ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2021 June: 27(2): 1~15 http://dx.doi.org/10.13088/jiis.2021.27.2.001

BERT 기반 감성분석을 이용한 추천시스템*

박호연
동국대학교_서울 일반대학원 경영정보학과
(hoyeonpark@dgu.ac.kr)

김경재
동국대학교_서울 경영대학 경영정보학과
(kjkim@dongguk.edu)

추천시스템은 사용자의 기호를 파악하여 물품 구매 결정을 도와주는 역할을 할 뿐만 아니라, 비즈니스 전략의 관점에서도 중요한 역할을 하기에 많은 기업과 기관에서 관심을 갖고 있다. 최근에는 다양한 추천시스템 연구 중에서도 NLP와 딥러닝 등을 결합한 하이브리드 추천시스템 연구가 증가하고 있다. NLP를 이용한 감성분석은 사용자 리뷰 데이터가 증가함에 따라 2000년대 중반부터 활용되기 시작하였지만, 기계학습 기반 텍스트 분류를 통해서는 텍스트의 특성을 완전히 고려하기 어렵기 때문에 리뷰의 정보를 식별하기 어려운 단점을 갖고 있다. 본 연구에서는 기계학습의 단점을 보완하기 위하여 BERT 기반 감성분석을 활용한 추천시스템을 제안하고자 한다. 비교 모형은 Naïve-CF(collaborative filtering), SVD(singular value decomposition)-CF, MF(matrix factorization)-CF, BPR-MF(Bayesian personalized ranking matrix factorization)-CF, LSTM, CNN-LSTM, GRU(Gated Recurrent Units)를 기반으로 하는 추천 모형이며, 실제 데이터에 대한 분석 결과, BERT를 기반으로 하는 추천시스템의 성과가 가장 우수한 것으로 나타났다.

주제어 : BERT, 딥러닝, 추천시스템, 감성분석, CRM

논문접수일 : 2020년 12월 3일 논문수정일 : 2021년 3월 4일 게재확정일 : 2021년 3월 8일
원고유형 : 일반논문 교신저자 : 김경재

1. 개요

오프라인 매장에서 의사결정을 내릴 수 없을 때, 친구나 주변 사람들에게 조언을 구하여 추천을 받고 구매하였던 과거와 달리, 현재는 온라인에서 정보 검색을 통해 물품 평점 데이터를 보고 물품 구매 결정을 하는 경우가 많다. 이와 같은 과정과 결정 시간을 줄여주기 위해 추천시스템을 이용하며, 추천시스템은 물품 구매 결정과 관련하여 사용자의 기호를 파악하여 추천하기 때문에 마케팅 관점에서 학계뿐만 아니라 산업계에서도 관심이 높다 (Kim and Park, 2018). 추천시스템은 경험적 직관이 아니라 데이터에서 파생된 사용자의 선호도 정보를 사용하는 시스템이며 사용자 기반으로 맞춤형 데이터를 생성한다. 추천시스템의 방식은 명시적 평가와 암시적 평가의 두 가지 평가 방식이 있으며, 주로 점수나 평점을 사용하는 명시적 평가방식을 사용한다 (Jawaheer et al., 2010).

명시적 평가 방식의 추천시스템에서는 사용자 평점이 가장 중요한 데이터이다. 사용자 평점은 아이템 항목과 함께 사용하여 아이템 항목에 대

* 이 논문은 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2019S1A5A2A01050194)

한 관심을 나타낼 수 있으며, 주로 Likert 5점 척도 또는 '좋아요/싫어요' 처럼 표현된다 (Chelliah and Sarkar, 2017). 추천시스템의 연구는 사용자 평점과 아이템 항목만 있으면 사용자 선호도를 통해 예측할 수 있는 장점이 있을 뿐 아니라, 사용자의 의사결정시간을 단축할 수 있는 장점이 있다. 추천시스템의 수 많은 연구들은 시간이 지나면서 평점에 대한 추천 항목 표현이 잘되는 모델로 진화되었지만, 사용자 평점만 이용한 추천시스템은 개인화 콘텐츠와 지능형 콘텐츠 필터에 대한 한계점이 존재했다 (Ku and Ahn, 2018).

선행 연구에서는 추천시스템에서 사용자 평점뿐만 아니라 사용자의 리뷰 데이터와 같은 기타의 감성 정보를 활용 (Hyun et al., 2019) 함으로써 추천 성과의 정교성을 제고할 수 있을 것으로 보고하였다 (Chen, 2019). 실제로 대부분의 소비자 리뷰는 제품의 사용자 평점 뿐 아니라 평가에 사용될 수 있는 주관적인 리뷰 데이터가 포함되어 있지만, 사용자 평점과 아이템 항목만 평가 데이터로 사용되어 왔다. 그 이유는 자연어에 대한 처리가 어렵기 때문에 리뷰데이터를 이용하더라도 문맥 정보를 무시하고 키워드나 해쉬태그(hash-tag)등으로 반영하여 연산해야 했기 때문이다. 현재는 딥러닝의 발전으로 인해, NLP(natural language processing)는 신경망 기반의 seq2seq (sequence to sequence) 구조를 통해 기계 번역(machine translation)이나 대화 모델(conversation model) 등 다양한 분야에서 활용되고 있다. Seq2seq는 언어의 encoder와 decoder를 RNN(recurrent neural network)을 기반으로 하나의 입출력을 이루는 구조이다. RNN의 은닉층의 입출력이 인공신경망에 좀 더 가깝게 모델을 이루게 되었음에도 불구하고, sigmoid와 같은 활성화 함수에서 그라디언트 소멸(vanishing gradient) 문제가 있다. RNN은 sigmoid 함수에서 값이 1에 가까워질수록 그라디언트가 0으로 수렴되기 때문에 초반에 들어온 정보가 충분히 전달되지 못하는 장기의존성 문제 (problem of long-term dependencies)가 나타나기 때문이다. RNN의 장기 의존성 문제는 LSTM(long short term memory networks)의 게이트 구조를 이용하여 해결할 수 있지만 텍스트 분류, 개체명 인식 등의 문장 임베딩까지는 고려하기 어려웠으나, 2018년 사전 학습 모델이 제안된 후 seq2seq 구조가 아닌 문법적 맥락이 포함된 임베딩에 대한 문제까지 고려되고 있다.

대표적인 사전 학습 model인 BERT(bidirectional encoder representations from transformers)는 구글에서 제안된 모델로서 label이 아니라 언어 자체를 학습할 수 있으며, 또한 transformer 모델이기 때문에 RNN의 seq2seq 같이 입출력이 encoder와 decoder로 조합되어 있지만 self-attention을 통해 문장 정보에서 생성된 유의미한 관계를 조합할 수 있다.

본 연구에서는 BERT를 이용하여 소비자 리뷰의 문맥정보까지 반영할 수 있는 새로운 추천시스템을 제안하고자 한다. 문맥 정보는 추천시스템의 기존 알고리즘을 통해서 해결하기 어렵기 때문에 BERT의 감성분석을 이용하여 분류하여 추천을 진행한다. 본 연구의 순서는 다음과 같다. 2장에서는 사전 학습 모델과 텍스트 기반 추천시스템 관련 선행연구를 소개하고, 3장에서는 BERT와 추천시스템의 조합 모델이 어떻게 구현되는지를 설명한다. 4장에서는 논문의 분석 과정과 결과를 설명하며, 5장에서는 결론과 향후 연구방향을 제시한다.

2. 추천시스템 선행연구

2.1. 텍스트 기반 추천시스템

텍스트 기반 추천시스템은 대부분 온라인 쇼핑몰에서 사용자가 제공한 상품에 대한 리뷰 및 피드백을 활용한다. 주로 추천 성과에서의 사용자 리뷰의 유용성을 평가할 수 있도록 하는 것을 목표로 하며, 임베딩 모델과 알고리즘에 따라 다양한 연구가 이루어져 왔다. 대표적인 것이 BoW, TF-IDF, HF-IHU 등을 활용한 연구인데 구체적인 선행연구는 다음과 같다.

BoW (Bag of Words)는 문서 순서와 관계없이 단어 출현 빈도에 따라 텍스트를 임베딩하는데, Yin et al. (2018)은 cold-start 문제를 해결하기 위해서 개인 사용자의 관심사 및 카테고리 인식, 선호도 등에 BoW를 적용했다. 한편, TF-IDF(Term Frequency-Inverse Document Frequency)는 특정 문서 내에서 가중치를 이용하는 임베딩이다. Zangerle et al. (2013)은 해시 태그를 사용하여 추천을 하기 위해 TF-IDF의 코사인 유사성 및 Jaccard 유사성 접근법을 비교하였으며, Otsuka et al. (2014)는 해시 태그 관련성을 평가하는 동안 Twitter 데이터 세트에서 계정 데이터의 희소성을 고려하는 TF-IDF 가중치가 있는 HF-IHU (Hashtag Frequency-Inverse Hashtag Ubiquity)를 제안했다. 토픽 모델링을 이용한 모델 중 하나인 Hashtag-LDA는 사용자 벡터의 차원을 줄이고, 사용자, 단어 및 해시 태그 간의 잠재적인 관계를 찾을 수 있도록 하여 추천 알고리즘을 설계하였다 (Zhao et al., 2016). 그 외에도 LDA와 결합된 모델인 CTR (collaborative topic regression)은 종종 리뷰 등급을 모델링하는 데 사용되었으며 (Wang and Blei, 2011), 추천시스템과 PMF (probabilistic matrix factorization)를 통합한 모델로서 차원 축소에 기여하였으나 과적합의 단점이 있었다. 텍스트 리뷰에서의 과적합 단점을 보완하기 위해, 잠재 토픽들을 클러스터링 하였고 (Zhang and Whang, 2016), Qian et al. (2014)은 사용자 텍스트 리뷰에서 관심사를 추출한 후 통합하여 사람들이 참석할 만한 이벤트 추천에 이용하였다. Du et al. (2020)은 이벤트 추천의 cold-start problem을 완화하고자, 이벤트와 주최자의 상관관계를 분석한 후, 텍스트 데이터 셋의 sparsity를 줄여서 확률적 생성 모델을 구축하였으나 잠재 주제에 대한 Gibbs sampling으로 인하여 모델이 효율적이지 못했다. 이상의 대부분의 연구들은 고차원의 벡터를 줄이는 차원 축소에 NLP를 적용하여 왔으나, 본 연구에서는 데이터의 정보 손실을 최소화하기 위하여 벡터의 차원을 줄이지 않고 딥러닝을 이용하여 소비자의 관심사와 평점을 그대로 활용하여 분석을 진행하고자 한다.

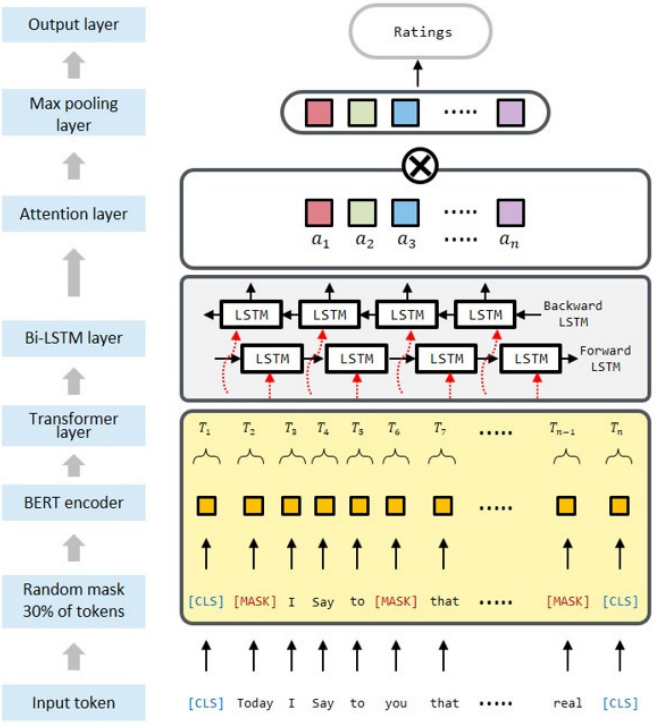
2.2. 감성분석을 이용한 추천시스템

감성분석은 텍스트에서 사용자의 감성을 추출하는 것으로 제품이나 장소에 대한 의견의 유사도를 참조할 때 유용하게 사용될 수 있다. 실제로 감성분석을 통해 마케팅 전략, 제품에 대한 피드백 등을 분석할 수 있기 때문에 최근에는 추천시스템과 감성분석을 결합하여 활용하는 연구가 진행되어 왔다.

일반적으로 감성분석은 긍정 및 부정적인 극성 단어가 포함된 어휘 사전을 통해 사용자 관점에서 텍스트를 식별하며, 딥러닝이 발전되기 전까지는 주로 나이브베이즈, SVM(support vector machine) 등을 이용하였다. Akhtar et al. (2017)은 tripadvisor에서 호텔 리뷰를 분석하여, LDA를

통해 잠재적인 주제를 도출한 후, SentiWordNet의 코퍼스를 사용하여 긍정 및 부정에 대한 극성 분석을 진행하였다. García-Cumbreras et al. (2013)은 사용자 리뷰에 대한 감성분석 레이블을 협업 필터링 알고리즘의 새로운 속성으로 추가하여 분석한 결과, 분류 정확도가 80%로 분석되었다. D' Addio et al. (2017)은 휴리스틱과 기계 학습을 접목하여 텍스트의 특징을 추출하여 태그를 레이블로 지정한 후, 근접이웃 기반의 협업필터링을 이용하여 아이템 별 유사점을 계산할 때 이용하였다.

최근에는 딥러닝 기술이 발전하면서 감성분석 기반 추천시스템이 NLP와 결합되어 사용자 리뷰를 사용자 선호도와 함께 모델링하고 있다. (Hu et al., 2020). Kumar et al. (2020)은 IMDB 데이터를 이용하여 콘텐츠 기반 필터링과 협업필터링을 결합한 하이브리드 추천시스템을 제안하였다. 이 연구에서는 사전 학습된 트윗을 이용하여 사용자 트윗에 대한 감성분석을 이용하는 방식으로 진행하였지만 데이터의 양이 많지 않았기 때문에 성능이 떨어지는 결과를 보였다. 한편, Da'u and Salim (2019)은 제품의 각 측면에 대한 사용자 감성 점수를 추정하여 가중치 등급을 산정한 후, CNN (convolution neural network)을 이용하여 사용자 감성에 대한 극성을 추출하고, Word2Vec의 CBOW (continues bag of word model) 아키텍처를 기반으로 대규모 Google 뉴스 코퍼스를 사전 학습하여 활용하였다.



〈Figure 1〉 BERT model this study

3. BERT 기반 감성분석을 활용한 추천시스템

본 논문에서는 추천시스템에 전술한 감성분석을 결합하여 활용하며 특히 감성분석의 성능을 제고하기 위하여 “사전 학습(pre-train)”과 “미세조정(fine-tuning)”이 가능한 BERT를 이용하고자 한다. <Figure 1>은 본 연구에서 사용하는 BERT의 개념도이다.

3.1. BERT를 이용한 감성분석

다른 딥러닝 모델과 달리 BERT는 입력 임베딩을 배치할 수 있는 3개의 레이어를 가지고 있다. 즉, Token embedding, Segment embedding, Position embedding으로 구성되어 있고, 그 중 token embedding layer는 기본적으로 tokenizer로 전처리된 입력 텍스트를 token이 포함된 레이어로 변환하는 역할을 한다.

BERT encoder는 양방향 언어 모델(bidirectional language model)을 사용하여 사전 학습된 모델을 조정할 수 있기 때문에 (Xie et al., 2020) 다중 클래스 문제 또한 연구자가 원하는 출력에 가깝게 표현할 수 있다 (Majumder et al., 2019 ; Zhou et al., 2019 ; Bai et al., 2020) Encoder의 단어 임베딩은 BERT에서 마스킹(masking)된 언어 모델을 사용할 수 있는데, 이는 주어진 시퀀스를 통해 전체 문장을 모델에 삽입한 후 공백에 해당하는 단어를 예측하여 학습하며, 문장을 학습한 후에는 해당 문장을 일련의 token으로 표시하고 token 배열에서 공백으로 설정할 부분은 "[MASK]"로 정의한다 (Zeng., 2020). 본 연구에서는 한 문장 token을 이용하여 30%의 마스킹을 하였다. 본 연구의 사전 학습은 사용자 리뷰와 평점을 벡터로 입력 받아 BERT 감성분석을 할 수 있도록 구현하였다. 사전 학습 모델은 Xie et al. (2020)와 Rothe et al. (2020)의 연구에 따라 BERT multilingual base model (Cased)를 활용하였다.

3.2. BERT를 활용한 추천시스템

BERT 모델의 장점은 포지션 임베딩을 이용하여 문장에서 한 쌍의 단어 관계를 정의하여 진행한다는 점이다. 한 쌍의 단어 관계를 나타내는 방법은 BERT의 메모리 관점에서 효율적인 scaled dot-product attention 방법을 주로 이용한다. Scaled dot-product attention은 $\sqrt{d_k}$를 scaling factor로 사용하여 나눈 뒤 softmax를 취한 결과로서 query (Q), key (K), value (V)를 기반으로 구성되며 식 (1)로 표기된다.

$$softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V = Attention(Q, K, V) \quad (1)$$

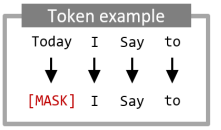
Scaled dot-product attention에서 Query와 key는 코사인 유사성을 이용한 내적(dot-product)을 제곱근으로 나눠 softmax를 적용하여 계산한다. 이 관계에서 문장 내의 단어들은 한 쌍의 행렬을 만들며 이에 대한 예시는 아래 <Figure 2>와 같다.

다음 단계는 Bi-LSTM과 attention 레이어를 진행하는 것인데 미세조정 상태에서 사전 학습은 Bi-LSTM 계층에 배치되기 때문에 출력 벡터에 대한 attention 레이어가 LSTM에 적용된 은닉층을 생성한다.

LSTM 레이어를 삽입하는 이유는 seq2seq에 대한 그라디언트 소멸 문제를 해결하고 문맥 정보를 활용하기 위해서다. 감성분석의 입력 데이터에 대해 LSTM을 훈련시키기 위해 기존 LSTM

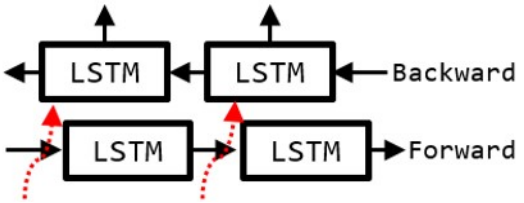
$$softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V = \begin{array}{c} \\ \text{[MASK]} \\ \text{I} \\ \text{Say} \\ \text{to} \end{array} \begin{array}{c} \begin{array}{cccc} \text{[MASK]} & \text{I} & \text{Say} & \text{to} \end{array} \\ \begin{pmatrix} 0.1 & 0.2 & 0 & 0.4 \\ 0 & 0 & 0 & 0 \\ 0.7 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{pmatrix} V_{[MASK]} \\ V_{[I]} \\ V_{[Say]} \\ V_{[to]} \end{pmatrix}$$

$$= \begin{array}{c} \\ \text{[MASK]} \\ \text{I} \\ \text{Say} \\ \text{to} \end{array} \begin{pmatrix} 0.1\,V_{[MASK]}+0.2V_{[I]}+0.4V_{[to]} \\ 0 \\ 0.7\,V_{[MASK]}+0.1V_{[to]} \\ 0 \end{pmatrix}$$



〈Figure 2〉 Scaled Dot-Product Attention Process

을 확장한 Bi-LSTM을 사용한다. LSTM 2개를 결합한 Bi-LSTM을 사용하는 이유는 양 방향으로 LSTM을 적용할 수 있기 때문이다. 또한, Bi-LSTM을 이용하게 되면 트랜스포머 계층에서 숨겨진 token을 연결하여 전체 모델을 다시 미세 조정할 수 있다 (Devlin et al., 2018). <Figure 3>은 Bi-LSTM의 일반적인 개념도이다.



〈Figure 3〉 Bi-LSTM Structure

Attention 레이어는 하나의 문장 정보를 문장 끝까지 입출력 할 수 있도록 하여 첫 단어가멀리 있는 단어와도 상관 관계를 이룰 수 있도록 고안되었다 (Bahdanau et al., 2014). 한편, Max pooling 레이어를 사용하면 attention 레이어 값을 하나로 줄일 수 있다. Max pooling 레이어는 계산량을 늘리거나 줄일 수 있는데 계산량을 줄임으로써 과적합을 방지하는 데 도움이 될 수 있다 (Vaswani et al., 2017).

Output 레이어는 softmax 활성화를 통해 전체 신경망 모델의 분류기 역할을 한다. Softmax 활성화 함수를 사용하여 분류된 문장 쌍을 이용하고, 각 평점 레이블의 확률 값에 따라 출력 값을 산출한다.

4. 실험 및 분석

4.1. 데이터 셋과 전처리

본 연구에서의 데이터 셋은 대표적인 온라인 쇼핑몰의 리뷰 데이터인 아마존 미국 사이트의 리뷰 데이터를 크롤링하여 진행하였다. Keras에서 제공하는 아마존 데이터 셋 리뷰는 2010 년부터 2015 년까지의 제품 및 리뷰 데이터로 구성되어 있기 때문에 본 연구에서 크롤링할 제품 및 리뷰 데이터는 2018년부터 수집하였다.

본 연구에서 사용하는 데이터는 음식 분야의 약 32만개의 데이터를 크롤링하여 구성하였다. 32만개의 데이터를 크롤링하였지만, 전체적으로

〈Table 1〉 Statistics of dataset

# users	# items	# ratings	Avg. # ratings / users	density ($\fallingdotseq \frac{\#ratings}{\#users \times \#items}$)
25,451	13,688	42,283	1.661	0.121 %

〈Table 2〉 Description of BERT Token

BERT token	Description
[CLS]	At the beginning of the sequence, it is indicated before the word-embedding(sentence) as a token to classify.
[UNK]	Used when the token is not in the vocabulary dictionary.
[SEP]	When the sequence ends, it is marked at the end of the word-embedding(sentence) as a token to classify.
[PAD]	Token to match the batch size, indicated when using sequences of different lengths.
[MASK]	Used to predict the masked language with a token used for masking.

평점과 부합하지 않는 리뷰들이 많이 있었다. 예를 들어, “Perfectly good. But think before you buy. (정말 좋다. 그러나 사기 전에 생각해봐라.)”의 리뷰에 평점이 1점으로 주어진 경우이다. 이 경우, BERT에서 마스크를 씌울 때 평점과 리뷰에 대한 효율적 매칭이 어렵기 때문에 약 32만개 데이터 중, sparsity를 고려하여 우선 사이트 내의 helpfulness가 5개 이하의 데이터는 제거한 후 평점과 리뷰를 매칭하였다. 매칭된 리뷰는 42,283개로 구성되었으며, 데이터 셋에 대한 상세정보는 <Table 1>과 같다.

전처리 후, 분석 결과의 성과를 확인하기 위하여 Training set 80% (33,826개)와 Test set 20% (8,457개)로 구성하였다.

4.2. 감성분석 결과

BERT를 통해 분석된 결과를 비교하기 위해서 confusion matrix의 대표적인 평가 지표인 Accuracy, Precision, Recall, F-measure를 이용하였다. BERT는 512byte의 제약이 있기 때문에 학습할 때, 리뷰 길이에 대한 설정이 필요하다. 512byte를 전부 설정하게 될 경우, 시간이 오래 걸리거나 무한 루프에 빠질 위험이 있기 때문이다. 본 논문에서 사용되는 리뷰 길이에 대한 설정은 BERT학습에서 config는 base BERT로 사용하여 90%를 128byte의 built-in 데이터 구조로 설정한 후, 10%는 512byte로 학습하도록 두었다. 이렇게 설정하면 분석 시간이 절감되면서 성능이 유지될 수 있기 때문이다. 리뷰의 길이를 설정한 다음에는 batch size를 64로, thread는 4로 설정하였다. Config, batch size, thread 과정이 완료된 후, 마스킹을 씌우는 과정을 진행하였다.

마스킹을 씌우기 위해서는 vocab이 있어야 하기 때문에 Wordpiece 알고리즘을 이용하였다. Wordpiece 알고리즘은 NLP를 이용하여 만들어진 tokenization 용도의 함수로서 일부 단어에 대한 전처리와 마스킹을 동시에 진행할 수 있다. 이 과정에서 생긴 사전 학습 모델의 단어를 token으로 적용하였으며 token이 [PAD]가 아닌 경우 0으로 처리 후 진행하였다. Token에 대한 자세한 설명은 아래 <Table 2>와 같다.

〈Table 3〉 BERT results

Measure	Accuracy (%)	F1 Score (%)	Sensitivity (%)	Specificity (%)	Precision (%)
%	90.81	90.81	91.60	90.04	90.05

〈Table 4〉 Comparative Models

Model	Description
Naïve-CF	Naïve-CF is a model that uses the most common user-based collaborative filtering (user-based CF). It does not consider review data and recommends through user ratings.
SVD-CF	SVD is an algorithm that has been considered by the most winners of the Netflix competition. SVD is a model that reduces the dimension by using the upper *n* diagonal elements (*s*), and it is possible to explain the Gaussian distribution or covariance aspect by reducing the dimension of high-dimensional data.
MF-CF	MF is a basic recommendation system technique of collaborative filtering. The MF creates a matrix according to the user's preference, but it is created as a sparse matrix.
BPR-MF-CF	BPR is a technique to obtain MF by estimating parameters using a Bayesian approach. The BPR-MF method is also referred to as the “Pairwise approach” because a pair of ranking loss plays a role in filling the empty space after optimizing the MF model.
LSTM	LSTM works by working on the encoder-decoder model. It has the advantage of improving the seq2seq gradient problem of RNN and enabling collaborative filtering and extension easily.
CNN-LSTM	CNN facilitates extended representation of data, and LSTM is excellent for natural language processing. CNN and LSTM are combined and called CNN-LSTM. CNN-LSTM has a high prediction rate, so it is widely used in other deep learning models.
GRU	GRU is a structure created through the transformation of LSTM and LSTM to solve the long-term dependency problem of RNN. LSTM generally has 3 gates, but RNN has 2 gates because there is no output gate. Despite having two gates, the reason GRU is used is because it can learn less data than LSTM in a short time.

마스킹 처리 후, 감성분석과 BERT 모델을 결합하기 위해서 파이토치의 tensor를 이용하였으며, 레이블은 감성 레이블이 아니라 아마존의 평점을 이용하여 진행하였다. 파이토치는 torchvision package에서 사전 학습된 모델을 통해 이용할 수 있으며, tensor는 파이토치의 자료형으로서 다차원 행렬을 단일 행렬로 만들 수 있는 자료구조이다. 그 다음, 감성분석의 가중치에 패널티를 부여하기 위해서 Adam (adaptive moment estimation) 함수를 이용하였다. Adam 함수는 크기에 따라 최적화를 적용하여 목적함수에 대한 최소값을 찾아주는 함수로서 딥러닝의 정확도 개선하는데 많이 이용된다. Adam 함수를 이용한 후 5000번의 반복 시행을 거쳐 생성된 데이터셋에 로지스틱 회귀분석을 적용하여 <Table 3>의 결과가 산출되었다.

4.3. 추천 성능 평가

상품 추천을 위해 사용되는 알고리즘의 성능을 평가하기 위해 잘 알려진 평가지표인 RMSE

〈Table 5〉 Experimental results of comparative models

Model	RMSE	MAE
Naïve-CF	1.909315	1.479220
SVD-CF	1.125820	0.915790
MF-CF	1.109006	0.740954
BPR-MF-CF	0.901572	0.621709
LSTM	1.070171	0.705493
CNN-LSTM	0.958817	0.644363
GRU	0.764277	0.344468
BERT	0.751685	0.326222

(Root Mean Squared Error)와 MAE (Mean Absolute Error) 를 이용하였다. 본 연구에서 제안하는 BERT와 추천시스템이 결합된 기존 추천시스템 과의 성과 비교를 위하여 기존 추천시스템 알고리즘인 Naïve-CF(collaborative filtering), SVD (singular value decomposition)-CF, MF(matrix factorization)-CF, BPR-MF(Bayesian personalized ranking matrix factorization)-CF를 선택하였다. 딥러닝 모델은 감성분석 선행연구에서 활용되어 온 LSTM, CNN-LSTM, GRU(Gated Recurrent Units) 알고리즘을 채택하였다. 딥러닝 모수는 공통적으로 1e-4, 125 unit, batch size 32, drop out 0.5, epoch 100으로 진행하였다. LSTM은 Bi-LSTM layer를 2개로 설계하였으며, CNN-LSTM은 convolution layer와 max-pooling layer를 4개씩, fully connected layer 2개, Bi-LSTM layer를 2개로 구성하였으며, GRU는 window size 24, hidden layer를 3개로 구성하였다. 각 비교 모형에 대한 구체적인 설명은 다음 <Table 4>와 같다.

본 논문에서는 추천시스템의 모형들을 비교하기 위해서 이상의 7가지 모형을 BERT와 비교하였다. 분석 결과, BERT가 RMSE와 MAE 측면에서 가장 우수한 추천 성과를 보여 주었다. 분석 결과는 <Table 5>와 같다.

Naïve-CF의 경우, 사용자 평점 만을 고려하는 모형이기 때문에 정보의 양이 상대적으로 부족하기에 RMSE, MAE 결과가 가장 크게 나온 것으로 추정된다. SVD-CF는 사용자 리뷰를 반영할 때, 관련된 아이템과 사용자에 대한 임베딩을 통해 리뷰 내의 용어와 내적 계산으로 진행될 수 있다. 그러나 이 연구의 SVD-CF 분석 결과는 MF-CF 실험결과보다 RMSE, MAE가 낮게 측정되었다. LSTM는 딥러닝 기반 기법 임에도 불구하고, RMSE가 1점이 넘게 나왔으며, MF의 실험결과와 비슷하게 측정되었다. 결합 모델도 LSTM 와 유사하게 BPR-MF가 CNN-LSTM보다 RMSE 결과가 높게 측정되었다. 본 연구에서 제안한 모델인 BERT는 다른 딥러닝 기반 모델보다 성과가 향상되었으나 그 차이는 매우 크지 않았다. BERT모델은 GRU를 제외한 다른 모델들보다는 성능이 많이 개선되는 것으로 나타났다.

5. 결론

본 연구에서는 BERT 기반의 감성분석과 추천모델을 통한 추천시스템을 제안하였으며 비교모형들에 비해 성능이 우수한 것을 확인하였다. 제

안 모형에서 우수한 추천 성능을 보인 이유는 전통적인 협업필터링 기반의 추천시스템에서 고려하지 않았던 사용자 리뷰 데이터의 정보를 고려하였고, 감성분석 과정에서도 여러 가지 장점을 가진 BERT 모형을 활용하였기 때문인 것으로 판단된다. 제안 모형과 비교 모형을 전통적인 CF를 활용하는 모형들 (Naïve-CF, SVD-CF, MF-CF, BPR-MF-CF)과 딥러닝 기반의 모형들 (LSTM, CNN-LSTM, GRU, BERT)로 분류하여 비교하면 딥러닝 기반의 모형들이 전반적으로 CF를 활용하는 모형들에 비해 추천 성능이 우수한 것으로 나타났다. 이는 딥러닝이 가진 예측의 정교성 등에 기반한 것으로 판단된다.

본 연구에서 제안하는 모형의 실무적인 의의는 전통적인 추천시스템에 사용자 리뷰와 같은 정성 정보를 반영하기 위해서는 이를 별도로 분석하여 고려하거나 별도의 감성분석을 진행한 후 고려하여야 하였으나 제안 모형은 BERT를 기반으로 감성분석과 추천을 동시에 진행할 수 있으므로 분석과 추천과정이 단순화된다는 것이다.

한편, 연구 결과를 통해 BERT 기반의 추천시스템이 추천 결과에서 우수한 성능을 보임을 확인하였으나 BERT의 연산과정이 난해하여 이에 대한 이해와 모델링 과정에서 상당한 시간이 소요됨은 제안 모형의 한계점이라고 할 수 있을 것이다. 딥러닝의 특성 상 결과를 해석하는 설명력에서의 한계점도 있다고 생각된다. 그리고 본 연구는 추천시스템의 정확도 제고를 목표로 설계되었기에 콜드스타트 문제를 해결할 수 있는 방안은 고려하지 못하였다는 점은 또 하나의 한계점이라고 할 수 있다. 또, 추천의 성능 평가에 있어서 추천시스템에서 많이 활용되고 있는 Top-N까지 고려하지 못한 점도 한계점으로 생각된다.

향후 연구에서는 향후 연구에서는 가짜 뉴스와 같은 문장 생성이 가능한 GPT(Generative Pre-trained Transformer)와의 성능 비교도 이루어져야 할 것이다. GPT의 초기 모델인 GPT-1은 BERT 모델과 함께 소개되었으나 (Devlin et al., 2018), attention 레이어가 GPT는 단방향인 반면에 BERT는 양방향이었기에 단순문장 생성분야 외에는 모든 NLP작업에서 BERT가 우수한 것으로 알려져 있었다. 그러나 최근에 개선된 GPT-3는 few-shot learning, zero shot learning, one shot learning 등의 학습 모델을 재구성하여 사전 훈련 없이 바로 훈련할 수 있는 장점 때문에 BERT의 사전 훈련 모델과 비교되고 있다. 따라서 향후에는 GPT와 BERT의 추천 성능 비교연구가 이루어져야 할 것이다. 또, 본 연구에서는 아마존 데이터셋만을 활용하여 검증하였으나 다른 특성을 가진 데이터셋에서는 현재와 다른 성과가 나올 수도 있을 것이다. 향후 연구에서는 본 연구 결과의 일반화를 위해 다양한 특성을 가진 데이터셋에서의 성능 평가가 이어져야 할 것이다.

참고문헌(References)

Akhtar, N., N. Zubair, A. Kumar and T. Ahmad, "Aspect Based Sentiment Oriented Summarization of Hotel Reviews", *Procedia computer science,* Vol.115, (2017), 563~71.

Bahdanau, D., K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate", *arXiv preprint arXiv: 1409.0473*, (2014)

Bai, P., Y. Xia and Y. Xia "Fusing Knowledge and Aspect Sentiment for Explainable Recommendation", IEEE Access Vol.8, (2020),

137150~137160.

Chelliah, M. and S. Sarkar, *Product Recommendations Enhanced with Reviews*. Proceedings of the Eleventh ACM Conference on Recommender Systems, 2017.

Chen, R.-C., "User Rating Classification Via Deep Belief Network Learning and Sentiment Analysis", *IEEE Transactions on Computational Social Systems,* Vol.6, No.3, (2019), 535~46.

D'Addio, R. M., M. A. Domingues and M. G. Manzato, "Exploiting Feature Extraction Techniques on Users' Reviews for Movies Recommendation", Journal of the Brazilian Computer Society, Vol.23, No.1, (2017), 7.

Da'u, A. and N. Salim, "Sentiment-aware deep recommender system with neural attention networks", IEEE Access, Vol.7(2019), 45472~45484.

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "Bert: pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv preprint arXiv:1810.04805*, (2018)

Du, Y. L., X. W. Meng and Y. J. Zhang, "Cvtm: A Content-Venue-Aware Topic Model for Group Event Recommendation", *IEEE Transactions on Knowledge and Data Engineering,* Vol.32, No.7, (Jul 1 2020), 1290~1303. <Go to ISI>://WOS:000543006000005

García-Cumbreras, M. Á., A. Montejo-Ráez and M. C. Díaz-Galiano, "Pessimists and Optimists: Improving Collaborative Filtering through Sentiment Analysis", *Expert Systems with Applications,* Vol.40, No.17, (2013), 6758~6765.

Hu, S., A. Kumar, F. Al-Turjman, S. Gupta and S. Seth, "Reviewer Credibility and Sentiment Analysis Based User Profile Modelling for Online Product Recommendation", *IEEE Access,* Vol.8, (2020), 26172~26189.

Hyun, J., S. Ryu and S.-Y. T. 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, (2019), 219~239.

Jawaheer, G., M. Szomszor and P. Kostkova, *Comparison of Implicit and Explicit Feedback from an Online Music Recommendation Service*. proceedings of the 1st international workshop on information heterogeneity and fusion in recommender systems,, (2010).

Kim, K.-W. and D.-H. Park, "Individual Thinking Style Leads Its Emotional Perception: Development of Web-Style Design Evaluation Model and Recommendation Algorithm Depending on Consumer Regulatory Focus", *Journal of Intelligence and Information Systems,* Vol.24, No.4, (2018), 171~196.

Ku, M. J. and H. Ahn "A Hybrid Recommender System Based on Collaborative Filtering with Selective Use of Overall and Multicriteria Ratings", *Journal of Intelligence and Information Systems,* Vol.24, No.2, (2018), 85~109.

Kumar, S., K. De and P. P. Roy "Movie Recommendation System Using Sentiment Analysis from Microblogging Data", *IEEE Transactions on Computational Social Systems,* (2020)

Majumder, N., S. Poria, H. Peng, N. Chhaya, E. Cambria and A. Gelbukh, "Sentiment and Sarcasm Classification with Multitask Learning", *IEEE Intelligent Systems,* Vol.34, No.3, (2019), 38~43.

Otsuka, E., S. A. Wallace and D., *Design and Evaluation of a Twitter Hashtag Recommendation*

System. Proceedings of the 18th International Database Engineering & Applications Symposium, 2014.

Qian, X. M., H. Feng, G. S. Zhao and T. Mei, "Personalized Recommendation Combining User Interest and Social Circle", *IEEE Transactions on Knowledge and Data Engineering* , Vol.26, No.7, (Jul 2014), 1763~77. <Go to ISI>://WOS:000340205700017.

Rothe, S., S. Narayan and A. Severyn, "Leveraging pre-trained Checkpoints for Sequence Generation Tasks", *Transactions of the Association for Computational Linguistics,* Vol.8, (2020), 264~280.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Attention Is All You Need.* Advances in neural information processing systems, 2017

Xie, Q., X. Zhang, Y. Ding and M. Song, "Monolingual and Multilingual Topic Analysis Using Lda and Bert Embeddings", *Journal of Informetrics,* Vol.14, No.3, (2020), 101055.

Xie, R., C. Ling, Y. Wang, R. Wang, F. Xia and L. Lin "Deep Feedback Network for Recommendation", *Proceedings of IJCAI-PRICAI*, (2020).

Yin, H., W. Wang, L. Chen, X. Du, Q. V. H. Nguyen and Z. Huang, "Mobi-Sage-Rs: A Sparse Additive Generative Model-Based Mobile Application Recommender System", *Knowledge-Based Systems,* Vol.157, (2018), 68~80.

Zangerle, E., W. Gassler and G. Specht, "On the Impact of Text Similarity Functions on Hashtag Recommendations in Microblogging Environments", *Social network analysis and mining* , Vol.3, No.4, (2013), 889~898.

Zeng, Z., C. Xiao, Y. Yao, R. Xie, Z. Liu, F. Lin, L. Lin and M. Sun, "Knowledge Transfer Via pre-training for Recommendation: A Review and Prospect", *arXiv preprint arXiv:2009.09226*, (2020)

Zhang, W. and J. Wang, "Integrating Topic and Latent Factors for Scalable Personalized Review-Based Rating Prediction", *IEEE Transactions on Knowledge and Data Engineering,* Vol.28, No.11, (2016), 3013~3027.

Zhao, F., Y. Zhu, H. Jin and L. T. Yang, "A Personalized Hashtag Recommendation Approach Using Lda-Based Topic Model in Microblog Environment", *Future Generation Computer Systems,* Vol.65, (2016), 196~206.

Zhou, Y., X. Wang, M. Zhang, J. Zhu, R. Zheng and Q. Wu, "Mpce: A Maximum Probability Based Cross Entropy Loss Function for Neural Network Classification", *IEEE Access,* Vol.7, (2019), 146331~146341.

Abstract

Recommender system using BERT sentiment analysis

Ho-yeon Park* · Kyoung-jae Kim**

If it is difficult for us to make decisions, we ask for advice from friends or people around us. When we decide to buy products online, we read anonymous reviews and buy them. With the advent of the Data-driven era, IT technology's development is spilling out many data from individuals to objects. Companies or individuals have accumulated, processed, and analyzed such a large amount of data that they can now make decisions or execute directly using data that used to depend on experts. Nowadays, the recommender system plays a vital role in determining the user's preferences to purchase goods and uses a recommender system to induce clicks on web services (Facebook, Amazon, Netflix, Youtube). For example, Youtube's recommender system, which is used by 1 billion people worldwide every month, includes videos that users like, "like" and videos they watched. Recommended system research is deeply linked to practical business. Therefore, many researchers are interested in building better solutions. Recommender systems use the information obtained from their users to generate recommendations because the development of the provided recommender systems requires information on items that are likely to be preferred by the user. We began to trust patterns and rules derived from data rather than empirical intuition through the recommender systems. The capacity and development of data have led machine learning to develop deep learning. However, such recommender systems are not all solutions. Proceeding with the recommender systems, there should be no scarcity in all data and a sufficient amount. Also, it requires detailed information about the individual. The recommender systems work correctly when these conditions operate. The recommender systems become a complex problem for both consumers and sellers when the interaction log is insufficient. Because the seller's perspective needs to make recommendations at a personal level to the consumer and receive appropriate recommendations with reliable data from the consumer's perspective.

In this paper, to improve the accuracy problem for "appropriate recommendation" to consumers, the recommender systems are proposed in combination with context-based deep learning. This research is to

* Dept. of MIS, Graduate School, Dongguk University_Seoul
** Corresponding author: Kyoung-jae Kim
Dept. of MIS, Dongguk University_Seoul
30, Pildong-ro 1-gil, Chung-gu, Seoul, 04620, Republic of Korea
Tel: +82-2-2260-3324, E-mail: kjkim@dongguk.edu

combine user-based data to create hybrid Recommender Systems. The hybrid approach developed is not a collaborative type of Recommender Systems, but a collaborative extension that integrates user data with deep learning. Customer review data were used for the data set. Consumers buy products in online shopping malls and then evaluate product reviews. Rating reviews are based on reviews from buyers who have already purchased, giving users confidence before purchasing the product. However, the recommendation system mainly uses scores or ratings rather than reviews to suggest items purchased by many users. In fact, consumer reviews include product opinions and user sentiment that will be spent on evaluation. By incorporating these parts into the study, this paper aims to improve the recommendation system. This study is an algorithm used when individuals have difficulty in selecting an item. Consumer reviews and record patterns made it possible to rely on recommendations appropriately. The algorithm implements a recommendation system through collaborative filtering. This study's predictive accuracy is measured by Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Netflix is strategically using the referral system in its programs through competitions that reduce RMSE every year, making fair use of predictive accuracy. Research on hybrid recommender systems combining the NLP approach for personalization recommender systems, deep learning base, etc. has been increasing.

Among NLP studies, sentiment analysis began to take shape in the mid-2000s as user review data increased. Sentiment analysis is a text classification task based on machine learning. The machine learning-based sentiment analysis has a disadvantage in that it is difficult to identify the review's information expression because it is challenging to consider the text's characteristics. In this study, we propose a deep learning recommender system that utilizes BERT's sentiment analysis by minimizing the disadvantages of machine learning. This study offers a deep learning recommender system that uses BERT's sentiment analysis by reducing the disadvantages of machine learning. The comparison model was performed through a recommender system based on Naïve-CF(collaborative filtering), SVD(singular value decomposition)-CF, MF(matrix factorization)-CF, BPR-MF(Bayesian personalized ranking matrix factorization)-CF, LSTM, CNN-LSTM, GRU(Gated Recurrent Units). As a result of the experiment, the recommender system based on BERT was the best.

Key Words : BERT, deep learning, recommender system, sentiment analysis, CRM

Received : December 3, 2020 Revised : March 4, 2021 Accepted : March 8, 2021
Corresponding Author : Kyoung-jae Kim

저 자 소 개

박 호 연

동국대학교에서 컴퓨터공학을 전공하여 공학사, 경영정보학을 전공하여 경영학 석사 및 경영학 박사를 취득하였다. 주요 관심분야는 딥러닝, 빅데이터, 비즈니스 애널리틱스, 텍스트마이닝 등이다.

김 경 재

현재 동국대학교 경영대학 경영정보학과 교수로 재직 중이다. KAIST에서 경영정보시스템을 전공으로 박사학위를 취득하였으며, 연구 관심분야는 비즈니스 애널리틱스, CRM, 추천기술, 빅데이터 분석 등이다.