Journal of KIIT. Vol. 20, No. 1, pp. 149-157, Jan. 31, 2022. pISSN 1598-8619, eISSN 2093-7571 149
http://dx.doi.org/10.14801/jkiit.2022.20.1.149

BERT를 활용한 AR-KNU+ 감성사전 기반의 대학 평판도 평가

정주현*[1], 김장원*[2], 온병원*[3], 정동원*[4]

University Reputation Evaluation based on AR-KNU+ Sentiment Dictionary using BERT

Juhyeon Jeong*[1], Jangwon Gim*[2], Byung-Won On*[3], and Dongwon Jeong*[4]

이 연구는 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임
(NRF-2019R1I1A3A01060826)

요 약

설문조사, 델파이, 데이터 마이닝 기법 등을 적용하여 대학 평판도를 평가하는 다양한 연구가 진행되었다. 그렇지만 지금까지의 대학 평판도 평가는 평가 항목, 평가 방법 그리고 사용하는 데이터의 기간과 그 규모에 따라 대학 순위가 변동될 수 있기 때문에 평판도 평가의 정확도 및 신뢰도에 대한 한계점들이 존재한다. 따라서 이 논문에서는 AK-KNU+ 감성사전과 함께 BERT 모델을 활용하여 대학 뉴스 기사에 포함된 감성어휘 기반 대학 평판도 평가 방법을 제안한다. 제안 방법을 통해 대학 평판도를 표현하는 감성 어휘를 정확하게 분류할 수 있으며, 분류된 감성 어휘를 기반으로 보다 객관적으로 대학 평판도 평가를 수행할 수 있음을 보인다.

Abstract

Various studies have been conducted to evaluate university reputation by applying surveys, Delphi, and data mining techniques. However, there are limitations to the accuracy and reliability of the reputation evaluation because university rankings can change depending on the evaluation items, evaluation methods, and the period and scale of the data used. Therefore, in this paper, to solve this limitation, we propose a method for evaluating university reputation based on sentiment vocabulary included in university news articles by using the BERT model and the AK-KNU+ sentiment dictionary. Through the proposed method, it is possible to accurately classify the sentiment words expressing the university reputation. It is shown that the university reputation evaluation can be performed more objectively based on the classified sentiment vocabulary.

Keywords
university reputation evaluation, sentiment vocabulary classification, BERT

* 군산대학교 소프트웨어융합공학과 (*[2,4] 교신저자)
- ORCID[1]: https://orcid.org/0000-0003-4704-6279
- ORCID[2]: https://orcid.org/0000-0002-4480-7944
- ORCID[3]: https://orcid.org/0000-0001-6929-3188
- ORCID[4]: https://orcid.org/0000-0001-9881-5336

· Received: Sep. 29, 2021, Revised: Jan. 11, 2022, Accepted: Jan. 14 2022
· Corresponding Author: Jangwon Gim and Dongwon Jeong
Dept. of Software Convergence Engineering, Kunsan National University,
558, Deahak-ro, Gunsan, Jeollabuk-do, Korea,
Tel.: +82-63-469-8911, Email: {jwgim, djeong}@kunsan.ac.kr

Ⅰ. 서 론

매년 국내 대학 평판도 순위가 각종 언론사를 통해 발표되고 있다. 대학 평판도를 평가하는 언론사 중 대표적으로 조선일보 평가, 중앙일보 평가가 있다[1,2]. 조선일보 평가는 대학의 연구 수준에 가장 중점을 두어 대학 평판도를 평가하고, 중앙일보 평가는 학생들의 사회 진출도에 비중을 많이 두어 대학 평판도를 평가한다. 이처럼 언론사들은 대학의 발전, 변화 등을 알리기 위해 대학 평판도 평가를 시행하고 있지만, 언론사마다 대학 평판도를 측정하는 기준은 천차만별이다[3]. 언론사마다 다른 대학 평판도 순위는 대학과 입시를 준비하는 수험생 및 학부모와 취업을 준비하는 졸업예정자는 대학 평판도에 민감한 반응을 보일 수 있다. 따라서 더욱더 객관적이고 신뢰성 있는 대학 평판도를 측정하고 평가하는 연구가 필요하다.

따라서 대학 평판도를 측정하기 위해 대표적으로 설문조사 기반의 평가 방법과 자연어 처리(Natural language processing)기반의 평가 방법의 연구가 진행되고 있다. 설문조사 방식의 평가 방법은 대표적 사례로 중앙일보 평가가 있다. 중앙일보는 매년 평가 기준을 조금씩 발전시켜 대학 평판도를 평가하고 있다[2]. 하지만 설문조사 방식의 평가 방법은 두 가지 한계점을 지닌다. 첫 번째는 응답률이다. 설문의 응답자가 관심이 없거나 대학의 대한 정보가 없는 사람이라면 성실히 응답하지 않을 확률이 높아 전체의 특성을 파악하기 어렵다. 두 번째는 설문을 종합하고 분석하는 데 높은 비용이 필요하게 된다.

이러한 설문조사 방법의 한계점을 해결하기 위해 자연어 처리 기반의 연구가 진행되고 있다. 그중에서도 온라인상에 실시간으로 생성되는 데이터를 통해 사람의 감성, 태도, 평가, 의견 등을 분석하는 감성 분석을 통한 대학 평판도 연구가 진행되고 있다. 특히 기존 설문조사 방법의 대학 평판도 평가의 문제점을 해결하기 위해 감성 분석 사전을 활용한다[4][5]. [4]는 대학 평판도를 평가하기 위해 1년 동안의 온라인에 게시된 뉴스 기사를 기반으로 AR-KNU 감성 사전(Academic Reputation-KNU sentiment dictionary)을 구축하고 대학 평판도 알고리즘을 제안하였다. 군집화 알고리즘을 사용하여 주제를 도출하고 각 대학의 기사를 주제별로 분류하였다. 그리고 감성 사전 기반의 긍정 및 부정 양상을 도출하여 대학 평판도 평가를 하였다. 그러나 사용한 각 대학의 데이터양이 큰 편차를 보였고 이슈가 한쪽으로 편향되는 문제점이 발생할 수 있어 부정확한 결과를 보였다.

[5]는 이전의 특정 기간에 한정된 데이터만을 사용하여 구축된 감성사전을 확장한 AR-KNU+를 구축하였다. 또한 대학의 뉴스 기사를 3가지 클래스로 분류하여 대학 평판도 순위를 측정하였다. 이를 위해 온라인 기사들에서 추출된 단어를 100차원으로 임베딩 하고 가우시안 기반의 기댓값 최대화(EM, Expectation-Maximization) 알고리즘을 사용하여 대학 뉴스 기사의 주요 주제 3가지를 도출 및 대학 평판도 순위를 측정하였다. 그렇지만 단순히 기존 모델과 비교하여 입력 데이터의 크기만 늘렸기 때문에 분류 모델의 학습 정확도에 대한 한계점이 존재한다[6].

본 논문에서는 대학 평판도 알고리즘의 정확도와 신뢰성 향상을 위해 있는 BERT(Bidirectional Encoder Representations from Transformers)를 활용하여 감성 어휘에 대한 주제 분류를 학습하고 이를 기반으로 대학 평판도를 평가하는 방법을 제안한다. 감성 어휘에 대한 학습 정확도를 높이기 위해 기존에 공개된 AR-KNU+ 감성 사전(Academic Reputation -KNU sentiment lexicon+)을 활용하여 대학 평판도의 순위를 도출하며 이때 기존의 평판도 평가 방법과 제안 방법을 비교 평가하여 대학 평판도 평가에서 제안 방법의 우수성을 보인다.

이 논문의 구성은 다음과 같다. 2장은 관련 연구 및 제안 연구의 필요성에 관해 기술한다. 3장은 제안 방법을 설명하고 4장에서 기존 연구와의 실험을 통해 결과를 비교 평가한다. 마지막으로 결론과 한계점에 관하여 5장에서 기술한다.

Ⅱ. 관련 연구

이 장에서는 분류 모델에서 가장 우수한 성능을 보이는 BERT 모델을 설명하고 대학 평판도 분석에 관한 기존 연구의 한계점을 기술하여 제안 방법의 필요성을 기술한다.

2.1 BERT

BERT는 대용량 코퍼스를 이용하여 학습하고 이를 다양한 도메인의 문제해결에 적용할 수 있는 Transformer 기반의 언어 모델이다[7]. BRET는 Self-Attention 개념을 활용한 transformer 모델을 적용하여 대용량 코퍼스에 대한 사전학습 후 미세조정을 통해 자연어처리에서 해결하고자 하는 다양한 분야의 문제(기계 번역, 분류, 개체 인식 등)를 해결하는데 널리 적용되고 있다[8]. BERT의 장점은 양방향성을 지닌다는 점이다. 기존 언어 모델은 다음 단어의 예측 확률을 계산하는 모델이지만 BERT는 양방향으로 예측할 수 있도록 문장 내에 단어 일부를 마스킹하고 마스킹 된 단어를 예측하게 한다. 그리고 두 문장이 주어졌을 때 문맥상 첫 번째 문장이 두 번째 문장 바로 뒤에 위치하는지 판별하는 문제를 통해 문장 간 의미 관계 즉, 양방향성을 파악할 수 있게 된다.

2.2 감성 분석 기반의 대학 평판도 평가

[4]는 대학 평판도를 평가를 위해 1년 동안 게시된 온라인 뉴스 기사를 기반으로 AR-KNU 감성 사전을 구축하고 대학 평판도를 측정하는 알고리즘을 제안하였다. 대표적인 군집화 알고리즘인 EM 군집화 알고리즘으로 상위 3개 대학의 대표 주제를 도출하고 상위 3개를 제외한 나머지 대학의 기사를 다중분류 신경망을 사용하여 주제별로 분류하였다. 분류된 주제에 대하여 감성 사전에 표현된 긍정·부정 양상을 도출하여 대학 평판도 평가를 수행하였다. [4]에서 제안된 평가 방법과 기존의 설문 조사 방법들을 적용한 대학 평판도 평가 연구들과의 비교 평가를 수행하여 감성 분석이 대학 평판 측정에 유의함을 보이고 있다. 그렇지만 실험에 사용된 뉴스 기사 데이터는 1년에 한정된 것으로 해당 1년 동안에 공개된 각 대학의 뉴스 데이터양이 큰 편차를 보였으며 주제 어휘가 특정 분야로 편향되는 문제점이 발생하여 대학 평판도 측정 결과 또한 정확하지 않을 수 있음을 보였다.

[5]는 [4]에서 보였던 실험 데이터의 한계점을 극복하기 위해 수집 기간을 확장하여 데이터를 수집하여 신규 감성 어휘를 반영한 AR-KNU+ 감성 사전을 구축하였다. 확장된 감성 어휘 사전을 통해 대학 평판도의 정확도가 향상됨을 보이고 대학 평판도의 추이를 분석하였다. 이를 위해 3년 동안의 데이터 사용하여 단어 임베딩을 통해 100차원으로 전처리 후 가우시안 기반의 EM 알고리즘을 사용하여 주요 주제 3가지를 도출하였다. 그리고 다중 클래스 분류로 대학 평판도 순위를 측정하였다. 마지막으로 연도별 데이터 크기 및 실험 방법에 따른 종합적인 평판도를 비교하고 3년 동안 나타난 시계열 데이터를 이용하여 대학 평판도의 추이 변화를 분석하였다. 그 결과 대학 평판도 추이 분석 및 감성 어휘에 대한 분류 정확도 향상에는 유의미한 결과를 도출하였지만 F1값(0.89)이 다소 낮은 결과를 보인다.

이 논문에서는 감성 분석 기반의 대학 평판도 평가의 신뢰도 및 정확도를 높이기 위해 BERT 모델을 활용하여 대학 평판도 실험을 수행한다. 또한, 기존의 감성 사전을 활용하여 대학 뉴스 기사의 감성 문장을 도출한다. 그리고 대학 뉴스 기사의 주제를 도출하고 BERT의 사전 학습과 파인 튜닝을 통해 각 대학의 주제별 분류를 수행한다. 마지막으로, 기존 연구와의 분류 정확도 비교 평가를 수행하고 각 대학의 감성 문장을 사용해 종합적인 평판도와 주제별 평판도 순위를 도출한다.

III. 제안 방안

이 장에서는 기존 논문의 정확도 향상 및 대학 평판도 순위 도출을 위해 제안 방안을 기술한다. 전체적인 제안 시스템의 프로세스는 그림 1과 같다. 그림 1은 대학 평판도 개선을 위한 데이터 수집, 감성 사전 기반의 감성 어휘 분류, BERT 사전학습 및 파인 튜닝, 각 대학의 주제 도출 및 주제에 따른 대학 평판도 순위 도출로 구성된다.

첫 번째로, 뉴스 데이터는 10년 동안의 데이터를 수집한다. 기존 연구와 동일한 데이터로 실험을 하기 위해 3년 동안의 대학 뉴스 데이터와 BERT 사전학습을 위해 7년 동안의 대학 뉴스 기사 데이터를 수집한다.

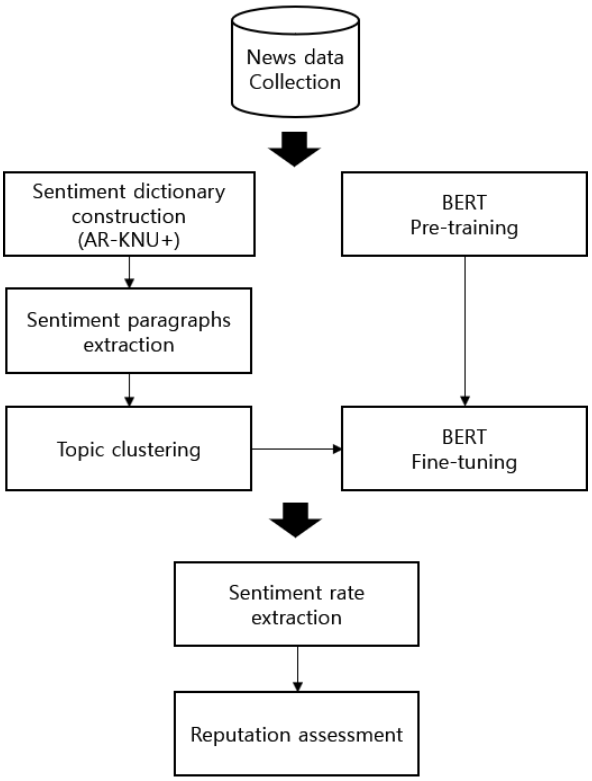

그림 1. 대학 평판도 평가의 전체적인 프로세스
Fig. 1. Overall process of university reputation assessment

그리고 3년 동안의 데이터는 대학 데이터에 적합한 AR-KNU+ 감성 사전 기반의 각 대학 뉴스 기사에 포함된 감성 문장의 점수를 파악하고 7년 동안의 데이터는 BERT에 사전학습을 수행한다. 두 번째로, 감성 문장의 개수가 가장 많은 상위 k개의 대학을 주제 추출을 위해 EM 알고리즘을 수행 후 도출된 군집을 대학 주제로 라벨링을 수행한다. 세 번째로, 상위 k개 대학을 제외한 나머지 대학의 주제 분류를 위해 군집에 포함된 문장을 학습 데이터로 군집의 레이블을 정답 데이터로 BERT에 파인 튜닝 수행 후 분류를 수행한다. 마지막으로, 군집에 포함된 각 대학의 평판을 긍정·부정 성향 파악하기 위해 감성 문장을 이용하여 긍정·부정 비율을 측정한다. 그리고 모든 대학의 종합적인 평판도 및 주제별 평판도 순위 측정을 위해 평판도 점수를 도출한다.

3.1 대학별 감성 분석

이 절에서는 대학의 평판도 평가를 위해 뉴스 기사 데이터를 수집하여 기존 논문에서 구축한 AR-KNU+ 감성 사전을 기반으로 각 대학의 문장에 감성 점수를 부여한다. AR-KNU+ 감성 사전은 3년 동안의 온라인 뉴스 기사를 기반으로 구축한 감성 사전이다. 형태소 분석을 수행하여 명사, 동사, 형용사, 부사, 감탄사로 품사별 감성 어휘에 대하여 표 1과 같은 5점 척도로 감성 점수를 측정한다.

표 1. 감성 어휘의 감성 정도
Table 1. Sentiment degree of sentiment words

Very neg	Neg	Neu	Pos	Very pos
-2	-1	0	1	2

AR-KNU+ 감성 사전을 이용하여 각 대학의 문장에 대해 감성 점수를 부여한다. 대학에 따른 문장의 감성 어휘의 합이 음수일 경우 부정적인 문장, 양수일 경우 긍정적인 문장으로 분류한다.

3.2 대학 주제 추출

이 절에서는 대학의 공통 주제 도출을 위한 방안을 기술한다. 대학별 감성 분석의 결과에서 나온 감성 문장의 개수가 많은 k개의 대학을 추출하여 기존 연구에서 사용한 가우시안 혼합 모델기반의 EM 알고리즘을 수행한다[5]. 그리고 클러스터링과 클러스터링 합병을 통해 상위 k개의 대학의 문장에 대한 공통 주제 3가지를 추출 후 라벨링을 진행한다.

3.3 BERT Pre-training

이 절에서는 BERT 모델을 이용하여 사전학습 방안을 기술한다. 먼저 사전학습을 위해 10년 동안의 뉴스 기사 데이터 중 7년 동안의 데이터를 이용하여 사전학습을 수행한다. 사전학습 방식으로는 그림 2와 같은 입력 문장을 Token embedding과 Sentence embedding, Position embedding으로 임베딩하여 합친 다음 임의의 토큰으로 마스킹한다. 학습 방식은 마스킹 된 토큰을 예측하는 방식인 MLM (Masked Language Model)와 두 문장이 있을 때 첫 번째 문장 뒤에 두 번째 문장이 바로 다음에 오는지 예측하는 방식인 NSP(Next Sentence Prediction)로 수행한다.

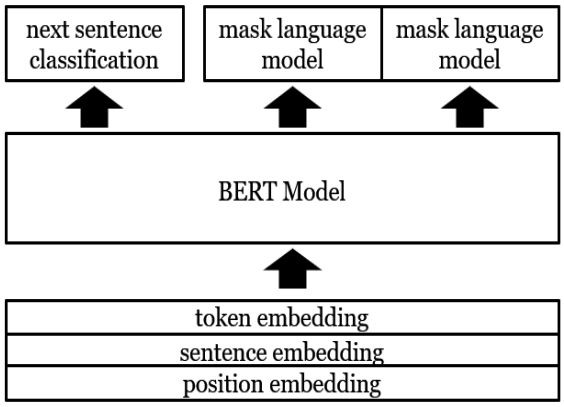

그림 2. BERT 사전 학습 모델
Fig. 2. BERT Pre-training model

3.4 BERT Fine-tuning

이 절에서는 BERT 파인 튜닝을 위해 학습 및 상위 k개를 제외한 나머지 대학의 3가지 공통 주제에 대한 분류하는 방안을 기술한다. 사전 훈련된 BERT를 통해 문장 임베딩 벡터 추출 후 추출된 벡터를 Fully connected layer와 소프트맥스 함수를 적용하여 공통 주제 3가지에 대해 분류할 수 있도록 구성한다. 감성 문장이 많은 상위 k개 대학의 문장과 EM 알고리즘을 통해 나온 라벨 값을 정답 데이터로 사용하여 학습을 수행한다.

3.5 대학 평판도 측정

이 절에서는 각 대학의 종합적인 평판도 및 주제별 평판도 측정을 위해 문장의 감성 점수에 따른 긍정·부정 비율과 감성 점수의 평균을 이용한다. AR-KNU+ 감성 사전을 기반으로 각 대학평판의 긍정·부정 양상과 순위를 측정할 수 있다.

Ⅳ. 실험 및 비교 평가

이 장에서는 기존 연구의 한계점을 해결하기 위해 BERT를 활용한 대학 평판도 평가 실험 및 기존 논문과의 비교 평가를 기술한다.

4.1 실험 데이터

이 절에서는 BERT 사전학습 및 기존 논문과의 비교 평가를 위해 실험 대상인 23개 대학의 키워드를 이용하여 10년 동안의 온라인 뉴스 기사의 사회, 교육면을 자동으로 수집한다. 그리고 실험을 위해 수집한 데이터를 문장 단위로 전처리한 개수는 표 2와 같다.

2004년부터 2014년 동안의 700,984개 문장은 BERT 사전학습을 위해 사용되며, 기존 논문과 같은 데이터인 2014년~2016년 동안의 620,593의 문장은 BERT 파인 튜닝 및 기존 논문과의 비교 평가를 위해 사용된다.

표 2. 연도별 수집된 데이터 개수
Table 2. Number of collected by year

Year	Number of paragraph	Number of sentence
2006~2014	20,035	700,984
2014~2016	14,012	620,593

4.2 대학별 감성 분석 및 주제 추출

각 대학의 감성 문장 도출 및 주제 추출을 위해 4.1절에서 수집된 2014년~2016년 동안의 뉴스 기사를 문장 단위로 파싱하고 AR-KNU+ 감성 사전을 이용하여 각 문장의 감성 점수를 추출한다. 그리고 각 대학 중 감성 문장이 가장 많은 상위 3개의 대학에서 공통 주제 추출을 위해 BERT를 활용하여 각 감성 문장을 768차원의 벡터로 전처리 후 군집화를 수행한다. 군집화를 위해 EM 알고리즘을 사용하며, 파이썬의 머신러닝 패키지 scikit-learn의 GaussianMixture 모델을 이용하여 구현한다. EM 알고리즘으로 도출된 군집 중 상위 3개의 공통 군집을 추출하고 단어 빈도수를 통해 각 군집의 주제를 파악 후 라벨링을 수행한다.

표 3은 각 군집을 레이블링한 결과이다. 주제 추출 결과는 대표적으로 '취업(Employment)', '입학(Entrance)', '학생활동(Student activities)' 주제로 분류된다. 또한, 입학이라는 주제에서 단어 빈도수를 측정해 나온 결과 중 '대입', '입시', '등록금', '수시'라는 단어가 상위 순위에 있어 단편적인 주제가 아닌 세부적인 주제임을 알 수 있다.

표 3. 주제 레이블링 결과
Table 3. Topic labeling result

Topic 1	Topic 2	Topic 3
Employment	Entrance	Student activities

표 4는 상위 3개 대학(U1, U2, U3)에 대한 군집화 및 3가지 주제별 긍정·부정 양상을 보여준다. 대학 관련 뉴스 기사에서 전체적으로 부정적인 견해가 높을 것을 알 수 있다.

표 4. 상위 3개 대학 주제 추출 결과
Table 4. Topic labeling result

Topic	Univ.	Num	Pos.	Neu.	Neg.
Employment	U1	77,901	26%	31%	43%
	U2	40,481	27%	31%	42%
	U3	39,819	26%	32%	42%
Entrance	U1	57,584	18%	14%	68%
	U2	22,224	19%	15%	66%
	U3	20,521	20%	15%	65%
Student activities	U1	11,855	27%	33%	40%
	U2	6607	26%	37%	37%
	U3	6501	26%	37%	37%

4.3 BERT Pre-training 실험

이 절에서는 대학 뉴스 기사로 구성된 7년 동안의 데이터를 사용하여 BERT 모델을 사전 학습한다. 표 2의 데이터로 SentencePiece를 사용하여 8,000개의 Vocabulary를 생성하고, 모델 타입은 BPE(Byte-pair Encoding) 방식을 사용한다. 입력 단어의 15%를 마스킹하고 마스킹 된 단어를 예측하는 MLM 방식과 두 문장이 있을 때 두 번째 문장이 첫 번째 문장의 바로 다음에 오는 문장인지 예측하는 방식을 사용하여 BERT 사전학습을 수행한다.

학습에는 GeFoce RTX 3080 2장을 사용하여 약 2주일 정도 소요되었고, 학습 수행 후 학습 데이터 중 30%인 테스트 데이터에 대해 NSP 정확도는 약 92%, MLM에 대해서는 약 63%의 정확도를 보였다. 학습에 사용되었던 대표적인 파라미터와 하이퍼 파라미터는 다음 표 5와 같다.

표 5. 사전학습 파라미터
Table 5. Pre-training parameters

Parameters, Hyper parameters	Value
Dropout prob	0.1
Activation function	Gelu
Hidden size	768
Max position embedding	512
Attention heads	12
Hidden layers	12
Vocab size	8000
Bach size	64
Learning rate	1e-4
Step	1,000,000

4.4 BERT Fine-tuning 및 분류 비교 평가

이 절에서는 3년 동안의 대학 기사로 구성된 데이터 중 감성 문장이 많은 상위 3개 대학의 문장을 이용하여 파인 튜닝을 진행한다. 먼저 앞에서 사전학습한 모델을 이용하여 '졸업', '입학', '학생활동'에 대한 3가지 주제 분류를 위해 출력층에 Fully connected layer와 소프트맥스 함수를 추가한다. 그리고 모델의 학습 데이터를 token index와 segment index를 합쳐 입력 데이터를 구성한다. 만든 학습 데이터에 대해 20번 반복하여 파인 튜닝을 수행한다. 그리고 10%의 검증 데이터를 구축하여 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수(F1 Score)로 측정한다.

표 6은 기존 논문의 대학 뉴스 기사를 다중 클래스 분류 모델로 분류한 검증 결과와 제안 논문의 BERT를 이용하여 분류한 모델의 검증 결과이다. 표 6에서 대표적으로 F1 점수를 보면 기존 논문보다 제안 논문의 결과가 약 2.4%p 더 좋은 성능을 보인다.

표 6. 모델 검증 결과
Table 6. Model verification results

Method \ Test	Accuracy	Precision	Recall	F1 Score
Previous	0.9001	0.9232	0.8725	0.8970
Proposed	0.9275	0.9279	0.9381	0.9230

표 7은 검증 데이터를 이용하여 주제별로 분류한 기존 논문의 다중 클래스 분류와 제안 논문의 분류

를 통해 도출한 감성 문장이 많은 상위 3개 대학의 대한 정량적인 지표이다. 표 7을 통해 기존 논문보다 제안 논문이 예측된 데이터의 개수가 많아 제안 논문의 분류 정확도에 대해 더 좋은 성능을 보인다.

표 7. 상위 3개 대학 예측된 데이터 개수
Table 7. Top-3 university predicted number of data

Topic	Univ.	Pos.num	Previous	Proposed
Employment	U1	7,790	7,112	7,233
	U2	4,048	3,556	3,570
	U3	3,982	3,451	3,766
Entrance	U1	5,758	5,055	5,256
	U2	2,224	1,984	2,127
	U3	2,052	1,708	1,902
Student activities	U1	1,186	1,095	1,126
	U2	661	640	643
	U3	650	517	532

4.5 대학 평판도 측정 결과

표 8은 AR-KNU+ 감성 사전을 이용하여 상위 3개 대학을 제외한 나머지 대학 중 주제별 감성 문장의 개수가 많은 상위 10개의 대학에 대한 긍정·부정 비율을 나타낸다. 표 8을 통해 주제별 대학 기사의 긍정·부정 양상을 알 수 있다. 각 대학의 주제별로 도출된 순위를 합하여 종합적인 평판도를 측정한다.

표 8. 상위 10개 대학 주제별 분석 결과
Table 8. Top-10 university analysis result by topic

Topic	Univ	Num	Pos.	Neu.	Neg.
Employment	U10	22,437	25%	31%	43%
	U5	19,867	26%	32%	42%
	U4	19,561	28%	32%	40%
	U7	18,890	26%	32%	42%
	U9	15,787	26%	33%	41%
	U6	15,539	26%	32%	42%
	U14	13,808	25%	32%	43%
	U16	13,725	24%	32%	44%
	U19	10,997	23%	30%	47%
	U15	8,766	24%	32%	44%
Entrance	U12	11,364	18%	15%	67%
	U21	10,937	16%	14%	71%
	U16	10,617	17%	14%	69%
	U6	10,451	21%	15%	64%
	U10	8,409	18%	15%	67%
	U8	7,800	19%	16%	65%
	U20	7,578	16%	13%	72%
	U5	6,890	20%	15%	65%
	U7	6,196	19%	16%	65%
	U14	3,863	17%	16%	67%
Student activities	U4	3,484	26%	42%	31%
	U10	3,393	24%	39%	37%
	U5	3,336	25%	41%	33%
	U7	2,894	26%	39%	35%
	U6	2,523	26%	40%	33%
	U9	2,461	23%	42%	34%
	U14	2,083	24%	40%	37%
	U16	1,972	23%	39%	37%
	U18	1,125	26%	36%	39%
	U19	1,122	23%	36%	41%

표 9. 실험에 따른 대학 평판도 순위 비교
Table 9. Comparison of university reputation ranking by experiment methods: ranking order

Methods / Ranking	Previous (2014~2016)	Proposed (2014~2016)
1	U21	U20
2	U19	U4
3	U10	U13
4	U17	U12
5	U8	U5
6	U6	U6
7	U5	U10
8	U11	U3
9	U9	U9
10	U4	U23
11	U14	U8
12	U22	U2
13	U3	U7
14	U1	U15
15	U13	U18
16	U7	U11
17	U2	U1
18	U15	U14
19	U20	U21
20	U16	U17
21	U12	U16
22	U18	U19
23	U23	U22

23개 대학의 평판도 순위 측정을 위해 주제별로 긍정, 부정 비율을 통합하여 최종적으로 대학 평판도 종합적인 순위를 도출한다. 표 9는 3년 동안의 데이터를 이용하여 기존 논문과 제안 논문의 종합적인 평판도 순위를 도출한 결과이다. 표 9를 통해 실험에 방법에 따른 순위가 다른 것을 알 수 있다. BERT를 활용하여 기존 연구의 알고리즘보다 분류 정확도 부분에서 우수한 성능을 보여 제안 논문의 대학 평판도 순위가 더 신뢰성 있다.

V. 결 론

이 논문에서는 기존 감성 분석 기반의 대학 평판도 평가 방식의 정확도와 신뢰도 향상을 위해 BERT를 활용한 새로운 방안을 제안하였다. 이를 위해 기존 연구에서 23개 대학을 키워드로 수집한 3년 동안에 대학 뉴스 기사보다 7년 동안의 대학 뉴스 기사를 더 수집하였다. 7년 동안의 대학 뉴스 기사는 BERT에 사전학습을 수행하였으며, 3년 동안의 뉴스 기사는 대학의 주제 도출을 위해 AR-KNU+ 감성 사전 기반으로 감성 문장을 도출하고 감성 문장이 많은 상위 3개 대학을 기준으로 군집화를 수행하여 3가지의 주제로 라벨링을 진행하였다. 그리고 나머지 대학에 대한 주제 분류를 실시하기 위해 파인 튜닝을 수행하였다. 먼저 사전 학습된 BERT의 출력층을 주제를 분류할 수 있게 Softmax 함수를 사용하여 3개의 노드로 구성하였다. 그리고 상위 3개 대학의 감성 문장을 학습 데이터로, 군집화를 통해 도출된 군집 레이블을 정답 데이터로 학습하였다. 학습된 BERT 모델을 이용하여 나머지 20개 대학 문장에 대한 주제 분류를 수행하였다. 마지막으로 감성 문장을 통해 3가지의 주제별로 긍정·부정 비율을 측정하여 종합적 대학 평판도를 도출하였다.

BERT를 활용하여 기존 연구의 알고리즘보다 분류 정확도 부분에서 우수한 성능을 보였지만 7년 동안의 데이터를 가지고 사전학습을 진행한 BERT의 MLM 정확도에서 낮은 성능을 보였다. 이것은 BERT 모델에서 상대적으로 적은 학습데이터의 사용이 영향을 미친 것으로 판단된다.

향후에는 대학 관련 뉴스 기사 데이터를 대용량으로 수집하여 추가로 학습할 예정이다. 제안 연구보다 더 많은 대학과 관련된 Vocabulary를 생성하여 MLM 및 NSP에 대한 정확도를 높이고 대학 평판도 모델의 신뢰도를 높일 것이다.

References

[1] Chosunilbo university reputation, https://www.chosun.com/national/qs_evaluation/ [accessed: Aug. 01, 2021]

[2] Joongangilbo university reputation, http://univ.joongang.co.kr. [accessed: Aug. 01, 2021]

[3] Korea University NewPaper, https://www.kunews.ac.kr/news/articleView.html?idxno=15792 [accessed: Aug. 02, 2021]

[4] S. M. Park, C. M. Eom, B. W. On, and D. W. Jeong, "An AR-KNU Sentiment Lexicon-based University Reputation Assessment Using Online News Data", The Journal of Korean Institute of Information Technology, Vol. 17, No. 3, pp. 11-21, Mar. 2019. http://doi.org/10.14801/jkiit.2019.17.3.11.

[5] S. H. Chae, D. W. Jeong, B, W, On, and J. W. Gim, "University Reputation Assessment through AR-KNU Expansion", The Journal of Korean Institute of Information Technology, Vol. 18, No. 11, pp. 35-45, Nov. 2020. http://doi.org/10.14801/jkiit.2020.18.11.35.

[6] Ariticial Intelligence Times, https://www.aitimes.kr/news/articleView.html?idxno=15036 [accessed: Aug. 02, 2021]

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT:Pre-training of Deep Bidirectional Transformers forLanguage Understanding", arXiv: 1810.04805, 2018.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention Is All You Need, NIPS, 2017.

저자소개

정 주 현 (Juhyeon Jeong)

2022년 1월 현재 : 군산대학교 소프트웨어융합공학과(학부생)
관심분야 : 빅데이터 분석, 감성 분석, 사물 인터넷

김 장 원 (Jangwon Gim)

2005년 : 상명대학교 컴퓨터소프트웨어공학과(학사)
2008년 : 고려대학교 컴퓨터학과(석사)
2012년 : 고려대학교 컴퓨터·전파·통신공학과(박사)
2013년 : 한국과학기술정보연구원 선임연구원
2017년 ~ 현재 : 군산대학교 소프트웨어융합공학과 교수
관심분야 : 자연어처리, 빅데이터 분석, 지식 그래프, 지식 그래프 임베딩

온 병 원 (Byung-Won On)

2007년 : The pennsylvania state U. 컴퓨터공학과(박사)
2008년 : 캐나다 U. of British Columbia 박사 후 연구원
2010년 : U. of Illinois at Urbana-Champaign ADSC 연구소 선임연구원
2011년 : 차세대융합기술연구원 공공데이터 연구센터 센터장
2014년 ~ 현재 : 군산대학교 소프트웨어융합공학과 교수
관심분야 : 데이터 마이닝, 빅데이터, 인공지능, 강화학습

정 동 원 (Dongwon Jeong)

1997년 : 군산대학교 컴퓨터과학과(학사)
1999년 : 충북대학교 전산학과(석사)
2004년 : 고려대학교 컴퓨터학과(박사)
2005년 ~ 현재 : 군산대학교 소프트웨어융합공학과 교수
관심분야 : 데이터베이스, 시맨틱 서비스, 빅데이터, 사물인터넷, 엣지컴퓨팅