Journal of KIIT. Vol. 18, No. 11, pp. 35-45, Nov. 30, 2020. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2020.18.11.35

AR-KNU 확장을 통한 대학 평판도 평가

채수현*[1], 정동원*[2], 온병원*[3], 김장원*[4]

University Reputation Assessment through AR-KNU Expansion

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

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

요 약

기존의 대학 평판도 평가 연구는 높은 비용과 감성의 편차가 발생하는 문제점을 지닌다. 따라서 이 논문에서는 이전 연구의 감성 사전 확장을 통한 대학 평판도 평가를 제안한다. 이를 위해, 우선 풍부한 데이터를 이용하여 기존의 감성 사전을 확장하고 군집화를 이용해 주제를 추출한다. 이러한 정보를 이용하여 대학의 평판도를 평가하고 추이를 분석한다. 연구 결과를 활용하여 기존의 설문 조사 기반 대학평가의 문제점을 개선할 수 있으며, 데이터의 크기와 대학 평판도의 관계를 파악할 수 있다. 또한, 주제별 대학평판의 개선을 위한 기초 자료로 활용될 수 있다.

Abstract

The existing university reputation assessments have several problems such as high cost and sentiment deflection. Therefore, this paper proposes a university reputation assessment through expanding the sentiment lexicon of the previous study. To achieve the goal, the previous sentiment lexicon is first expanded using rich data, and topics are extracted by clustering. This paper evaluates the university reputation and analyzes trends. With the outcome of this paper, we can improve the problem of the survey-based university reputation assessment and can also identify the relationships between data size and university reputation. In addition, this study can be used basic data to improve university reputation by topic.

Keywords

university reputation, sentiment lexicon, expansion, sentiment analysis, trend analysis

* 군산대학교 소프트웨어융합공학과
- ORCID[1]: https://orcid.org/0000-0001-5154-3986
- ORCID[2]: https://orcid.org/0000-0001-9881-5336
- ORCID[3]: https://orcid.org/0000-0001-6929-3188
- ORCID[4]: https://orcid.org/0000-0002-4480-7944

· Received: Oct. 29, 2020, Revised: Nov. 22, 2020, Accepted: Nov. 25, 2020
· Co-Corresponding Author: Dongwon Jeong and Jangwon Gim
Dept. of Software Convergence Engineering, Kunsan National University, 558, Daehak-ro, Gunsan, Jeollabuk-do, Korea.
Tel.: +82-63-469-8911, Email: {djeong, jwgim}@kunsan.ac.kr

Ⅰ. 서 론

4차 산업혁명 이후 정보통신의 발달과 함께 다양한 온라인 플랫폼의 서비스가 활발해졌으며, 그로 인해 많은 사회 이슈에 대한 접근성이 증가하고 있다[1][2]. 특히 온라인 뉴스와 SNS는 방송 및 신문과 같은 매체를 대신하여 실시간으로 증가하는 많은 양의 정보 전달이 가능하므로 사용자의 삶의 질을 높여주고 있다[3]. 이러한 비정형 및 반정형 데이터의 증가와 함께 데이터 마이닝(Data mining) 기법을 기반으로 다양하고 많은 데이터를 활용하여 의미 있는 통찰력 도출을 위한 연구가 활발히 진행중이다. 그중에서도 실시간으로 증가하는 온라인 뉴스 및 SNS와 같은 텍스트 데이터에 대해 유용한 정보를 도출하고 분석하는 자연어 처리 기법이 대두되고 있다[4]. 자연어 처리는 텍스트 데이터에 대해 정보검색 및 문서 분류 등의 기술을 적용하여 양질의 정보를 분석하며, 특히 감성을 도출하는 감성 분석(Sentiment analysis)이 다양한 도메인에서 활용되고 있다[5].

이러한 상황에서 실시간으로 증가하는 온라인 데이터에 대한 대중의 감성을 분석하는 연구가 활발히 진행되고 있으며, 이를 통해 기관 또는 시장은 분석 결과를 적극적으로 반영하여 여론의 관심에 빠르게 대응하고 있다[6]. 대표적으로 입시를 준비하는 수험생과 보호자 및 취업을 준비하는 대학의 졸업생은 매년 변하는 대학평판에 민감한 반응을 보인다. 따라서 다양한 방법을 통해 대학 평판도를 측정하고 평가하는 연구들이 수행되고 있다.

대학의 평판 측정 및 분석을 위해 대표적으로 사용되는 방법은 설문 조사 기반의 평가와 자연어 처리 기법을 이용한 평가 방법으로 분류된다. 기존의 설문 조사 기반의 대학평판 측정은 중앙일보에서 매년 수행하는 대학 종합평가가 있다[7]. 중앙일보는 다양한 지표에 대한 설문 조사를 진행하여 여론의 의견을 기반으로 종합적인 대학의 평판도를 측정하고, 순위를 도출한다. 하지만 설문 조사 방법은 설문의 결과를 종합 및 분석을 위한 많은 시간과 높은 비용이 필요하므로 설문 표본의 크기가 제한적이다. 또한, 응답자가 설문 내용에 성실하게 응하지 않는다면 설문 조사 결과를 신뢰하기 어렵다.

이러한 설문 조사를 이용한 대학 평판도 측정 방법의 문제해결을 위해 텍스트 데이터에 대하여 자연어 처리를 활용한 대학평판 분석 연구가 진행되었다[8-9]. 그중에서 감성 분석을 이용한 대학의 평판도 분석 방법 연구가 수행되었으며, 대학의 평판도 측정을 위해 각 대학에 대한 여론의 관심을 분석하였다[10]-[12]. 감성 분석을 위해 온라인 뉴스 및 다양한 SNS 등에서 데이터를 수집하고 결과를 바탕으로 대학을 평가하였다. 이 논문은 설문 조사 기반 대학평가의 문제해결을 위한 기존의 감성 분석 기반의 대학 평판도 평가 연구를 활용한다[12]. [12]는 감성 사전을 구축하여 뉴스 기사에 대한 감성 문단을 추출하고, 감성 사전을 통해 뉴스 기사의 주제를 분류한다. 분류된 주제에 대한 대학 평판도 평가를 위해 주제별 감성 문단의 긍정·부정 비율을 측정하여 종합 평판도를 도출한다. 그러나 텍스트 데이터의 크기가 충분하지 않으므로 대학별 데이터의 편차가 존재하며, 그로 인해 구축된 감성 사전의 크기가 작아 주제별 대학의 극성이 편향되는 문제가 존재한다. 이러한 문제는 분석 결과가 일정하지 않으며, 실제 여론의 관심과 다른 차이를 보일 수 있다.

따라서 이 논문은 데이터의 크기를 추가하여 더욱 객관적인 대학의 평판도 측정 실험을 위해 연도별 뉴스 기사를 추가하고, 추가된 데이터를 활용하여 감성 사전을 확장한다. 그리고 데이터 크기에 따른 대학 평판도 변화를 확인하고, 시계열 데이터로부터 대학평판의 순위에 관한 추이 분석을 수행한다. 이를 위해 분석 대상 대학의 키워드를 통해 연도별 온라인 뉴스 기사를 수집하고, 기존의 AR-KNU 감성 사전(Academic Reputation-KNU sentiment lexicon)을 기반으로 대학 도메인에 의존적인 감성 사전을 확장한다. 확장된 감성 사전을 기반으로 각 대학의 감성 문단을 추출하고 감성 문단의 개수가 많은 상위 k개의 대학에 대해 군집화 기법을 적용하여 뉴스 기사의 주제(이슈)를 도출한다. 군집화 대상 대학을 제외한 나머지 대학의 뉴스 기사에서 주제 도출을 위해 다중 클래스 분류를 수행한다. 그리고 각 대학의 평판도 평가를 위해 주제별 대학에 대한 긍정·부정 비율을 도출하여 종합적인 평판도

를 측정한다. 마지막으로, 데이터 크기에 따른 대학 평판도 비교 및 순위에 관한 추이 분석을 통해 연도별 대학평판의 변화를 확인한다.

이 논문의 구성은 다음과 같다. 제2장은 관련 연구로서 기존 연구의 문제와 감성 분석 기반의 대학 평판도 측정 연구의 필요성에 관해 기술한다. 제3장에서는 대학 평판도 측정 방법을 정의하고, 제4장에서는 실험을 수행하고, 결과에 대한 분석을 서술한다. 마지막으로, 제5장에서는 연구의 결론과 한계점 및 향후 연구에 관하여 기술한다.

II. 관련 연구

대학의 평판도를 측정하고 평가를 위하여 대표적으로 설문 조사 기반의 평가와 자연어 처리 기반의 평가 연구들이 수행되었다. 이 장에서는 대학 평판도 평가를 위한 기존 연구를 소개하고 기존 연구의 문제점을 통해 감성 분석 기반의 대학 평판도 평가 연구의 필요성을 제시한다.

2.1 설문 조사 기반의 대학 평판도 평가

대학의 평판도 평가를 위해 설문 조사 기반의 연구들이 수행되었다. 대표적으로 중앙일보에서는 다양한 지표를 통해 설문 조사를 진행하고 이를 바탕으로 매년 대학에 대한 종합평가를 진행한다[7]. 설문 조사에 사용되는 지표는 양질의 교육을 수행하는 대학 및 향후 발전 가능성이 보이는 대학 등이 있으며, 이러한 지표의 설문 조사를 기반으로 대학의 평판을 측정하고 순위를 도출하여 결과를 공개한다.

[8]은 설문 조사를 통해 교직원과 학생들 사이의 대학에 대한 인식 차이를 분석하고 대학평판에 영향을 미치는 요인에 관해 연구를 수행하였다. 이를 통해 대학평판에 유의미한 영향을 미치는 요소를 파악하고, 대학평가 및 평가 결과가 수요자에 대해 어떠한 영향력을 가지는지 분석하였다. 하지만 평가 대학과 설문 조사 대상의 제한으로 인해 대학평가의 신뢰도에 대하여 부정적인 견해가 많은 것을 확인하였다.

[9]는 해외 대학의 명성과 이미지가 대학 진학에 미치는 영향 분석을 위해 설문 조사를 진행하였다. 설문지를 통한 대학의 순위와 평판 측정을 위해 교육, 연구 및 서비스 등의 척도를 조사하고, 커뮤니티 사용자와 무작위로 선정한 응답자에 대해 설문 조사를 진행하였다. 분석 결과, 부적합한 응답자 선정과 설문 조사 표본의 크기가 작은 문제로 인해 특정 대학의 인지도는 그 대학의 선호도에 미치는 영향이 적다는 결과를 도출하였다.

이처럼 설문 조사 기반의 대학 평판도 측정은 설문에 대한 응답자의 태도와 무분별한 표본 선정 및 크기로 인해 분석 결과에 대해 신뢰하기 어렵다는 문제점이 있다.

2.2 자연어 처리 기반의 대학 평판도 평가

기존의 설문 조사를 통한 문제점 해결을 위해 온라인 데이터를 활용한 자연어 처리 분석 기반의 대학 평판도 평가 연구가 필요하며, 다양한 연구가 수행되었다.

[11]은 특수목적대학에 대한 인식 분석을 위해 다양한 SNS 데이터를 수집하여 연구를 수행하였다. 이를 위해 빅데이터 분석 도구를 사용하여 특수목적대학에 관한 연관 주제, 여론의 인식 및 호감도를 분석하고, 구글 트렌드를 통해 결과를 비교하였다. 그 결과, SNS 데이터를 사용하는 사용자의 연령층과 단기간에 형성된 대학의 평판에 대한 파급효과를 확인하였으며, 다양한 온라인 매체에서 생성되는 데이터에 관한 연구의 중요성을 제시하였다. 하지만 통계분석을 활용하여 어휘의 감성을 도출하기 어렵고, 적은 데이터의 크기로 인해 간접적인 이미지 추측의 문제가 존재한다.

[12]는 대학의 평판도 평가를 위해 온라인 뉴스 데이터를 기반으로 AR-KNU 감성 사전을 구축하고, 텍스트 마이닝 기법의 하나인 군집화 알고리즘을 사용하여 뉴스 기사의 주제를 분류하였다. 각 주제에 대한 대학의 긍정·부정의 양상을 도출하여 대학의 종합적인 평판도를 측정한 뒤, 기존 연구들과 결과를 비교 분석하였다. 그러나 적은 데이터의 크기로 인해 긍정·부정의 양상이 한쪽으로 편향되고 그

로 인한 이슈의 부정확한 결과가 발생하는 문제점을 지닌다.

이 논문에서는 설문 조사 기반의 대학 평판도 평가 연구의 문제점 해결을 위해 제시된 온라인 데이터를 활용한 감성 분석 연구 기반의 대학 평판도 평가 실험을 수행한다. 또한, 데이터를 추가하여 [12]에서 구축한 AR-KNU 감성 사전을 확장하고, 확장된 감성 사전을 통해 뉴스 기사의 주제를 분류한다. 그리고 감성 문단을 사용해 주제별 각 대학의 긍정·부정 양상을 도출하고, 데이터 크기에 따른 종합적인 평판도를 비교 분석한다. 마지막으로, 시계열 정보를 포함한 뉴스 기사를 사용하여 대학 평판도의 추이를 분석한다.

III. 제안 방법

이 논문에서는 기존 연구의 대학 평판도 알고리즘에 대한 정확도 향상과 대학 평판도의 추이 분석을 위해 데이터를 추가하여 실험한다. 그림 1은 제안 방법의 전체 과정을 나타내며 추이 분석을 위한 데이터 수집, 감성 분석을 통한 감성 어휘 분류, 각 대학의 주제 도출 및 주제에 따른 대학의 평판도 측정 및 추이 분석으로 구성한다.

첫 번째로, 추이 분석에 필요한 연도별 대학의 뉴스 기사를 수집하여 기존의 AR-KNU 감성 사전 확장을 위해 뉴스 기사에 대한 감성 어휘를 도출한다. 그리고 감성 어휘에 대한 긍정·부정 정도를 5점 척도를 이용하여 분류하고, 각 대학의 뉴스 기사에 포함된 감성 어휘의 점수를 계산하여 감성 문단의 점수를 파악한다. 두 번째로, 각 대학에 대한 감성 문단의 개수가 가장 많은 상위 k개의 대학을 추출하여 유사한 문단으로 이루어진 주제 군집화를 위해 EM(Expectation Maximization) 알고리즘을 수행한다. 도출된 주제 군집의 의미 파악을 위해 군집마다 레이블링을 수행한다. 세 번째로, 나머지 대학의 주제 분류를 위해 레이블링 된 주제 군집에 포함된 문단을 학습 데이터로, 군집의 레이블을 정답 데이터로 입력하여 다중 클래스 분류를 수행한다. 마지막으로, 특정 주제 군집에 대한 각 대학평판의 긍정·부정 성향 파악을 위해 감성 문단을 이용하여 긍정·부정 비율을 계산한다. 그리고 모든 대학의 종합적인 평판 순위 파악을 위해 평판도 점수를 도출하고, 연도별 평판도 점수를 사용하여 각 대학의 추이를 분석한다.

3.1 대학별 감성 분석

이 절에서는 대학의 평판도에 따른 추이 분석을 위해 연도별 온라인 뉴스 기사 데이터를 수집하고, 감성 분석을 위해 감성 사전을 구축한다. 이를 위해 기존에 구축한 AR-KNU 감성 사전을 기반으로 감성 어휘를 추가하여 확장된 AR-KNU+ 감성 사전을 구축한다. 기존의 AR-KNU 감성 사전은 대학 도메인에 대해 1년 동안의 뉴스 기사를 기반으로 구축된 감성 사전이며, 뉴스 기사에 포함된 감성 어휘를 분류하여 감성점수를 측정한다. 이를 활용한 감성 어휘 도출을 위해 수집한 데이터에 대한 형태소 분석을 수행하여 품사를 구분한다. 구분된 품사는 명사, 동사, 형용사, 부사 및 감탄사이며, 해당 품사에 속한 어휘 중 AR-KNU 감성 사전에 포함되지 않은 어휘에 대하여 감성점수를 측정한다.

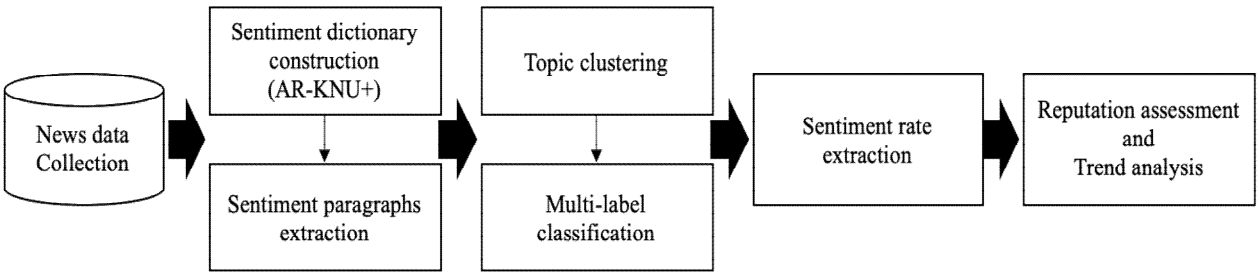


그림 1. 대학 평판도 평가의 전체적인 프로세스
Fig. 1. Overall process of university reputation assessment

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

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

어휘의 감성점수는 5점 척도 방식으로 표 1과 같이 구분하며, 감성점수 측정 결과를 통해 확장된 AR-KNU+ 감성 사전을 구축한다. 이처럼 구축된 AR-KNU+ 감성 사전을 이용하여 각 대학의 뉴스 기사에 대한 감성 분석 실험을 위해 감성 문단을 도출한다. 감성 문단 도출은 각 대학의 뉴스 기사를 대학에 따라 문단으로 구분하고, AR-KNU+ 감성 사전을 기반으로 각 문단에 포함된 감성 어휘의 점수를 모두 더하여 계산한다. 대학에 따른 문단에 포함된 감성 어휘의 합이 양수일 경우 긍정적인 문단, 합이 음수일 경우 부정적인 문단으로 분류한다.

3.2 뉴스 기사 주제 추출

이 절에서는 대학의 뉴스 기사에 대해 공통주제 도출을 위한 방법을 제안하며, 주제 도출을 위하여 이전에 수행한 감성 분석 결과를 사용한다. 감성 분석의 결과인 감성 문단의 개수가 많을수록 뉴스 기사에 주제(이슈)가 포함될 가능성이 크기 때문에, 감성 문단의 개수가 많은 상위 k개의 대학을 추출하여 가우시안 혼합 모델(Gaussian mixture model) 기반의 EM 알고리즘을 적용해 주제 군집화를 수행한다.

EM 알고리즘은 최대 가능성(Maximum likelihood)을 예측하거나, 모델의 파라미터를 귀납 추론에 수렴하도록 조정하는 방식으로 주어진 데이터 중 유사한 데이터를 군집하는 군집화 알고리즘이며, 예상 단계(E-step, Expectation step)와 최대화 단계(M-step, Maximization step)를 반복하여 최적의 군집을 찾는 방법이다[13]. 예상 단계는 군집의 중심과 가장 가까운 데이터를 예측하여 군집화하는 단계이며, 최대화 단계는 군집화된 데이터들과 중심의 거리에 대한 총합이 최소가 되도록 군집의 중심을 초기화하는 단계이다. 이 논문에서는 상위 k개 대학의 감성 문단을 이용해 군집된 주제 도출을 위해 그림 2의 알고리즘과 같이 가우시안 혼합 모델 기반의 EM 알고리즘을 구현한다. 초기에 군집의 중심을 정하는 파라미터들을 임의로 초기화한 후, 예상 단계인 7~9줄에서 입력 데이터들의 가장 유사한 군집을 사후 확률(γ)로 계산한다. 그리고 10~13줄의 최대화 단계에서 데이터와 중심의 거리에 대한 분산(Σ)이 최소가 되는 최적의 중심을 찾아 군집을 개선한다. 마지막으로 14~15줄에서 사후 확률이 가장 높은 군집을 출력하여 알고리즘을 종료한다.

Algorithm 1: EM algorithm of GMM

Input : a given data $X = \{x_1, x_2, ..., x_n\}$
Output : class labels $Y = \{y_1, y_2, ..., y_n\}$

1 Randomly initialize c, μ, Σ
2 γ : posterior probability
3 c : mixture component
4 μ : mean vector
5 Σ : covariance matrix
6 **for** i **in** I
7 **for** n **in** N
8 **for** k **in** K
9 $$\gamma_{nk}^{(i)} = \frac{c_k^{(i-1)} N(x_n|\mu_k^{(i-1)}, \Sigma_k^{(i-1)})}{\sum_{k=1}^{K} c_k^{(i-1)} N(x_n|\mu_k^{(i-1)}, \Sigma_k^{(i-1)})}$$
10 **for** k **in** K
11 $$c_k^{(i)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk}^{(i)}$$
12 $$\mu_k^{(i)} = \frac{\sum_{n=1}^{N} \gamma_{nk}^{(i)} \cdot \ x_n}{\sum_{n=1}^{N} \gamma_{nk}^{(i)}}$$
13 $$\Sigma_k^{(i)} = \frac{\sum_{n=1}^{N} \gamma_{nk}^{(i)} \cdot \ (x_n - \mu_k^{(i)})(x_n - \mu_k^{(i)})^T}{\sum_{n=1}^{N} \gamma_{nk}^{(i)}}$$
14 **for** n **in** N
15 $$y_n = \arg\max_k \gamma_{nk}^{(i)}$$

그림 2. 주제 도출을 위한 GMM 기반 EM 군집화 알고리즘
Fig. 2. EM clustering algorithm based on GMM for topic extraction

3.3 대학별 주제 분류

이 절에서는 EM 알고리즘의 결과를 통해 도출된 주제에 대하여 상위 k개 대학을 제외한 나머지 대학의 주제 분류를 위해 다중 클래스 분류를 수행한다. 다중 클래스 분류 모델 학습을 위해 상위 k개 대학의 감성 문단을 학습 데이터로, EM 알고리즘을 통해 도출된 해당 감성 문단의 주제인 군집 레이블을 정답 데이터로 모델을 학습한다. 다중 클래스 분류는 신경망을 통해 수행하며 학습된 다중 클래스 분류 모델의 입력층에 나머지 대학의 감성 문단을 입력한다. 그리고 은닉층을 통해 입력 데이터의 주제를 예측한 뒤, 출력층에서 각 주제에 대해 감성 문단이 포함될 확률을 나타낸다. 이때, 출력된 결과 중 가장 확률이 높은 레이블이 입력 데이터의 주제가 되며 다중 클래스 분류 모델을 사용하여 각 대학의 주제를 분류한다.

3.4 대학의 평판도 측정 및 추이 분석

이 절에서는 각 대학의 주제에 대한 평판도 측정을 위해 문단의 감성점수에 따른 긍정·부정 비율을 이용한다. AR-KNU+ 감성 사전을 기반으로 주제에 따른 대학의 긍정·부정 비율을 통해 해당 주제에 관한 각 대학평판의 양상을 세부적으로 확인할 수 있으며, 각 대학의 종합적인 평판도를 측정하여 순위를 나열한다. 먼저 대학별 감성 분석의 결과인 각 문단의 감성점수를 주제에 따라 분류하고, 감성점수를 통해 주제별 대학의 긍정·부정의 비율을 계산한다. 주제별 각 대학의 긍정·부정 비율은 해당 대학의 긍정적인 견해와 부정적인 견해를 나타내므로 종합적인 대학의 평판이라고 할 수 있으며, 이를 합하여 각 대학의 평판도를 측정한다.

대학의 평판에 대학 추이 분석을 위해 시계열 정보를 사용하여 연도별 대학 평판도를 측정하고, 그 결과를 이용하여 대학 평판도 변화의 추이를 분석한다.

IV. 실험 및 평가

4.1 대학별 감성 분석 및 주제 추출

기존 연구의 데이터 크기에 대한 문제를 해결하고 대학의 평판도 측정 및 추이 분석을 위해 2014년부터 2016년까지 3년 동안의 온라인 뉴스 기사 중 실험 대상 23개 대학의 키워드를 이용하여 데이터를 수집한다. 수집된 데이터의 개수는 표 2와 같으며 연도별 모든 대학의 뉴스 기사를 문단 단위로 구분하여 측정한다. 이를 활용하여 AR-KNU+ 감성 사전 구축을 위해 수집한 데이터에 대한 파이썬 형태소 분석 패키지인 KoNLPy의 MeCab 형태소 분석기를 사용하고 명사, 동사, 형용사, 부사 및 감탄사 품사에 해당하는 어휘를 추출한다. 추출된 어휘 중 기존의 AR-KNU 감성 사전에 포함되지 않은 어휘에 대해 4명의 연구자가 투표와 토론을 진행하여 수작업으로 감성점수를 측정한 후 확장된 AR-KNU+ 감성 사전을 구축한다. 각 대학의 감성 문단 도출을 위해 뉴스 기사를 문단 단위로 파싱하고 구축된 감성 사전을 이용하여 각 문단에 포함된 감성 어휘 점수의 총합을 계산한다. 그리고 군집화 및 주제 분류 학습을 위해 fasttext의 skipgram 모델을 사용하여 단어 임베딩(Word embedding)을 수행하며, 각 감성 문단을 100차원의 벡터로 전처리한다.

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

Year	Number of paragraph
2014	4,100
2015	6,372
2016	3,720
Total	14,012

다음 뉴스 기사의 주제(이슈) 추출을 위해 감성 문단의 개수가 많은 상위 3개 대학 U1, U2, U3(표 3의 Univ.)에 대해 군집화를 수행한다. 유사한 감성 문단의 군집화를 위해 가우시안 혼합 모델 기반의 EM 알고리즘을 사용하며, 파이썬의 기계 학습 패키지인 scikit-learn의 GaussianMixture 모델을 이용해 구현한다. 구현한 모델을 통해 U1, U2, U3 대학에 대하여 3개의 주제 군집화를 수행한다. 도출된 3개의 군집에 대하여 각 군집이 대표하는 주제 파악을

위해 상위 k개 단어를 통하여 주제 레이블링을 수행한다. 주제 분류 및 레이블링 결과, 대학에 대한 온라인 뉴스 기사는 대표적으로 '졸업·취업(Graduation and employment)', '입학(Entrance)' 및 '학생 활동(Student activities)'의 주제로 분류된다. 또한, '입학'이라는 주제는 상위 k개 단어를 통해 단순히 입학 전형이나 경쟁률뿐만 아니라 등록금과 같은 세부적인 주제도 포함하는 것을 확인할 수 있다. 이는 표 3의 도출된 주제에 대한 대학평판의 긍정·부정 양상이 단편적인 주제에 관한 결과가 아닌 세부적인 주제를 포함한 복합적인 결과임을 알 수 있다.

표 3. 주제별 U1, U2, U3의 감성 비율
Table 3. Sentiment rates of U1, U2, U3 by topic

Topic	Univ.	Num	Pos.	Neu.	Neg.
Graduation and employment	U1	846	0.25	0.25	0.5
	U2	445	0.22	0.2	0.58
	U3	375	0.26	0.19	0.55
Entrance	U1	457	0.16	0.28	0.56
	U2	279	0.19	0.23	0.59
	U3	241	0.22	0.17	0.6
Student activities	U1	1,595	0.13	0.04	0.83
	U2	610	0.16	0.05	0.79
	U3	579	0.15	0.05	0.79

4.2 대학별 주제 분류 실험

감성 문단의 개수가 많은 상위 3개의 대학을 제외한 나머지 대학에 대한 주제 분류를 위해 다중 클래스 분류를 수행한다. 다중 클래스 분류 모델은 파이썬의 딥러닝 패키지인 Keras를 사용하며, 학습과 분류는 GPU 환경에서 진행한다. 학습 모델의 입력층에는 감성 문단에 대한 100차원의 벡터를 입력하며, 은닉층은 Relu 함수를 사용하는 100개의 노드로 구성한다. 그리고 출력층은 입력 데이터에 대한 세 가지 주제 분류를 위해 Softmax 함수를 사용하는 3개의 노드로 구성한다.

다중 클래스 분류 모델의 분류 정확도 평가를 위해 학습 데이터 중 10%를 검증 데이터(Validation data)로 구축하고, 표 4와 같이 정확도(Accuracy), 정밀도(Precision), 재현율(Recall) 및 F1 점수(F1 Score)를 측정한다. 모델의 학습 결과 문단의 주제를 약 86%~91%의 정확도로 분류한다.

표 4. 다중 클래스 분류 모델의 정확도
Table 4. Accuracy of multi-label classification model

Year	Accuracy	Precision	Recall	F1 Score
2014	0.8602	0.8907	0.8095	0.8481
2015	0.8932	0.9063	0.8655	0.8853
2016	0.8699	0.8977	0.8425	0.8691
Total	0.9001	0.9232	0.8725	0.8970

4.3 대학 평판도 측정 결과

표 5는 U1, U2, U3를 제외한 나머지 대학 중 감성 문단의 개수가 많은 상위 10개의 대학에 대한 AR-KNU+ 감성 사전 기반 주제별 감성 문단의 긍정·부정 비율이다. 표 5를 통해 특정 주제에 대한 각 대학의 긍정·부정 양상을 알 수 있다. 이를 활용하여 대학의 종합적인 평판도를 도출한다.

표 5. 주제별 상위 10개 대학의 감성 비율
Table 5. Sentiment rate of Top-10 university by topic

Topic	Univ.	Num	Pos.	Neu.	Neg.
Graduation and employment	U4	361	0.2	0.16	0.63
	U5	321	0.17	0.15	0.67
	U6	282	0.2	0.16	0.64
	U7	338	0.18	0.17	0.65
	U8	305	0.27	0.21	0.52
	U9	234	0.18	0.24	0.57
	U10	267	0.14	0.2	0.66
	U11	212	0.33	0.25	0.42
	U12	204	0.13	0.18	0.69
	U13	151	0.3	0.17	0.52
Entrance	U4	169	0.2	0.3	0.5
	U5	180	0.21	0.25	0.54
	U6	165	0.16	0.28	0.56
	U7	161	0.2	0.23	0.57
	U8	178	0.18	0.23	0.48
	U9	212	0.16	0.28	0.56
	U10	90	0.24	0.38	0.38
	U11	174	0.2	0.35	0.45
	U12	101	0.16	0.31	0.53
	U13	114	0.18	0.26	0.55
Student activities	U4	382	0.11	0.05	0.84
	U5	390	0.07	0.04	0.89
	U6	334	0.11	0.04	0.85
	U7	225	0.16	0.08	0.76
	U8	235	0.14	0.1	0.76
	U9	251	0.08	0.03	0.89
	U10	255	0.07	0.06	0.87
	U11	209	0.23	0.14	0.63
	U12	153	0.16	0.09	0.75
	U13	178	0.12	0.04	0.83

각 대학의 평판에 대한 순위 측정을 위해 주제별 대학의 긍정·부정 비율을 통합하여 최종적으로 종합 평판도를 계산한 뒤, 평판도를 기반으로 각 대학의 순위를 도출한다. 도출한 대학 평판도 순위는 2016년도와 데이터를 추가한 종합 연도를 기준으로 중앙일보에서 수행한 대학평가 순위와 비교한다. 표 6은 순위를 기준으로 실험 방법에 따른 23개 대학과 중복되는 중앙일보의 대학을 비교한 것이며, 표 7은 대학을 기준으로 실험 방법에 따른 순위를 비교한 것이다. 이를 통해 연도별 데이터 크기 및 실험 방법에 따른 대학의 평판도 순위를 알 수 있다.

그림 3은 연도별 각 대학의 평판도 추이 결과를 나타낸다. 그림 3의 (a)는 감성 문단의 개수가 많은 상위 5개 대학의 2014년~2016년도 및 종합 연도에 대한 평판도 순위를 나타낸다. 그림 3의 (b)는 감성 문단의 개수가 적은 하위 5개 대학의 평판도 순위 추이 결과이며, 그림 3을 통해 대학 평판도 순위의 추이 변화를 확인할 수 있다.

표 6. 실험에 따른 대학 평판도 순위 비교 : 순위 순서
Table 6. Comparison of university reputation ranking by experiment methods: ranking order

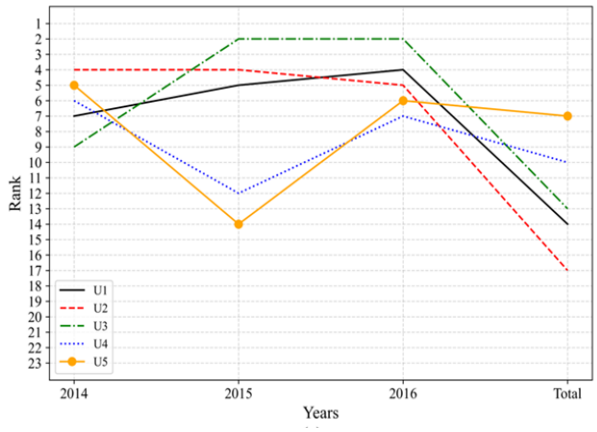
Methods / Ranking	Previous (‘16)	Proposed (‘16)	Proposed (‘14~’16)
1	U1	U9	U21
2	U7	U3	U19
3	U11	U6	U10
4	U2	U1	U17
5	U3	U2	U8
6	U8	U5	U6
7	U4	U4	U5
8	U6	U14	U11
9	U16	U16	U9
10	U22	U10	U4
11	U13	U11	U14
12	U15	U8	U22
13	U19	U17	U3
14	U5	U21	U1
15	U12	U20	U13
16	U14	U7	U7
17	U10	U13	U2
18	U17	U18	U15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

표 6 및 7과 같이 이 논문에서 제안한 대학 평판도 순위와 기존 연구의 대학평가 순위의 비교 결과를 통해 단일 연도는 물론 종합 연도에서도 차이가 나타남을 알 수 있다. 이는 단순히 데이터의 크기가 대학 평판도에 대해 크게 영향을 미치지 않는 것을 알 수 있다. 데이터가 많을수록 각 대학에 관한 뉴스 기사의 편차 또한 커지기 때문에 데이터의 크기는 대학 평판도에 영향이 적을 수 있다. 표 3과 5를 통해 특정 주제에 대한 대학의 긍정·부정 양상이 어느 한쪽으로 치우쳐 있지 않고, 데이터의 크기와 관계없이 다양하게 나타난다.

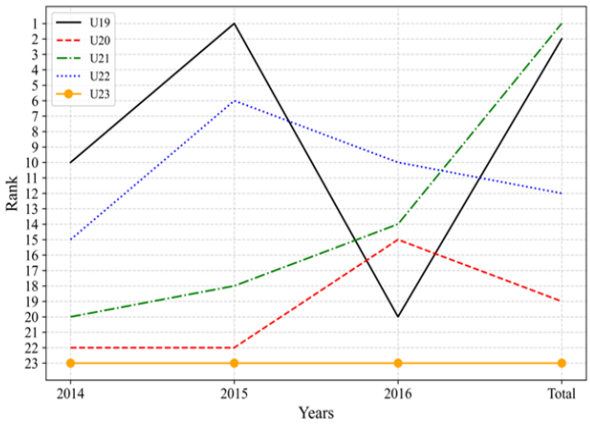
그리고 그림 3과 같이 연도별 대학평판의 순위 변동이 있는 것으로 보아 뉴스 매체의 특성상 기관에 따라 평가에 차이가 있고 데이터의 양도 다르므로 데이터 크기로 인한 분석 결과를 신뢰하기 어렵다는 것을 알 수 있다.

표 7. 실험에 따른 대학 평판도 순위 비교 : 대학 순서
Table 7. Comparison of university reputation ranking by experiment methods : University order

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



(a)



(b)

그림 3. 연도별 대학의 평판도 추이 결과
Fig. 3. Results of university reputation trend by year

V. 결론 및 향후 연구

이 논문은 대학 평판도 측정과 순위의 추이 분석을 위해 정성적인 데이터 크기에 따른 평판도 변화 실험을 제안하였다. 대학의 평판도 평가 및 추이 분석을 위해 연도별 온라인 뉴스 기사로부터 23개 대학의 데이터를 수집하였다. 감성 문단 도출을 위해 수집한 데이터로부터 감성 어휘를 추출하여 기존 AR-KNU 감성 사전 기반의 확장된 AR-KNU+ 감성 사전을 구축하였다.

그리고 각 대학의 공통적인 주제 도출을 위해 감성 문단의 개수가 많은 상위 3개의 대학을 기반으로 군집화를 수행하였으며, 추출된 군집에 대해 주제를 레이블링하였다. 그 다음 나머지 20개 대학에 대한 주제 분류를 위해 다중 클래스 분류 모델을 적용하였다. 마지막으로 감성 문단을 통해 주제별로 분류된 각 대학의 긍정·부정 비율을 측정하여 종합적인 평판도를 도출하고, 평판 순위에 대해 추이 분석을 수행하였다.

제안 방법을 통해 주제별 대학평판의 긍정·부정 비율을 확인하고, 종합적인 대학의 평판도를 기계적으로 도출하였다. 이는 기존의 설문 조사 기반 대학 평가의 문제점을 개선할 수 있을 것이다. 또한, 데이터 크기에 따른 대학 평판도 측정 실험을 수행하여 데이터 크기와 대학 평판도의 관계를 알 수 있으며, 연도별 대학 평판도를 측정하여 순위 변동 추이를 확인할 수 있다. 이를 통해 관심 있는 대학의 주제별 평판을 알 수 있으며, 각 대학은 평판 개선을 위한 기초 자료로 사용할 수 있다. 그리고 제안 방법을 기반으로 다양한 도메인의 감성 분석 및 여론 조사에 활용할 수 있을 것이다.

이 논문에서는 대학의 종합적인 평판도 측정을 위해 실험 데이터로 온라인 뉴스 기사를 활용하였다. 뉴스 매체의 특성으로 보아 기관에 따른 평가의 차이 및 데이터 크기의 차이가 존재하므로 대학의 평판 양상이 편향되는 문제점을 발생시킬 수 있다. 또한, 하나의 뉴스 기사는 다양한 대학에 관한 내용을 다룰 수 있으므로 특정 대학의 평판이 다른 대학에 영향을 미칠 수 있는 문제가 있다. 이러한 문제점들로 인해 대학의 주제를 분류하는 군집화 기법의 결과가 달라질 수 있다는 한계점이 존재한다.

따라서 이 논문은 향후 연구로서 뉴스 기사뿐만 아니라 SNS와 같은 다양한 매체의 데이터를 활용하여 한계점을 개선할 것이다. 그리고 단순히 데이터의 크기만을 늘리는 것이 아니라 전처리 작업을 통해 신뢰할 수 있는 데이터를 구축하여 대학 평판도를 도출할 것이다. 또한, 대학의 주제 추출을 위해 평균 및 분산 기반의 군집화 기법과 확률 분포 기반의 토픽 모델링 알고리즘을 활용하여 주제별 대학평판에 대한 비교 평가 실험을 수행할 것이다.

References

[1] J. W. Byun and J. H. Kim, "Market Definition and Market Dominance of the On-Line Two-Sided Transaction Platforms", Journal of Korean

Competition Law, Vol. 37, pp. 119-147, May 2018.

[2] W. J. Choi, "Hate expression and responsibility of platform service operators spreading through online video services", Law Research Institute Chonbuk National University, Vol. 62, pp. 125-150, May 2020.

[3] W. Shang and S. K. An, "The Effect of Online News Use Motivation on Acceptance and Satisfaction A Comparative Study on Korean and Chinese University Students", Journal of The Korea Contents Association, Vol. 20, No. 6, pp. 293-311, Jun. 2020.

[4] H. Y. Jeon, "Analysis of Employment Trends of Disabilities through Big Data: Focusing on SNS and Online News", Korea Employment Development Institute, Vol. 29, No. 2, pp. 55-82, May 2019.

[5] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S. F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis", Image and Vision Computing, Vol. 65, pp. 3-14, Aug. 2017.

[6] Z. Li, Y. Fan, B. Jiang, T. Lei, and W. Liu, "A survey on sentiment analysis and opinion mining for social multimedia", Multimedia Tools and Applications, Vol. 78, No. 6, pp. 6939-6967, Aug. 2018.

[7] Joongang, "University reputation evaluation", http://univ.joongang.co.kr/. [accessed: Jun. 04, 2020]

[8] S. W. Kim, B. L. Cho, and S. P. Han, "The Effects of University Evaluation on Its Image Transformation", Journal of Communication Science, Vol. 10, No. 2, pp. 139-178, Jun. 2010.

[9] A. Brewer and J. Zhao, "The impact of a pathway college on reputation and brand awareness for its affiliated university in Sydney", International Journal of Educational Management, Vol. 24, No. 1, pp. 34-47, Jan. 2010.

[10] E. A. Kim and Y. S. Lee, "The College Reputation System using Public Data and Sentiment Analysis", Journal of Information and Security, Vol. 18, No. 1, pp. 103-110, Mar. 2018.

[11] Y. K. Kim and W. S. Kang, "A Study of Special-purpose Academy Image Positioning Strategy Utilizing Big Data Analysis", Journal of Social Science, Vol. 35, No. 1, pp. 33-70, Jun. 2018.

[12] 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.

[13] F. Tian, Q. Zhou, and C. Yang, "Gaussian mixture model-hidden Markov model based nonlinear equalizer for optical fiber transmission", Optics Express, Vol. 28, No. 7, pp. 9278-9737, Mar. 2020.

저자소개

채 수 현 (Soohyeon Chae)

2019년 : 군산대학교 소프트웨어융합공학과(학사)
2019년 ~ 현재 : 군산대학교 소프트웨어융합공학과(석사과정)
관심분야 : 자연어 처리, 텍스트 마이닝, 빅데이터 분석, 지식 그래프

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

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, 정보검색, 빅데이터, 인공지능

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 : 텍스트 마이닝, 빅데이터 분석, 메타데이터, 공공데이터, 지식 그래프