와 평가용 데이터셋 2,000개를 구성했다. 또한, 긍정 및 부정의 비율이 1:1이 되도록 데이터셋을 구성했다.

본 실험에서는 한국어에 특화된 사전 학습된 BERT기반 모델인 KoBERT와 기존 한국어 감성 분석 관련 연구[30]에서 높은 성능을 보여준 LMKor_BERT, LMKor_ALBERT, LMKor_ELECTRA를 사용하였다. LMKor 기반 모델들은 국내 주요 상업 리뷰 1억개와 나무위키, 모두의 말뭉치 등 다양한 한국어 데이터 70GB를 학습에 사용하여 Pre-train되었다. 또한, Naver cloud에서 제공하는 서비스를 기반으로 Clova는 ShallowBert[38]를 활용하는 감성분석 api로 실험을 진행하였다. Clova의 경

1) https://github.com/bab2min/corpus/tree/master/sentiment

2) https://github.com/e9t/nsmc

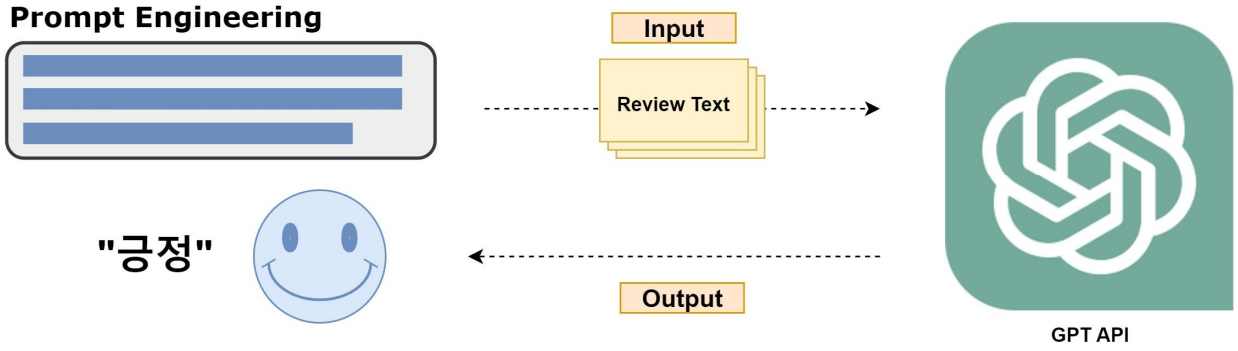

Fig. 2. Sentiment analysis process using GPT API

Table 1. Hardware environment and hyperparameters used in the experiment

Category	Decsription
GPU	Nvidia Geforce RTX 3090
OS	Windows 10
Epoch	5
Batch Size	128
Learning Rate	5e-5
Optimizer	AdamW
Scheduler	Cosine Scheduler
Warm Up Ratio	0.1

우 '중립'으로 판단을 내리는 경우가 종종있어 해당 값들은 모두 제거하고 평가하였다. LLM은 GPT-3.5 기반 가장 최근 모델인 GPT-3.5-turbo-1106과 GPT-4를 대상으로 실험을 진행하였다. 또한, 대규모 언어 모델의 zero-shot 감성 분류 성능에 대한 평가가 주 목적이기에 few-shot 방식이나 CoT(Chain of Thought)와 같은 프롬프트 형식을 사용하지 않았다. 실험에 사용된 데이터셋과 모델을 비롯한 전반적인 프레임워크는 Fig. 1에서 확인할 수 있다.

생성 모델의 한국어 감성 분석 성능과 기존의 사전 학습된 한국어 언어모델을 미세 조정한 성능과 비교하는 실험을 진행하였다. 모든 실험은 NVIDIA Geforce RTX 3090 GPU에서 진행하였다. 사전 학습된 한국어 모델은 5 에포크 동안 배치 크기 128, 학습률 5e-5로 미세 조정되었으며 AdamW를 사용하여 최적화를 진행했다. 효율적인 학습을 위해 Cosine 스케줄러를 사용했으며, wram up ratio는 0.1로 설정하였다. 실험에 사용한 하이퍼파라미터 설정은 사전에 진행된 한국어 감성 분석에 관한 연구[30]를 참고하여 최적의 값으로 선정하였다. 하드웨어 환경 및 하이퍼파라미터 설정은 Table 1에 요약되어 있다.

Clova의 경우 Naver cloud platform의 Clova Sentiment API를 활용하여 감성 분석을 진행하였다. GPT-3.5-turbo와 GPT-4는 ChatGPT API를 활용하였고 데이터셋에 따른 적절한 감성 분석 결과를 도출 하도록 프롬프트를 조정하였다. ChatGPT API를 활용한 감성분석 과정은 Fig. 2에서 볼 수 있다.

감성 분석 성능을 평가하는데 일반적으로 사용되는 평가지표인 정확도(Accuracy), F1-Score를 사용하여 모델의 감성 분석 성능을 평가하였다. 정확도는 모델이 전체 샘플 중에서 얼마나 많은 샘플을 정확하게 분류했는지를 의미하는 지표이다. F1 점수는 모델의 정밀도(Precision)와 재현율(Recall)의 조화 평균을 의미한다. 정밀도는 모델이 긍정으로 분류한 항목 중 실제로 긍정인 경우의 비율을, 재현율은 실제 긍정 항목 중 모델이 긍정으로 정확하게 분류한 비율을 뜻한다. 정확도가 종합적인 평가를 제공한다면 F1 score는 클래스 불균형이 있을 때 유용하게 사용될 수 있다.

IV. Experiments

본 연구에서는 생성 모델 및 기존의 사전 학습된 한국어 언어 모델들의 감성 분석 성능을 평가하였으며, 모든 실험은 한 번씩 수행하였다. 실험 결과는 Table 2에서 확인할 수 있다.

성능 비교 결과에 따르면, 정확도와 F1-score 모두에서 LMKor-ELECTRA 모델이 타 모델들에 비해 우수한 성능을 나타내었다. 이 모델은 데이터셋에 대한 미세 조정 과정을 거쳤으며, 실험에 사용된 사전 학습 언어 모델 중에서 가장 큰 크기를 가진 모델이다. Game 데이터셋에서는 0.83의 정확도를 기록하여 가장 높은 성능을 보였고, 나머지 두 데이터셋에서도 두 번째로 높은 성능을 나타내었다.

Table 2. Table showing the accuracy and F1 values of the models according to the datasets. The Steam dataset is named 'Game' and the NSMC dataset is named 'Movie'. Accuracy is labeled as 'A' and F1 value is labeled as 'F', each expressed in % and rounded to two decimal places. Scores in bold indicate the highest score, while scores in underline indicate the second highest score.

Dataset	KoBERT		Lmkor-BERT		Lmkor-ALBERT		Lmkor-ELECTRA		Hyper Clova		ChatGPT 3.5-turbo		ChatGPT4	
	A(%)	F(%)	A	F	A	F	A	F	A	F	A	F	A	F
Shopping	91.6	91.5	93.6	93.7	92.8	92.8	94.0	**94.1**	91.5	91.9	92.2	92.1	**94.1**	94.0
Game	78.7	78.6	82.8	82.8	80.1	79.5	**82.9**	**82.9**	74.1	66.6	75.0	68.7	79.4	75.4
Movie	84.6	84.5	88.1	88.0	85.5	85.1	88.5	**88.6**	81.6	80.2	83.8	83.0	**88.7**	88.0

F1-score 기준으로는 모든 데이터셋에서 각각 0.89, 0.83, 0.94의 성능을 보여, 가장 높은 성능을 기록하였다. 반면 제로 샷 학습을 기반으로 하는 ChatGPT-4는 Movie와 Shopping 데이터셋에서 각각 0.89와 0.94의 높은 정확도를 보였으며, F1-score에서도 두 데이터셋 모두에서 두 번째로 높은 성능을 나타내었다. 이러한 결과는 특정 데이터로 학습되지 않았음에도 불구하고 Zero-shot 기반의 대규모 언어 모델이 한국어 감성 분석에 있어 사전학습된 언어 모델과 유사하거나 더 높은 성능을 얻을 수 있음을 시사한다.

다만, Movie 데이터셋과 Shopping 데이터셋에 비해 Game 데이터셋에서의 상대적으로 낮은 성능은 게임 리뷰의 특성상 은어나 비속어의 사용이 예측에 어려움을 주었기 때문으로 분석된다. 마찬가지로 제로 샷 기반 모델들이 미세 조정된 모델들에 비해 Game 데이터셋에서 낮은 성능을 보인 것도 추가 학습 없이는 높은 성능을 기록하는데 어려움을 주기 때문으로 해석할 수 있다.

V. Conclusions

본 연구는 GPT-3.5, GPT-4, Naver Hyper Clova, Bard와 같은 대규모 언어 모델과 기존에 사전 학습된 한국어 언어 모델들의 한국어 감성 분석 성능을 비교하는 실험을 진행하였다. 이를 위해 한국어 감성 분석 연구에 주로 사용되는 Naver Shopping, Steam, NSMC 데이터셋을 사용하였다.

연구 질문 2에 대한 실험 결과, 대규모 언어 모델들 중 특히 LMKor-ELECTRA 모델이 한국어 감성 분석에서 우수한 성능을 보임을 시사한다. 이 모델은 정확도와 F1-score 모두에서 높은 성과를 기록하였으며, 특히 Game 데이터셋에서의 정확도가 0.83으로 가장 높았다. 모든 데이터셋에서 F1-score는 각각 0.89, 0.83, 0.94로 가장 높은 성능을 보여주었다. 반면, 제로 샷 기반인 ChatGPT-4는 Movie와 Shopping 데이터셋에서 0.89와 0.94로 가장 높은 정확도를 나타냈다. 이러한 결과는 연구 질문 1에 대하여 학습 데이터에 포함되지 않은 언어 표현에도 불구하고, 영어가 아닌 한국어 데이터에서도 대규모 언어 모델의 zero-shot 기반 방법이 높은 성능을 얻을 수 있음을 입증한다. 그러나 Game 데이터셋에서의 상대적으로 낮은 성능은, 게임 리뷰의 특성상 은어나 비속어의 사용이 예측에 어려움을 준 것으로 해석된다. 이러한 어려움은 제로 샷 기반 모델들이 미세 조정된 모델들에 비해 낮은 성능을 보이는 원인이 될 수 있다.

이와 같은 실험 결과는 다음과 같은 방법으로 성능을 개선 시킬 여지가 남아 있다. 사전 학습된 언어 모델의 경우, 미세 조정에 사용된 데이터의 수를 증가시킨다면 성능이 증가할 것이다. 특히 Game 데이터셋의 경우 사전 학습된 텍스트 데이터와 상대적으로 상이하므로 미세 조정용 데이터의 수가 증가할 경우 더 큰 성능 변화를 얻을 수 있을 것이다. Zero-shot 기반 대규모 언어 모델의 경우, 보다 적절한 프롬프트 입력과 CoT 방식의 활용을 통해 성능을 개선할 수 있을 것이다. 또한, zero-shot 기법 대신 few-shot 기법의 활용을 고려한다면 다른 결과를 기대할 수 있다.

본 연구는 대규모 언어 모델과 미세 조정된 사전 학습된 한국어 모델의 한국어 감성 분석 성능을 비교함으로써, 한국어 감성 분석에 대한 대규모 언어 모델의 활용 가능성을 발견하였다. 이는 한국어 감성 분석 연구와 응용에 있어 추후 연구에 중요한 방향을 제시한다.

반면 본 연구는 다음과 같은 한계점이 존재한다. 첫째, 대규모 언어 모델의 다양성을 고려하지 않았다. Google의 BARD를 비롯해 다양한 대규모 언어 모델이 존재하지만 가장 대표적인 ChatGPT의 성능만을 비교하였다. 둘째, 평가용 데이터의 양이 상대적으로 작다, API 가격 문제로 인해 평가용 데이터의 양을 적게 산정하였는데, 이는 실험의 신뢰성을 저하시킬 수 있다.

ACKNOWLEDGEMENT

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (No. RS-2023-00273751).

REFERENCES

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017). DOI: 10.48550/arXiv.1706.03762

[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018). DOI: 10.48550/arXiv.1810.04805

[3] Min, Bonan, et al. "Recent advances in natural language processing via large pre-trained language models: A survey." ACM Computing Surveys 56.2 (2023): 1-40. DOI: 10.1145/3605943

[4] Open AI, ChatGPT, https://chat.openai.com/

[5] Google, BARD, https://bard.google.com/chat

[6] Achiam, OpenAI Josh et al. “GPT-4 Technical Report.” (2023). DOI: 10.48550/arXiv.2303.08774

[7] Anil, Rohan, et al. "Palm 2 technical report." arXiv preprint arXiv:2305.10403 (2023). DOI: 10.48550/arXiv.2305.10403

[8] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams engineering journal 5.4 (2014): 1093-1113. DOI: 10.1016/j.asej.2014.04.011

[9] Al-Thubaity, Abdulmohsen, et al. "Evaluating ChatGPT and Bard AI on Arabic Sentiment Analysis." Proceedings of ArabicNLP 2023. 2023. DOI: 10.18653/v1/2023.arabicnlp-1.27

[10] Zhao, Wayne Xin, et al. "A survey of large language models." arXiv preprint arXiv:2303.18223 (2023). DOI: 10.48550/arXiv.2303.18223

[11] Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

[12] Radford, Alec, et al. "Better language models and their implications." OpenAI blog 1.2 (2019).

[13] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." The Journal of Machine Learning Research 21.1 (2020): 5485-5551. DOI: 10.48550/arXiv.1910.10683

[14] Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461 (2019). DOI: 10.48550/arXiv1910.13461

[15] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019). DOI: 10.48550/arXiv.1907.11692

[16] Beltagy, Iz, Kyle Lo, and Arman Cohan. "SciBERT: A pretrained language model for scientific text." arXiv preprint arXiv:1903.10676 (2019). DOI: 10.48550/arXiv.1903.10676

[17] Luo, Renqian, et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining." Briefings in Bioinformatics 23.6 (2022): bbac409. DOI: 10.48550/arXiv.2210.10341

[18] Floridi, Luciano, and Massimo Chiriatti. "GPT-3: Its nature, scope, limits, and consequences." Minds and Machines 30 (2020): 681-694. DOI: 10.1007/s11023-020-09548-1

[19] Lund, Brady D., and Ting Wang. "Chatting about ChatGPT: how may AI and GPT impact academia and libraries?." Library Hi Tech News 40.3 (2023): 26-29. DOI: 10.1108/LHTN-01-2023-0009

[20] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023). DOI: 10.48550/arXiv.2302.13971

[21] Ouyang, Xi, et al. "Sentiment analysis using convolutional neural network." 2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing. IEEE, 2015. DOI: 10.1109/CIT/IUCC/DASC/PICOM.2015.349

[22] Yoosin Kim, Namgyu Kim, and Seong Ryoul Jeong, "Stock-Index Invest Model Using News Big Data Opinion Mining," Journal of Intelligence and Information Systems, Vol. 18, No. 2, pp. 143-156, 2012. DOI: 10.13088/JIIS.2012.18.2.143

[23] Sang Hoon Lee, Jing Cui, and Jong Woo Kim, "Sentiment analysis on movie review through building modified sentiment dictionary by movie genre," Journal of Intelligence and Information Systems, Vol. 22, No. 2, pp. 97-113, 2016. DOI: 10.13088/JIIS.2016.22.2.097.

[24] Sang-Min Park, Chul-Won Na, Min-Seong Choi, Da-Hee Lee, and Byung-Won On, "KNU Korean Sentiment Lexicon - Bi-LSTM-based Method for Building a Korean Sentiment Lexicon -," Journal of Intelligence and Information Systems, Vol. 24, No. 4, pp. 219-240, 2018. DOI: 10.13088/JIIS.2018.24.4.219.

[25] Hyun-jung Park, Min-chae Song, and Kyung-shik Shin, "Sentiment Analysis of Korean Reviews Using CNN - Focusing on Morpheme Embedding -," Journal of Intelligence and Information Systems, Vol. 24, No. 2, pp. 59-83, 2018. DOI: 10.13088/JIIS.2018.24.2.059.

[26] Jinju Hong, Sehan Kim, Jeawon Park, and Jaehyun Choi, "A Malicious Comments Detection Technique on the Internet using Sentiment Analysis and SVM," Journal of the Korea Institute of Information and Communication Engineering, Vol. 20, No. 2, pp. 260-267, 2016. DOI: 10.6109/jkiice.2016.20.2.260.

[27] Jiyeon Hyun, Sangyi Ryu, and Sang-Yong Tom Lee, "How to improve the accuracy of recommendation systems : Combining ratings and review texts sentiment scores," Journal of Intelligence and Information Systems, Vol. 25, No. 1, pp. 219-239, 2019. DOI: 10.13088/JIIS.2019.25.1.219.

[28] SK Telecom, KoBERT, https://github.com/SKTBrain/KoBERT

[29] Kiyongkim1, LMkor, https://github.com/kiyoungkim1/LMkor

[30] Jae-hong Lee "Comparison of Sentiment Classification Performance of for RNN and Transformer-Based Models on Korean Reviews" Journal of The Korea Institute of Electronic Communication Sciences 18.04 pp.693-700 (2023) : 693. DOI: 10.13067/JKIECS.2023.18.4.693

[31] Qin, Chengwei, et al. "Is ChatGPT a general-purpose natural language processing task solver?." arXiv preprint arXiv:2302.06476 (2023). DOI: 10.48550/arXiv.2302.06476

[32] Kocoń, Jan, et al. "ChatGPT: Jack of all trades, master of none." Information Fusion (2023): 101861. DOI: 10.1016/j.inffus.2023.101861

[33] Wang, Zengzhi, et al. "Is ChatGPT a good sentiment analyzer? A preliminary study." arXiv preprint arXiv:2304.04339 (2023). DOI: 10.48550/arXiv.2304.04339

[34] Amin, Mostafa M., Erik Cambria, and Björn W. Schuller. "Will affective computing emerge from foundation models and general ai? A first evaluation on chatgpv:2303.03186 (2023). DOI: 10.48550/arXiv.2303.03186

[35] Bang, Yejin, et al. "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity." arXiv preprint arXiv:2302.04023 (2023). DOI: 10.48550/arXiv.2302.04023

[36] Al-Thubaity, Abdulmohsen, et al. "Evaluating ChatGPT and Bard AI on Arabic Sentiment Analysis." Proceedings of ArabicNLP 2023. 2023. DOI: 10.18653/v1/2023.arabicnlp-1.27

[37] HyoungDo Kim, "Zero-shot Sentiment Analysis of Movie Reviews Based on GPT-4," JOURNAL OF THE KOREA CONTENTS ASSOCIATION, Vol. 23, No. 11, pp. 185-196, 2023. DOI: 10.5392/JKCA.2023.23.11.185

[38] Li, Xiang, et al. "A shallow BERT-CNN model for sentiment analysis on MOOCs comments." 2019 IEEE International Conference on Engineering, Technology and Education (TALE). IEEE, 2019. DOI: 10.1109/TALE48000.2019.9225993

Authors

Soon-Chan Kwon received the B.S degree in public administration & business from Kwangwoon University, Seoul, Korea in 2023. He is currently pursuing the M.S. degree in Graduate School of information, Yonsei University, Seoul, Korea. Soon-Chan Kwon is interested in Altificial Intelligence, NLP and Recommendation System.

Dong-Hee Lee received the B.S degree in business administration from Dongguk University, Seoul, Korea in 2023. He is currently pursuing the M.S. degree in Graduate School of information, Yonsei University, Seoul, Korea. Dong-Hee Lee is interested in Deep Learning, NLP and Causal Machine Learning.

Beak-Cheol Jang received B.S. degree in Computer Science from Yonsei University, Korea in 2001, M.S. in Computer Science from Korea Advanced Institute of Science and Technology, Korea in 2002, Ph.D in Computer Science from North Carolina State University, NC, USA in 2009. Beak-Cheol Jang is currently a Professor in the Graduate School of information, Yonsei University. He is interested in Wireless Networking, Artificial Intelligence.