Regular paper | Journal of the KIECS. pp. 693-700, vol. 18, no. 4, Aug. 31. 2023, t. 120, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2023.18.4.693

RNN과 트랜스포머 기반 모델들의 한국어 리뷰 감성분류 비교

이재홍*

Comparison of Sentiment Classification Performance of for RNN and Transformer-Based Models on Korean Reviews

Jae-Hong Lee*

요 약

텍스트 문서에서 주관적인 의견과 감정을 긍정 혹은 부정으로 분류하고 식별하는 자연어 처리의 한 분야인 감성 분석은 고객 선호도 분석을 통해 다양한 홍보 및 서비스에 활용할 수 있다. 이를 위해 최근 머신러닝과 딥러닝의 다양한 기법을 활용한 연구가 진행되어 왔다. 본 연구에서는 기존의 RNN 기반 모델들과 최근 트랜스포머 기반 언어 모델들을 활용하여 영화, 상품 및 게임 리뷰를 대상으로 감성 분석의 정확도를 비교 분석하여 최적의 언어 모델을 제안하고자 한다. 실험 결과 한국어 말뭉치로 사전 학습된 모델들 중 LMKor-BERT와 GPT-3가 상대적으로 좋은 정확도를 보여주었다.

ABSTRACT

Sentiment analysis, a branch of natural language processing that classifies and identifies subjective opinions and emotions in text documents as positive or negative, can be used for various promotions and services through customer preference analysis. To this end, recent research has been conducted utilizing various techniques in machine learning and deep learning. In this study, we propose an optimal language model by comparing the accuracy of sentiment analysis for movie, product, and game reviews using existing RNN-based models and recent Transformer-based language models. In our experiments, LMKorBERT and GPT3 showed relatively good accuracy among the models pre-trained on the Korean corpus.

키워드

Deep Learning, Transformer, BERT, GPT, Sentiment Analysis, Natural Language Processing
딥러닝, 트랜스포머, BERT, GPT, 감성 분석, 자연어 처리

Ⅰ. 서 론

텍스트 문서에서 주관적인 의견과 감정을 긍정 혹은 부정으로 분류하고 식별하는 자연어 처리의 한 분야로 감성 분석은 기존 CNN(Convolution Neural Network), RNN(: Recursive Neural Network)과 같은 딥러닝 모델에서 트랜스포머(transformer) 기반 모델들이 출현하면서 좋은 성능을 보여주고 있다. 상품 구입 후기, 영화나 콘텐츠에 대한 비평 등에 대하여 감성 분석 결과는 고객 선호도 분석을 통해 서비스, 홍

* 교신저자 : 전남도립대학교 보건의료과
• 접 수 일 : 2023. 06. 27
• 수정완료일 : 2023. 07. 20
• 게재확정일 : 2023. 08. 17
• Received : Jun. 27, 2023, Revised : Jul. 20, 2023, Accepted : Aug. 17, 2023
• Corresponding Author : Jae-Hong Lee
Dept. of Health & Medical Science, Jeonnam State University
Email : jhlee@dorip.ac.kr

보, 브랜드 마케팅 등에 응용될 수 있어 보다 좋은 분석 결과를 얻기 위하여 적절한 언어 모델을 선택하는 것이 중요하다.

본 연구에서는 네이버의 영화 및 쇼핑 리뷰, 게임 사용 리뷰를 대상으로 자연어 처리 중 감성분류에 사용될 수 있는 RNN과 트랜스포머 기반 파생모델을 대상으로 정확도를 비교하여, 한국어 감성 분류에 적합한 언어 모델을 선정하고자 한다. Ⅱ장에서는 감성 분석 분야의 관련 연구를 소개하고, Ⅲ장에서는 한국어 처리를 위한 언어 모델들에 대해 설명한다. Ⅳ장에서는 감성 분석을 위한 데이터 셋을 살펴본 후, 다양한 언어 모델들을 대상으로 실험 방법과 결과를 설명하고, 마지막으로 한국어 감성 분석을 위한 언어 모델을 제안한다.

Ⅱ. 관련 연구

자연어 처리, 이미지 분류와 검색, 빅데이터 분석 등의 분야에서 CNN, RNN과 같은 딥러닝 기술이 우수한 결과를 보여주었다[1,2].

자연어 처리의 한 분야인 감성 분석 분야에서는 CNN, RNN과 그 파생 모델을 이용한 딥러닝 모델과 트랜스포머 기반 BERT(: Bidirectional Encoder Representation from Transformer) 모델을 활용한 다양한 연구가 진행되었다. [3]에서는 RNN 모델 구조를 개선한 양방향 LSTM(BiLSTM)을 병렬로 쌓아올리고, 단어, 음절, 음소 임베딩을 앙상블 처리하여 기존 RNN, LSTM(: Long Short-Term Memory) 보다 분류 정확도를 개선하였다. [4]에서는 EEG 기반으로 효율적인 감성 분류를 위해 LSTM을 위한 최적의 하이퍼파라미터를 파악하고자 하였다. [5]에서는 형태소 분석기를 사용하는 기존 방법에 신조어와 같은 미등록어 문제를 보완한 하이브리드 토크나이저를 제안하고 이를 한국어 RoBERTa 모델에 적용하여 다양한 자연어 처리 분야에서 성능을 개선하였다. [6]에서는 네이버 영화평 말뭉치를 확장한 데이터 셋을 추가 구축하여 SKTBrain에서 공개한 KoBERT를 학습하여 LSTM과 다국어 버전의 BERT와 성능을 비교하여 한국어 말뭉치로 학습한 KoBERT가 우수한 성능을 보여주었다. [7]에서는 네이버 영화 리뷰 데이터를 대상으로 다층 퍼셉트론, CNN, LSTM과 트랜스포머 계열의 BERT를 적용하고 이들의 예측 결과에 대해 다수결 방식의 하드 보팅(voting) 기법을 적용한 앙상블을 수행하여 예측 정확도를 높이고자 하였다. [8]에서는 모델들을 학습시키기 위해 네이버 영화 리뷰 데이터에 군산대학교의 “KNU 한국어 감성 사전”[1]을 추가하여 훈련 데이터로 단일 데이터 셋을 적용하는 것보다 감성 분류 정확도를 높였음을 보여 주었다. [9]에서는 BERT의 8개의 공개된 사전학습 자연어 처리 모델을 구현하여 다양한 데이터 셋에 적용하여 성능 검증을 한 후, LMkor-BERT가 상대적으로 평균 정확성이 좋은 모델로 평가하고 있다.

Ⅲ. 자연어 처리

3.1 언어 모델

기존 딥러닝 연구에서 자연어 처리를 위해 사용되던 RNN 모델과 파생 모델인 LSTM과 GRU(: Gated Recurrent Units)를 대신하여 2017년 6월 허깅페이스(Huggingface)의 트랜스포머[2]가 등장하며 자연어 처리 분야에서 획기적인 도약이 이루어졌다[3]. 트랜스포머의 인코더만을 사용하는 BERT와 디코더를 사용하는 GPT(: Generative Pre-trained Transformer)가 두 축을 이루며 텍스트 분류, 감성 분석, 기계 번역, 문서 요약, 개체명 인식, 질의 응답, 문장 생성 등에서 괄목할만한 업적을 가져왔으며, 2022년 11월 OpenAI에서 공개된 ChatGPT는 GPT-3를 기반으로 한 채팅로봇으로 기존 검색 분야를 벗어나 시, 소설, 회화, 작곡에 이르는 창작 분야에 까지 그 활용 범위를 확대하고 있다.

트랜스포머 모델은 인코더(Encode) 기반의 BERT 계열과 디코더(Decoder) 기반의 GPT 계열, 그리고 인코더와 디코더를 모두 사용하는 계열로 크게 구분할 수 있다.

표 1의 트랜스포머 모델들은 데이터에 레이블을 지정할 필요가 없는 자가지도(self-supervised) 학습방식으로 많은 양의 데이터에 대해 학습된 언어 모델들

1) https://github.com/park1200656/KnuSentiLex
2) https://arciv.org/abs/1706.03762
3) https://jalammar.github.io/illustrated-transformer

이다. 이런 모델들은 학습된 언어에 대한 통계적 이해를 가능하게 하지만 문장 분류, 감성 분석, 질의 응답 등 실제 태스크에는 유용하지 않다. 따라서 이런 사전 학습된(pretrained) 언어 모델은 주어진 태스크에 맞는 데이터 셋(dataset)을 사용하여 추가 학습하는 방식으로 미세 조정(fine-tuning)되게 된다. 이때 사전 학습된 모델에 사용된 파라미터들이 재사용되므로 전이 학습(transfer learning)이라 한다. 사전 학습된 언어 모델을 사용하지 않고 새로운 태스크의 데이터 셋에 처음부터 특정 언어 모델을 학습하는 데에는 엄청난 컴퓨팅 파워와 시간이 소요되므로 일반적으로 다양한 데이터 셋에 대해 충분히 학습된 사전 학습된 모델을 사용하여 특정 태스크에 맞은 작은 데이터 셋에 대해 전이 학습시키는 방법을 사용하게 된다.

표 1. 트랜스포머 기반 모델들[4]
Table 1. Transformer based Models

Model	Derived Model	Tasks
Encoder based	BERT, ALBERT, DistilBERT, RoBERTa, ELECTRA, XLM	Text classification, Named entity recognition, Extractive Question and Answering
Decoder based	GPT-2, GPT-3, CTRL, GPT-J	Text generation
Encode-Decoder based	BART, T5, M2M-100, BigBird	Text summarization. Translation, Generative Question and Answering

BERT는 트랜스포머 기반 양방향 인코더로 MLM(Masked Language Modeling, 마스크 언어 모델링)과 NSP(Next Sentence Prediction, 다음 문장 예측)라는 두 가지 태스크를 이용해 거대한 말뭉치를 기반으로 사전 학습되었다[10]. ALBERT는 기존 BERT 모델의 2가지 단점을 개선했다. 먼저 모델의 파라미터 수를 줄이고 계층 사이에 파라미터를 공유하도록 구조를 변경하여 모델 크기를 줄였고, NSP 학습 방법을 변경하여 자연어 추론에서 기존보다 높은 성능을 보였다. DistillBERT는 BERT의 파생 모델로 BERT 보다 속도 향상과 메모리 소비 감소를 줄였지만 BERT 성능의 97%를 유지하고 있다. RobERTa는 기존 BERT에서 다음 문장을 예측하는 NSP는 사용하지 않고 MLM과 매 학습 때마다 임의의 단어가 동적으로 마스킹되는 방식을 사용하고 있다.

GPT는 주어진 텍스트를 기반으로 다음 단어를 잘 예측할 수 있도록 학습된 언어 모델이며, 특히 문장 생성에 최적화되어 있다.

3.2 한국어의 자연어 처리

KoNLPy(Korean NLP in Python)[5]는 한국어의 자연어 처리를 위한 파이썬 패키지로 Hannanum, Kkma, Komoran, Okt, Mecab의 형태소 분석기를 지원한다. KoNLPy 내부 모듈 간 비교에서 Mecab과 Okt(Open-source Korean Text Processor)가 상대적으로 실행 시간과 형태소 분석에서 조금 좋은 성능을 보이고 있다[6]. 이 중 Mecab은 윈도우즈 환경을 지원하지 않고 있으며, Okt[7]는 트위터에서 개발한 한국어 처리기에서 파생된 오픈소스 한국어 처리기이다.

3.3 한국어 언어모델

본 연구에서는 게임 플랫폼 스팀의 한국어 리뷰에 대해 RNN의 향상된 버전인 양방향의 BiLSTM(Bidirectional LSTM)[8]과 GRU, 사전 학습된 언어 모델인 M-BERT(Multilingual-BERT)[9], KLUE-BERT[10], GPT-2를 사용하여 예측 정확도를 비교하였다. BiLSTM과 Bi-GRU 모델에서는 형태소 분석기로는 KoNLPy의 Okt를 사용하였으며, 사전 학습된 언어 모델인 M-BERT, KLUE-BERT, GPT2는 자체 토크나이저를 사용하였다.

4) https://wikidocs.net/166785

5) http://konlpy.org/en/latest
6) http://konlpy.org/ko/v0.6.0/morph
7) https://github.com/open-korean-text/open-korean-text
8) https://arXiv:1508.01991
9) https://huggingface.co/bert-base-multilingual-uncased
10) https://github.com/KLUE-benchmark/KLUE

BERT 기반 사전 학습된 언어 모델의 한국어 버전에는 KLUE-BERT 외에도 Pytorch 기반의 KoBERT와 KoBART[11], KoELECTRA[12] 등이 있으나 본 논문에서는 Tensorflow/Keras 버전으로 코딩하여 사용하였다. KoBART는 다양한 문서 요약 공개 데이터로 튜닝 하여 문서 요약, 분류, 번역, 질의 응답 등에 좋은 성능을 보여준다.

BiLSTM은 2개의 독립적인 LSTM 구조를 함께 사용한다. 순방향 LSTM 마지막 시점의 은닉 상태와 역방향 LSTM 첫 시점의 은닉 상태를 출력층으로 보내는 구조이다. Bi-GRU는 BiLSTM에서 LSTM을 GRU로 대체한 것이다.

본 연구에서는 트랜스포머 모델이 등장하기 전 텍스트 분류 등에 사용되어 왔던 RNN의 파생 모델인 LSTM과 GRU를 양방향으로 입력받도록 수정된 BiLSTM을 감성 분류에 우선 적용하여 트랜스포머 기반 모델들과 성능을 비교하였다. 트랜스포머 기반 모델들 중에서는 BERT와 GPT에서 파생된 모델들 중 한국어를 처리할 수 있도록 한 사전 학습된 모델들을 선택하였다.

BERT 모델들로는 BERT의 다국어 지원 모델인 M-BERT, BERT 모델의 한국어 버전 중 대표적인 KLUE-BERT, KLUE-RoBERTa, LMKor-BERT, LMkor-ALBERT를 선정하였다. MLM을 이용하여 세계 언어 중 상위 102개 언어에 대해 사전 학습된 모델인 M-BERT(Multilingual BERT)을 기반으로 한국어 리뷰에 대한 가중치를 조정하였다. KLUE-BERT (Korean Language Understanding Evaluation)[13]는 BERT의 한국어 버전 중 하나로 MODU 말뭉치[14], CC-100-Kor(CC-Net 웹 크롤러를 사용하여 수집된 대규모 다국어 말뭉치), 나무위키, 2011년~2020년의 1,280만 뉴스 기사, 사회 문제에 대한 청와대 진정서 모음 등 대략 62GB 말뭉치로 학습하여, 주제 분류, 문맥 유사도, 자연어 추론, 개체명 인식 등에 사용될 수 있다. LMkor-BERT, LMkor-ALBERT[15]는 국내 주요 상업 리뷰 1억 개, 블로그 형 웹사이트 2천만 개에서 획득한 75GB, 모두의 말뭉치 18GB, 위키백과와 나무위키 6GB,총 99GB 중 불필요하거나 너무 짧은 문장과 중복된 문장들을 제외하여 최종적으로 70GB의 텍스트 데이터를 학습에 사용하여 40만개의 어휘 사전을 작성하여 사전 학습되었다. XLM-R[16]은 다중 언어 입력으로 확장한 XLM과 RoBERTa의 뒤를 이어 2.5테라바이트의 텍스트 데이터를 이용하여 MLM으로 인코더를 훈련하여 데이터가 부족한 언어에서 XLM과 다중 언어 BERT 변종인 M-BERT에 비해 큰 차이의 성능을 보였다. XLM-R은 XLM에서 사용하는 언어 임베딩도 제거하고 SentencePiece를 사용하여 원시 텍스트를 직접 토큰화하여 기존의 M-BERT를 대체하였다.

GPT의 한국어 버전인 KoGPT2[17]는 1억 2,500만개의 변수를 사용하고 한국어 위키피디아 외에 뉴스, 모두의 말뭉치 v1.0, 청와대 국민청원 등 약 40GB 이상의 한국어 텍스트를 이용해 사전 학습되었다. 이를 위해 바이트 쌍 인코딩(BPE, Byte Pair Encoding) 토크나이저로 학습했고 어휘 크기는 51,200이며, 대화에 자주 사용하는 이모티콘, 이모지 등도 추가했다. GPT-3는 아직 공개되지 않았지만 GPT-2를 기반으로 GPT-3의 공개된 특성을 반영한 GPT-3 시험버전도 감성 분류 성능 비교에 사용하였다.

Ⅳ. 감성 분석 비교

4.1 데이터 셋

본 연구에서는 네이버로 부터 수집한 감성 분석용 말뭉치인 쇼핑 후기[18]와 영화 리뷰[19], 그리고 게임 플랫폼 스팀(STEAM)의 리뷰[20]를 이용하였다. 공개된 말뭉치에는 감성 분석에 사용할 수 있도록 극성

11) https://github.com/SKT-AI/KoBART
12) https://huggingface.co/jaehyeong/koelectra-base-v3-generalized-sentiment-analysis
13) https://arXiv:2105.09680v4
14) https://github.com/bab2min/corpus/tree/master/sentiment
15) https://github.com/kiyoungkim1/LMkor
16) https://arXiv:1911.02116v2
17) https://github.com/SKT-AI/KoGPT2
18) https://raw.githubusercontent.com/bab2min/corpus/master/sentiment/naver_shopping.txt
19) https://github.com/e9t/nsmc
20) https://github.com/bab2min/corpus/tree/master/sentiment/steam.txt

(긍정·부정)이 라벨링된 텍스트 데이터로 되어 있으며, 네이버 쇼핑에서 제품별 후기를 별점과 함께 수집된 것과 게임 유통 서비스인 스팀의 각종 게임에 댓글로 달린 한국어 리뷰를 수집한 것으로 이루어져 있다.

네이버 쇼핑 제품후기 데이터는 그림 1에서와 같이 탭으로 분리되어 있으며, 첫 번째 필드에는 1점에서 5점의 별점, 두 번째 필드에는 리뷰가 있다. 별점 중 4~5점은 긍정, 1~2점은 부정으로 분류하였고 각각 99,963개와 100,037개로 모두 20만개의 리뷰로 구성되어 있으며, 긍정과 부정 1:1 비율로 구성되어 있다. 20만개의 리뷰 중 한글과 공백을 제외하는 전처리를 거쳐 훈련용 149,931개, 훈련용 49,997개를 사용하였다.

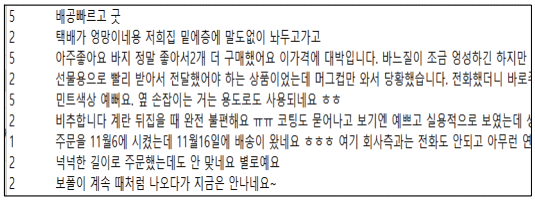
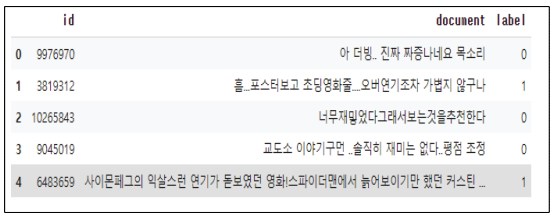

5 배공빠르고 굿
2 택배가 엉망이네용 저희집 밑에층에 말도없이 놔두고가고
5 아주좋아요 바지 정말 좋아서2개 더 구매했어요 이가격에 대박입니다. 바느질이 조금 엉성하긴 하지만
2 선물용으로 빨리 받아서 전달했어야 하는 상품이었는데 머그컵만 와서 당황했습니다. 전화했더니 바로
5 민트색상 예뻐요. 옆 손잡이는 거는 용도로도 사용되네요 ㅎㅎ
2 비추합니다 계란 뒤집을 때 완전 불편해요 ㅠㅠ 코팅도 묻어나고 보기엔 예쁘고 실용적으로 보였는데
1 주문을 11월6에 시켰는데 11월16일에 배송이 왔네요 ㅎㅎㅎ 여기 회사측과는 전화도 안되고 아무런 연
2 넉넉한 길이로 주문했는데도 안 맞네요 별로예요
2 보풀이 계속 때처럼 나오다가 지금은 안나네요~

그림 1. 네이버 쇼핑 리뷰의 내용
Fig. 1 Contents of naver shopping reviews

네이버 영화 리뷰(NSMC, Naver Sentiment Movie Corpus)는 그림 2에서 와 같이 리뷰(document)와 평점(label)이 한 쌍으로 이루어져 있으며, 실험에 사용된 리뷰 데이터에서는 1~10점의 평점 중 1~4점인 경우 부정적 레이블(0)로, 9~10점인 경우 긍정적 레이블(1)로 재할당하고 5~8점은 제외했다. 리뷰 데이터는 훈련용 리뷰로 15만개, 테스트용 리뷰는 5만개로 긍정과 부정 평점이 1:1이다.



	id	document	label
0	9976970	아 더빙.. 진짜 짜증나네요 목소리	0
1	3819312	흠...포스터보고 초딩영화줄...오버연기조차 가볍지 않구나	1
2	10265843	너무재밋었다그래서보는것을추천한다	0
3	9045019	교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정	0
4	6483659	사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 ...	1

그림 2. 네이버 영화 리뷰의 내용
Fig. 2 Contents of naver movie reviews

게임 플랫폼 스팀의 한국어 리뷰는 5만개의 긍정리뷰와 5만개의 부정 리뷰의 총 10만개의 리뷰로 구성되었다. 각 리뷰는 그림 3과 같이 긍정(1)과 부정(0)의 라벨과 리뷰 내용으로 이루어져 있다. 전체 리뷰 10만개에서 중복된 내용을 제거한 99,892개의 리뷰를 학습에 사용하였다.

전이 학습용으로 전체 10만개의 리뷰 중 중복된 내용을 제거한 후, 훈련용 74,919개와 검증용 24,973개로 75:25의 비율로 분할하였다. 학습, 검증, 테스트용 데이터 셋 모두 긍정과 부정 리뷰의 비율은 1:1로 구성하였다.

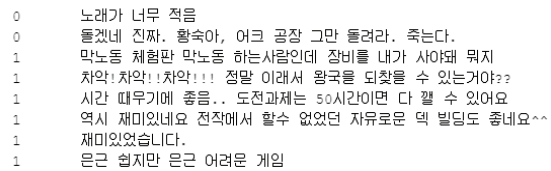

0 노래가 너무 적음
0 돌겠네 진짜. 황숙아, 어크 공장 그만 돌려라. 죽는다.
1 막노동 체험판 막노동 하는사람인데 장비를 내가 사야돼 뭐지
1 차악!차악!!차악!!! 정말 이래서 왕국을 되찾을 수 있는거야??
1 시간 때우기에 좋음.. 도전과제는 50시간이면 다 깰 수 있어요
1 역시 재미있네요 전작에서 할수 없었던 자유로운 덱 빌딩도 좋네요^^
1 재미있었습니다.
1 은근 쉽지만 은근 어려운 게임

그림 3. 게임 플랫폼 Steam의 한국어 리뷰
Fig. 3 Korean reviews of the gaming platform Steam

4.2 실험 환경

한국어 리뷰 감성 분석을 위한 실험 환경으로 구글의 코랩(Colab Pro)을 사용하였으며, 런타임 환경으로 GPU(Graphics Processing Unit) 유형 A100와 TPU(Tensor Processing Unit)를 활용하여 학습속도를 높였다.

BiLSTM과 Bi-GRU에서는 임베딩 크기로 100, 은닉상태 크기로 128, 옵티마이저로 rmsprop와 손실함수로는 BinaryCrossentropy, 임베딩 크기로 100을 사용하였다.

BERT 언어모델에서는 옵티마이저로 Adam(학습률은 5e-5), 손실함수로 BinaryCrossentropy, 입력 시퀀스의 최대 길이로 128을 사용하였으며, 2에포크(epoch) 전이 학습하였다.

GPT-3는 2023년 현재 공개되지 않았지만 GPT-2를 기반으로 작성된 체크포인트를 사용하였다. GPT는 훈련에 2에포크를 사용하였다.

각 모델의 감성 분류 클래스에서는 모델의 마지막 출력에 완전 연결층(Dense층) 1층을 적용하고 과적합(overfitting)을 줄이기 위해 Dropout층을 추가하였다.

BiLSTM과 Bi-GRU에서는 한국어 형태소 분석을 위해 Okt를 사용하였으며, 트랜스포머 기반 모델들은 각 모델 고유의 토크나이저를 사용했다. 단지 리뷰 데

이터 셋들에 불용어(stopwords)나 학습에 영향을 주지 않을 빈도수 적은 단어들은 제외하는 등의 전처리 과정을 거쳤다.

표 2와 표 3은 실험에 사용된 언어 모델들과 하이퍼파라미터들을 보여준다.

표 2. 실험에 사용된 언어모델들
Table 2. Language models used in the experiment

Model	Derived model (Pre-trained language model)	checkpoint
RNN	BiLSTM	-
	Bi-GRU	-
BERT	M-BERT	bert-base-multilingual-cased
	KLUE-BERT	klue/bert-base
	LMkor-BERT	kykim/bert-kor-base
	LMkor-ALBERT	kykim/albert-kor-base
	KLUE-RoBERTa	klue/roberta-base
	XLM-R	xlm-roberta-base
GPT	KoGPT2	skt/kogpt2-base-v2
	GPT-3	kykim/gpt3-kor-small_based_on_gpt2

표 3. 언어 모델들을 전이학습하기 위한 하이퍼파라미터들
Table 3. Hyper-parameters for transfer-learning language models

Batch size	32(training)	64 or 1024(evaluating)
learning rate	5e-5	
Optimizer	Adam	
Epoch	2~3	
loss function	BinaryCrossentropy	

4.3 분석 및 고찰

네이버 쇼핑과 영화, 게임 스팀 리뷰에 대하여 딥러닝 모델인 RNN의 파생모델인 양방향 LSTM과 GRU, 트랜스포머 기반 BERT 계열과 GPT 계열의 한국어 처리가 가능한 체크포인트들을 선별하여 분류 정확도를 측정한 결과는 표 4와 같다.

표 4의 성능비교를 보면, 일반적으로 LSTM, GRU의 딥러닝 모델에 비해 트랜스포머 기반 모델들이 좋은 정확도를 보여주고 있다. BERT 계열에서는 LMkor-BERT와 GPT 계열의 GPT-3가 다른 모델들보다 큰 차이는 아니지만 약간 높은 정확도를 보여주었다. 이는 보다 월등히 많은 학습 데이터에 대해 훈련된 모델들이기 때문이며, 실험에 사용된 GPT-3는 아직 그 세부 사항이 공개되어 있지 않은 것이지만 BERT 계열과 다른 태스크에 적합함에도 방대한 훈련 데이터와 하이퍼파라미터들로 향후 그 성능이 기대된다.

표 4. 한국어 리뷰 감성분석 성능 비교
Table 4. Korean review sentiment analysis performance comparison

Model	Derived model (Pre-trained language model)	accuracy(val_acc) unit : %		
		NSMC	Steam	Movie
RNN	BiLSTM	91.37	77.48	85.74
	Bi-GRU	91.43	77.51	85.28
BERT	M-BERT	88.87	77.46	85.76
	KLUE-BERT	91.20	82.12	90.25
	LMkor-BERT	93.86	83.41	90.30
	LMkor-ALBERT	93.15	82.53	88.97
	KLUE-RoBERTa	93.01	82,23	89.66
	XLM-R	93.04	80.92	87.21
GPT	KoGPT2	92.01	79.87	86.90
	GPT-3	94.04	78.03	90.35

딥러닝 모델들이 NSMC 데이터를 대상으로 한 성능 비교에서는 M-BERT와 KLUE-BERT 보다 다소 높은 성능을 보여준다. 일부 데이터 셋에 대해서는 트랜스포머 기반 모델들이 각 모델 고유의 토크나이저를 사용하는 것보다 한국어 처리에 더 적합한 형태소 분석기를 사용하는 것이 좋은 성능을 기대할 수 있음을 보여준다고 할 수 있다. 따라서 한국어 처리에 적합한 형태소 분석기를 기반으로 트랜스포머 기반 모델들의 토크나이저를 보완하는 것이 성능 개선에 도움이 될 수 있을 것으로 보인다.

KoGPT2 모델은 BERT 모델들과 비교하여 낮은 정확성을 보이고 있는데, GPT는 생성형 AI로 문장

생성에 특화된 모델이기 때문인 것으로 보인다.

다국어 기반의 M-BERT와 이를 대체하는 XLM-R은 영어 외의 다양한 언어에도 활용할 수 있도록 훈련된 모델임에도 한국어가 가지는 고유한 언어 특성 때문에 다른 모델들에 비해 정확도가 다소 낮지만 딥러닝 모델들에 비해서는 좋은 성능을 보여 주었다.

그림 4는 게임 플랫폼 스팀의 리뷰에 대해 GPT-3의 잘못된 감성 분류의 예를 보여준다. 3에포크로 전이 학습시킨 결과 정확도 78%이어서 "부정" 리뷰의 경우에도 "긍정"으로 판단하는 경우들이 다수 발견되고 있다. 5에포크로 전이 학습시킨 결과 정확도가 83%로 향상되어 그림 5와 같이 판단 결과가 개선되었다.

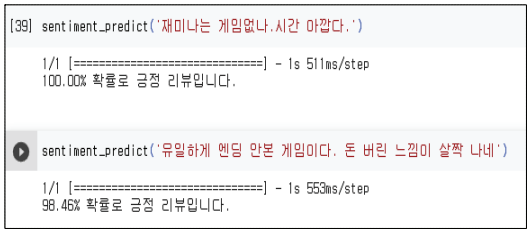


그림 4. GPT-3의 잘못된 감성분석 예
Fig. 4 An example of incorrect sentiment analysis in GPT-3

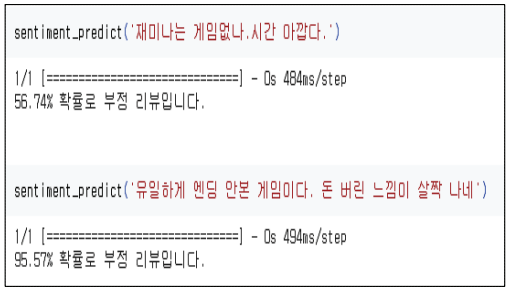


그림 5. 그림 4에 대한 수정된 분석
Fig. 5 Revised analysis for Figure 4

Ⅴ. 결론

본 연구에서는 3종의 한국어 감성 리뷰를 대상으로 다양한 언어 모델들을 대상으로 감성의 정확도를 비교하였다.

RNN 기반의 딥러닝 모델에서 보다 진화한 트랜스포머 기반 모델들이 더 좋은 성능을 보여주었다. 트랜스포머 기반 모델들 중에서도 인코더 기반의 BERT 계열이 디코더 기반의 GPT 계열보다 좀 더 좋은 정확도를 보였으며, BERT 계열 중에서 LMKor-BERT가 상대적으로 조금 높은 정확도를 보여주었다. 다국어 기반의 M-BERT와 XLM-R도 있으나 방대한 한국어 말뭉치로 사전 학습된 모델들이 한국어 감성 리뷰에서는 우수한 성능을 보여 주었다. GPT-3은 아직 세부 사항이 공개되어 있지 않았으나 GPT-2를 기반으로 한 모델도 BERT와 비교하여도 좋은 성능을 보여주어 향후 GPT-3 모델의 세부 내역이 공개되면 더 좋은 성능을 보여줄 것으로 기대된다.

감성 분석에서 보다 좋은 성능 개선을 위해서는 향후 한국어 특성을 반영한 토크나이저의 설계와 이를 적용한 한국어 고유 언어 모델의 개발이 필요하며, 사전 학습 언어 모델들을 전이 학습할 때 하이퍼파라미터를 미세 조정하는 등의 연구가 이루어져야겠다.

References

[1] J.-H., Seo, "Image Classification, Deep Learning, Convolutional Neural Network, Transfer Learning," *J. of The Korea Institute of Electronic Communication Sciences*, vol. 18, no. 3, 2023, pp. 413-420.

[2] J.-M. Jo, "Time Series Data Processing Deep Learning system for Prediction of Hospital Outpatient Number," *J. of The Korea Institute of Electronic Communication Sciences*, vol. 16 no. 2, 2021, pp. 313-318.

[3] Y.-T. Oh, M.-T. Kim, and W.-J. Kim, "Korean Movie-review Sentiment Analysis Using Parallel Stacked Bidirectional LSTM," *J. of The Korean Institute of Information Scientists and Engineers*, vol. 46, no. 1, 2019, pp. 45-49.

[4] C.-G. Lim, I. Aliyu, and R.-M. Mahmood, "LSTM Hyperparameter Optimization for an EEG-Based Efficient Emotion Classification in BCI," *J. of The Korea Institute of Electronic Communication Sciences*, vol. 14, no. 6, 2019, pp. 1171-1180.

[5] J.-W. Min, S.-H. Na, J.-H. Shin, and Y.-K. Kim, "RoBERTa for Korean Natural Language

Processing: Named Entity Recognition, Sentiment Analysis, Dependency Parsing," *Proc. of Korea Software Congress 2019*, , 2019, pp. 407-409.
[6] T.-H. Kim, D.-B. Cho, H.-Y. Lee, H.-J. Won, and S.-S. Kang, "Sentiment Analysis System by Using BERT Language Model," *Proc. of Annual Conference,* Online, Korea, 2020, pp. 975-977.
[7] M.-H. Lee, "Text sentiment analysis using deep learning and ensemble technique," Proc. of Korea Computer Congress 2021, Jeju, Korea, 2021, pp. 451-453.
[8] S.-H. Hwang, "Text sentiment Analysis Based on Transformer Models using an emotional dictionary," Proc. of Korea Computer Congress 2021, Jeju, Korea, 2021, pp. 876-878.
[9] K.-R. Lim and H.-S. Lim, "Comparative Analysis of Emotional Classification Performance by Domain between Korean Natural Language Processing Models," Proc. of Korea Computer Congress *2022*, Jeju, Korea, 2022, pp. 332-334.
[10] S. Ravichandiran, *Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT.* Birmingham: Packt Publishing, 2021.

저자 소개

이재홍(Jae-Hong Lee)

1986년 충남대학교 전자공학과 (공학사)
1988년 충남대학교 대학원 전자공학과(공학석사)
1999년 충남대학교 대학원 컴퓨터공학과(공학박사)
1988년-1994년 국방과학연구소 연구원
1994년-1995년,1999년 (주)한국인식기술 연구원
2000년-현재 전남도립대학교 보건의료과 교수
※ 관심분야 : 자연어 처리, 유비쿼터스 컴퓨팅