Journal of KIIT. Vol. 17, No. 3, pp. 11-21, Mar. 31, 2019. pISSN 1598-8619, eISSN 2093-7571 11
http://dx.doi.org/10.14801/jkiit.2019.17.3.11

온라인 뉴스 데이터를 활용한 AR-KNU 감성사전 기반의 대학 평판도 평가

박상민*[1], 엄창민*[2], 온병원**[1], 정동원**[2]

An AR-KNU Sentiment Lexicon-based University Reputation Assessment Using Online News Data

Sang-Min Park*[1], Chang-Min Eom*[2], Byung-Won On**[1], and Dongwon Jeong**[2]

이 논문은 2016년도 정부(미래창조과학부)와 2017년 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2016R1A2B1014843, NRF-2017M3C4A7068188)

요 약

지금까지의 대학 평판도 평가는 설문 조사를 통한 평판도 분석의 한계, 평판도 분석에 활용된 공공 데이터와 텍스트 데이터의 기간적 불일치 그리고 심층적인 분석 결과의 부재라는 문제점을 지닌다. 이 논문에서는 감성 분석 기반의 새로운 대학 평판도 평가 알고리즘을 제안하고 기존 연구와의 비교 분석을 수행한다. 이러한 목표 달성을 위해 감성 사전을 구축하고 감성 문단 및 군집을 추출한다. 정제된 군집 생성을 위해 추출된 군집에 대한 군집화, 군집 정제 연산을 수행하고, 최종적인 대학 평판도 평가 결과를 생성한다. 이 연구의 결과는 구체적인 대학의 이슈 발견에 활용될 수 있으며, 정확도가 향상된 대학 평판도 평가의 새로운 알고리즘 개발에 기여할 수 있다. 또한 다양한 평판도 평가 분야에서 유의미한 기초 자료로 활용될 수 있다.

Abstract

Although there have been various research efforts to evaluate the reputation of universities so far, previous studies are limited because of the inherent limitation, temporal inconsistency, and restricted insight discovery. This paper presents a new sentiment analysis-based university reputation evaluation algorithm that is verified through a comparative study with existing methods. To achieve the goal, a sentiment lexicon is manually constructed in the first step and the proposed method extracts sentiment sentences and clusters in the next steps. Specifically, the proposed method first generates fine-grained clusters through clustering, cluster merging, cluster labelling, and then evaluates the reputation results of universities. Finally, the paper shows the comparison between the existing and proposed methods. The outcome of this research can be used to identify specific university insights and contribute to the development of advanced algorithms with high accuracy for evaluating the reputation of universities. In addition, this study can be used as meaningful basic data to address a wider range of issues including university reputation assessment.

Keywords
sentiment lexicon, university reputation mining, sentiment analysis, text mining

* 군산대학교 소프트웨어융합공학과
- ORCID[1]: https://orcid.org/0000-0003-4730-1997
- ORCID[2]: https://orcid.org/0000-0003-1463-4111
** 군산대학교 소프트웨어융합공학과 교수(교신저자)
- ORCID[1]: https://orcid.org/0000-0001-6929-3188
- ORCID[2]: https://orcid.org/0000-0001-9881-5336

· Received: Nov. 21, 2018, Revised: Mar. 03, 2019, Accepted: Mar. 06, 2019
· Corresponding Author: Byung-Won On, Dongwon Jeong
Dept. of Software Convergence Engineering, Kunsan National University,
558, Daehak-ro, Gunsan, Jeollabuk-do, Korea,
Tel.: +82-63-469-8911, Email: {bwon, djeong}@kunsan.ac.kr

Ⅰ. 서 론

가트너는 '빅데이터에 대한 시장의 관심이 큰 가운데 새로운 데이터 소스로부터 통찰력을 추출, 활용하고 이를 기반으로 행동하는 것의 잠재적 가치가 증대되었다'라고 언급하였다[1]. 현재 빅데이터에서 통찰력을 추출하고 활용하기 위한 연구가 활발하게 진행되고 있으며 특히 비정형 데이터의 폭발적인 증가와 함께 이러한 비정형 데이터에서 의미 있는 정보를 추출하기 위한 텍스트 마이닝(Text Mining)의 중요성이 대두되고 있다. 텍스트 마이닝은 텍스트 데이터를 분석하고 의미 있는 정보를 추출하는 마이닝 기법이다[2]. 비정형 데이터에서 필요한 정보를 추출하기 위해 다양한 텍스트 마이닝 기법이 활용되고 있으며 그 중에서도 텍스트에서 작성자의 감성을 분석하기 위한 감성 분(Sentiment Analysis) 기법이 시장조사, 마케팅 그리고 여론조사 등 다양한 분야에서 중요하게 활용되고 있다[3].

온라인상의 데이터는 대중의 관심이 바로 투영되어 나타난다. 이와 같은 이유로 온라인상의 반정형 또는 비정형 데이터를 통해 기업이나 조직을 평가 및 분석하여 대중의 여론을 조사하는 연구가 진행되고 있다. 이러한 연구 동향은 대학 평판도를 평가하는 분야에서도 동일한 추세를 보인다. 즉, 일반적으로 널리 활용되던 설문조사 방식을 이용한 대학 평판도 평가의 문제점을 해결하기 위해 텍스트 마이닝 기법이 활용되고 있다.

대학 평판도 평가를 위한 방식은 크게 설문조사 기반 평가 방법과 텍스트 마이닝 기반 평가 방법으로 분류할 수 있다. 먼저, 설문조사 기반 대학 평판도 평가의 대표적인 사례는 중앙일보 대학 평판도 평가이다[4]. 중앙일보에서는 매년 대학 종합 평가를 실시하고 있으며 이를 위해 설문조사를 수행하고 이를 토대로 대학 평판도를 평가해왔다. 그러나 설문조사 방식에 기반한 평가 방법은 몇 가지 한계를 지닌다. 우선 설문 표본의 크기가 작고 응답률이 높지 않으며 응답자가 설문에 성실히 응답하지 않을 경우 모집단 추정에 있어 신뢰할 수 없다. 또한 설문조사 결과를 종합하고 분석하기 위해 많은 시간과 높은 비용을 필요로 하며 결과 분석에 있어 분석가의 주관이 투영될 수 있다는 문제점이 있다.

설문조사 기반 대학 평판도 평가 방법의 문제점을 해결하기 위해 텍스트 마이닝 기법을 이용한 평가 방법이 연구되었다[5][6]. 기존 연구 중에서, [5]는 감성 분석을 이용한 대학 평판도 평가 방법을 제안한다. 이 연구는 감성 분석 기법을 통한 정성적인 평가와 함께 정량 데이터를 이용한 정량 평가를 종합하고 활용하여 대학 평판도를 평가한다. 즉, '대학 공공알리미'에서 제공하는 정량 데이터를 이용한 정량적인 평가 결과와 '네이버 뉴스' 데이터를 이용하여 감성 분석 연산을 수행한 정성적인 평가 결과를 혼합하여 대학 평판도 평가를 수행한다. 그러나 이 연구는 대학 평판도 평가를 위해 사용된 정량 데이터와 정성 데이터의 기간이 불일치한다는 문제점과 함께 감성사전의 단어 개수가 충분하지 않으며 구체적이고 다양한 분석 내용이 제공되지 않는다는 한계를 지닌다.

무엇보다 앞서 기술한 기존 연구 결과는 실제 사람들이 느끼는 평판도와 차이가 크다는 공통적인 문제점을 지닌다. 순위의 정확성 및 일치성을 차치하고라도 일정한 형태의 패턴을 보이지 않는다.

따라서 이 논문에서는 각 대학의 평판도를 객관적이고 신속하게 평가하기 위한 알고리즘을 제안한다. 특히 제안 알고리즘과 기존 알고리즘의 비교를 통해 각각의 유의미성에 대한 분석을 수행한다.

이 연구의 목적을 달성하기 위해 우선 네이버 뉴스의 사회-교육면 기사를 자동으로 수집한다. 수집된 데이터를 키워드를 통해 23개의 대학으로 분류하여 형태소 분석을 수행한다. 형태소 분석 후 보다 정확한 감성 분석을 위해 KNU 한국어 감성사전(KNU Sentiment Lexicon)을 활용하여 대학 도메인에 적합한 감성사전을 구축한다[7]. 구축된 감성사전을 통해 23개 대학에 대하여 감성 분석을 수행하고 감성 분석을 통해 추출된 감성문단의 수가 가장 많은 상위 3개의 대학을 대상으로 군집화 및 군집 분석을 통해 주제를 추출한다. 주제의 예로 '연구·교육', '입학' 그리고 '학생활동' 등을 둘 수 있으며, 이는 각 대학에 관한 뉴스 기사에서 공통적으로 다뤄지는 주제를 의미한다. 추출된 주제에 따라 각 대학 뉴스 기사들을 자동으로 분류하고 평판도 평가를 수행하여 긍정·부정 비율을 보여준다. 마지막으로, 제안 방법에 의해 도출된 23개 대학의 종합 평판도

평가 결과를 보이고 기존 연구와 제안 연구와의 비교 평가를 수행한다.

이 논문의 구성은 다음과 같다. 제2장에서는 기존의 대학 평판도 평가 연구를 소개하고 주요 문제점을 정의한다. 제3장에서는 제안하는 대학 평판도 평가 알고리즘을 보이고 제4장에서는 실험환경, 결과 및 비교 분석 결과를 서술한다. 마지막으로, 제5장에서는 결론 및 향후 연구를 기술한다.

II. 관련 연구

이 장에서는 대학 평판도 평가를 위한 기존 연구를 소개하고 기존 연구가 지니는 주요 문제점을 기술한다. 이 논문에서 제안하는 대학 평판도 평가 알고리즘은 'KNU 한국어 감성사전'을 활용하여 대학이라는 도메인에 알맞은 감성사전을 구축하여 활용한다. 따라서 이 장에서는 'KNU 한국어 감성사전'에 대하여 간략하게 소개한다.

대학 평판도 평가에 대한 연구는 크게 설문조사 기반 평가 방법과 텍스트 마이닝 기반 평가 방법으로 분류된다. 이 절에서는 대학 평판도 평가와 관련된 기존 연구를 앞서 기술한 두 가지 분류에 따라 기술한다.

먼저 설문조사 기반 대학 평판도 평가의 대표적인 사례는 중앙일보에서 실시한 대학 평판도 조사이다[4]. 중앙일보는 대학 평판도 평가를 위해 각 대학에 대한 설문조사를 수행한 뒤 이를 바탕으로 대학의 평판도 결과를 발표한다. 중앙일보의 주요 설문조사 내용인 설문 항목은 채용하고자 하는 신입사원의 대학 출신, 전공 또는 교양 교육이 제대로 되어있는 대학 그리고 발전 가능성이 있다고 판단되는 대학이다. 중앙일보에서 실시한 설문조사 기반 대학 평판도 평가 결과는 사회 구성원이 인식도를 반영한다는 측면에서 의미가 있다. 하지만 중앙일보의 대학 평판 설문조사는 단순하게 각 문항에 대해 응답자들이 순위를 나열해서 각 대학 순위를 나열한다는 점에서 대학 평판에 대한 세부적인 정보를 확인할 수 없다는 문제점이 있다. 또한 응답자들이 불성실하게 설문에 응답할 경우 설문조사를 통해 수집된 데이터가 질적으로 떨어질 수 있는 문제점이 있다.

중앙일보에서 실시한 대학 평판도 평가와 달리 텍스트 마이닝 기반의 대학 평판도 평가 방법에 대한 연구가 진행되어 왔다[5][6][8].

[6]는 트위터의 SNS 데이터를 활용하여 국내 대학에 대한 평판도를 분석한다. 이 연구에서 제안한 국내 대학 평판 분석 방법은 다음과 같다. 첫 번째, SNS 데이터 안에서 특정 대학명이 얼마나 언급되었는지 측정하고 언급된 수치가 높을수록 사람들이 해당 대학에 관심이 많다고 분석한다. 두 번째, 많이 언급된 상위 10개의 대학을 대상으로 연관 키워드를 추출한다. 연관 키워드는 빈도수가 높은 k개의 단어이며 이는 연구자에 의해 긍정·부정으로 레이블링이 수행된다. 마지막으로 각 대학의 키워드 감성 분석을 수행한다. 이 연구는 대학 평판도 분석을 위해 6주 동안의 SNS 데이터를 활용하였다. 이 연구의 분석 결과를 6주라는 기간에 편향된 분석 결과이며 이를 대학의 평판 결과로 제공하기에 한계가 있다. 또한 감성 분석은 문장 수준의 감성 분석이 아닌 키워드 수준의 감성 분석을 수행하였다는 점에서 세부적인 대학의 평판 내용을 알 수 없는 문제가 존재한다.

[8]은 대학이 생존하고 운영되기 위해서 일반 기업과 같이 마케팅 개념이 도입되어야 한다고 주장하며 대학 이미지 평가 측정 척도 7개를 제안한다. 제안방법은 다음과 같다. 우선 대학 이미지 평가를 위한 28개의 대학 이미지 평가 항목을 선정하고 전문가를 통해 21개로 축소한다. 설문분석, 주성분 분석(PCA, Principal Component Analysis), 베리맥스(Varimax) 회전을 적용하여 총 7개의 대학 이미지 평가 측정 척도를 제안한다. 이 연구는 실제 데이터에 의해 객관적으로 수행된 것이 아니라 사람의 설문을 통해 이미지 평가 측정 척도를 제안하였다는 점에서 주관적인 척도가 포함되어 있을 수 있는 문제가 존재한다.

[5]는 정량 데이터와 정성 데이터를 이용한 혼합형 대학 평판도 평가 방법을 제안한다. 이 연구는 대학 알리미의 공공데이터인 대학 평가 지표를 활용하여 정량적 지수를 추출하고 네이버 뉴스 기사를 활용하여 정성적 지수를 산출, 이를 종합하여 혼합 대학 평판 지수를 산출한다. 혼합 대학 평판 지수 산출을 위해 이 연구에서 제안한 구체적인 방안

은 다음과 같다. 우선 정량적 지수는 대학 알리미에서 제공하는 2015년부터 2017년 12월 31일까지의 데이터를 사용하여 산출한다. 두 번째 단계로 네이버의 사회-교육면의 2017년 12월, 2018년 1월의 뉴스 기사를 수집하여 키워드 별로 각 대학을 분류한다. 분류된 뉴스 기사에서 자주 출현되는 단어를 활용하여 감성사전을 자체적으로 구축하고 이를 통해 각 대학의 감성 분석을 수행하여 정성 지수를 산출한다. 마지막 단계로 산출된 정량 지수와 정성 지수를 혼합하여 각 대학의 평판 지수를 산출하고 평판 지수가 높은 대학부터 나열하여 보여준다. 하지만 이 연구에서는 정량적 지수 산출을 위해 사용한 데이터의 기간과 정성적 지수 산출을 위해 사용된 데이터의 기간이 불일치하며 이로 인해 평판도 분석 결과의 신뢰성을 담보할 수 없다. 대학 평판 지수는 종합적인 하나의 점수로 환산되기 때문에 각 대학의 세부적인 주제에 대한 평판은 확인할 수 없다는 문제점이 존재한다. 또한 30개의 감성 어휘로 구성된 감성사전을 사용한다는 점에서 적절한 감성 분석 결과를 도출하기 어렵다.

KNU 한국어 감성사전은 다양한 도메인에 공통적으로 사용될 수 있는 감성 어휘로 구성되어 있다. 각 도메인의 감성사전을 빠르게 구축하기 위한 기초 자료로 개발되었으며 1-gram, 2-gram, 어구(n-gram), 문형, 신조어, 이모티콘 형태의 감성 어휘로 구성되어 있다. 각 감성 어휘는 긍정·부정, 중립에 대한 감성으로 분류되어 있으며 각 감성에 대한 정도 값과 어근을 포함하고 있다[7].

구축 방법으로는 표준국어대사전을 구성하는 형용사, 부사, 동사, 명사의 모든 뜻풀이에 대한 긍정·부정, 중립을 Bi-LSTM(Bidirection Long Short-Term Memory) 딥러닝 모델을 통해 분류하고 긍정으로 분류된 뜻풀이에서는 긍정에 대한 감성 어휘를, 부정으로 분류된 뜻풀이에서는 부정에 대한 감성 어휘를 추출한다. 또한 다양한 외부 소스에서 감성 어휘를 추출하였으며 온라인 텍스트 데이터에서 주로 사용되는 신조어와 이모티콘에 대한 정보를 포함하고 있다.

이 논문에서는 대학 평판도 평가를 위해 KNU 한국어 감성사전을 활용하여 대학이라는 도메인에 사용되는 특정 감성 어휘를 추가하여 AR-KNU 감성사전을 구축한다.

군집화는 주어진 데이터들의 특성을 고려하여 군집(데이터 집단)을 정의하고 데이터 집단을 대표할 수 있는 대표점을 찾는 데이터 마이닝 기법이다[9]. 군집이란 비슷한 특성을 가진 데이터들의 집단이며 데이터의 특성이 다르면 다른 군집에 속해야 한다. 이 연구에서는 대학이 공통적으로 가지고 있는 군집에 대한 사전 지식이 없기 때문에 EM 군집화(Expectation-Maximization clustering) 기법을 사용하여 자동으로 군집을 찾아 군집화를 수행한다.

신경망(Neural Network)은 생물학적 뉴런과 유사한 구조를 띄며 입력층(Input Layer), 은닉층(Hidden Layer) 그리고 출력층(Output Layer)로 구성되어 있는 비순환 그래프이다[10]. 이와 같은 다중 클래스 신경망(Multiclass Neural Network)을 통해 이 연구에서 각 대학 감성 문단의 주제를 분류한다.

III. 제안 방안

이 논문에서는 앞서 기술한 기존 연구의 문제점을 개선하기 위해 감성 분석 기반의 새로운 대학 평판도 분석 알고리즘을 제안한다.

그림 1은 제안 방안의 개요를 나타내며 감성 분석, 주제 추출, 주제별 분류 그리고 평판도 분석 등으로 구성된다.

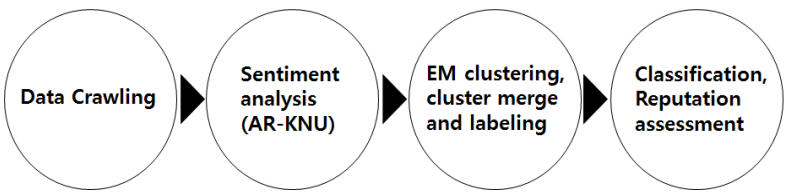

그림 1. 제안 방안의 전체적인 프로세스
Fig. 1. Overall process of proposed method

제안 방안은 다음과 같은 순서에 의해 수행된다. 첫 번째, 분석하고자 하는 대학 뉴스 기사를 수집하고 AR-KNU 감성사전을 구축하여 문단 단위의 감성 분석을 수행한다.

두 번째, 감성 분석에 의해 추출된 감성 문단의 개수가 많은 상위 k개의 대학은 다양한 주제(이슈)를 가지고 있는 대학일 가능성이 높다. 대학에 관한 다양한 주제를 추출하기 위해 주제가 많이 포함되어 있을 것으로 예상되는 상위 k개 대학의 감성 문단으로 EM 군집화를 수행하여 유사한 문단으로 이루어진 군집을 추출한다. 추출된 군집들의 상위 k개의 단어를 통해 유사한 군집들을 찾아내고 이를 하나의 군집으로 합병한다. 이와 같은 방안으로 추출된 각 군집은 군집을 이루는 유사한 문단들이 어떠한 주제를 의미 하는지 알 수 있도록 레이블링을 수행하여 군집명을 부여한다.

세 번째, 추출된 각 군집명을 정답으로, 군집에 속해있는 문단들을 입력으로 하는 학습 데이터를 구축하고 구축된 학습 데이터를 활용하여 다중 클래스 신경망을 학습, 나머지 대학의 주제 분류를 수행한다. 분류된 각 대학의 긍정·부정 비율을 통해 특정 레이블에서 각 대학의 평판이 얼마나 긍정·부정의 성향을 띄는지 보여준다. 또한 대학의 종합 평판도 순위를 추출하여 보여준다.

3.1 감성사전 구축 & 대학별 감성 분석

이 논문에서는 감성사전 기반의 감성 분석을 위해 군산대에서 구축한 KNU 한국어 감성사전을 활용하여 자체적으로 감성사전을 구축한다. KNU 한국어 감성사전은 도메인에 독립적인 감성 단어를 정의한 사전으로 대학이라는 도메인에 의존적인 감성 단어들이 포함되어 있지 않을 가능성이 높다.

그렇기 때문에 대학이라는 도메인에 의존적인 새로운 감성사전의 구축이 필요하다. 새로운 감성사전 구축을 위해 KNU 한국어 감성사전을 활용하여 대학 뉴스 기사에 포함되어있는 도메인에 독립적인 감성 어휘를 추출하고 감성 레이블링을 수행한다. 그 후에 대학이라는 도메인에서 의존적인 감성 어휘 확장을 위해 형용사, 명사, 부사 그리고 동사를 품사로 갖는 어휘들 중 감성 레이블링이 수행되지 않은 어휘에 대해 3명의 평가자가 수작업으로 감성 레이블링을 투표와 토론을 통해 수행한다.

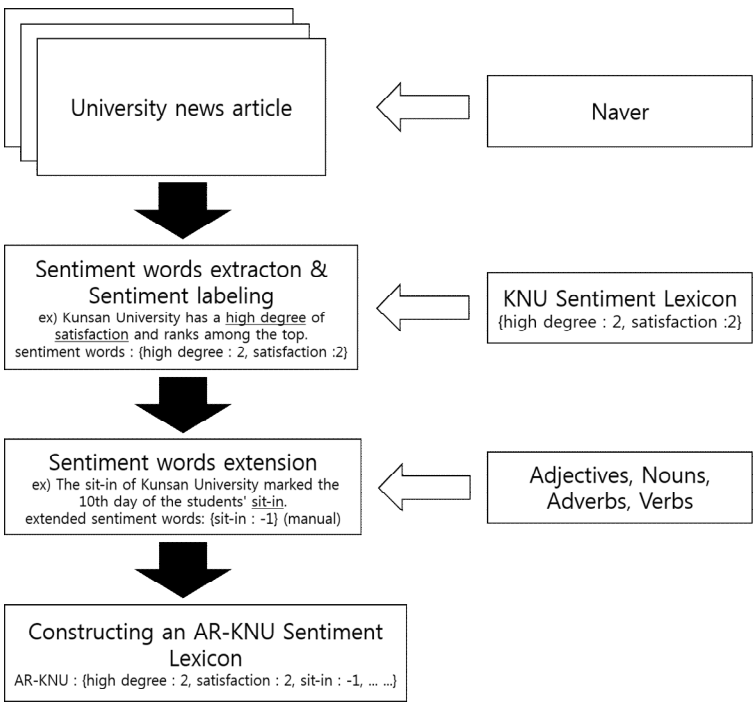

그림 2. AR-KNU 감성사전 구축 방안

Fig. 2. Method of construct AR-KNU sentiment lexicon

이와 같은 방안으로 구축한 감성 어휘들을 통해 Academic Reputation(AR)-KNU 감성사전)을 구축한다. 그림 2는 AR-KNU 감성사전 구축 방안에 대해 나타낸다. 표 1은 감성 어휘를 표현하는 감성 정도 값을 나타낸다.

표 1. 감성에 따른 감성 정도 값
Table 1. Sentiment degree by sentiment

Degree	-2	-1	0	1	2
Sentiment	Ver neg	Neg	Obj	Pos	Very pos

이와 같은 방법을 통해 구축한 AR-KNU 감성사전을 사용하여 각 대학 뉴스 기사 문단에 대하여 감성 분석을 수행하며, 감성을 지니는 문단들을 향후 실험을 위한 후보 문단으로 추출한다. 감성 분석은 다음과 같은 과정을 통해 측정된다. 첫 번째, 특정 문단을 구성하는 단어의 집합 중 감성사전에 포함된 단어가 있으면 해당 단어에 대한 감성 값을 모두 더한다. 두 번째, 특정 문단의 스코어가 양수인 경우 이를 긍정 문단으로, 음수인 경우 이를 부정 문단으로 분류한다.

3.2 주제 추출

이 절에서는 대학이 공통적으로 갖는 주제를 추출하기 위한 기법을 제안한다. 주제를 추출하기 위한 제안 방안은 다음과 같다. 3.1절에서 제안한 대학 뉴스 기사 문단의 감성 분석을 결과를 통해 추출된 감성 문단의 개수가 많은 상위 k개 대학을 군집화 수행을 위한 대학으로 선정한다. 감성 문단의 개수가 많은 상위 k개 대학은 감성 문단이 많기 때문에 다양한 주제(이슈)를 포함하고 있을 가능성이 높다. 이 연구에서는 대학에 대한 다양한 주제 추출을 위해 이와 같이 다양한 주제를 포함하고 있을 가능성이 높은 상위 k개 대학을 대상으로 EM 군집화를 수행하여 유사한 문단으로 이루어진 군집을 추출한다. 군집들의 중복을 방지하기 위해 추출된 군집들의 상위 k개의 단어들을 확인하고 이를 통해 유사한 군집들을 하나의 군집으로 합병을 수행한다. 이와 같은 과정을 통해 최종적으로 n개의 군집을 추출하고 군집을 대표할 수 있는 군집명을 레이블링 해준다. 그리고 이와 같은 군집명을 주제라고 한다.

3.3 대학별 주제 분류

이 절에서는 문단을 입력으로 하고 문단의 주제(군집명)를 정답으로 하는 학습 데이터를 구축하여 다중 클래스 신경망을 학습한다. 학습된 다중 클래스 신경망 모델을 통해 주제 추출 수행에 활용된 상위 k개 대학을 제외한 나머지 대학에 대해 주제 분류를 수행한다.

입력층은 100개의 노드로 구성되어 있으며 문단을 표현하는 100차원의 벡터 값이 입력으로 들어간다. 은닉층은 50개의 노드로 구성되어 있으며 출력층은 분류할 주제의 개수인 3개로 구성되어 있다.

3.4 주제별 대학 평판도 긍정·부정 비율 측정

이 연구에서는 각 대학의 주제별 평판도를 긍정·부정 비율로 표현한다. 주제별 평판도 측정을 통해 특정 대학의 긍정·부정 비율을 보여주면 해당 주제에 대한 평판이 어떠한 양상을 보이는지 세부적으로 알 수 있다. 분류된 각 대학의 주제별 평판도 긍정·부정 비율은 다음과 같은 과정을 통해 연산된다. 첫 번째, 각 대학의 문단을 분류된 주제에 따라 나눈다. 두 번째, 3.1절을 통해 추출된 각 문단의 극성을 통해 분류된 주제별 긍정·부정의 비율을 구한다. 세 번째, 각 주제에 대한 긍정·부정의 비율은 해당 주제에 대하여 어떠한 비율로 긍정적으로 또는 부정적으로 언급되는지에 대한 결과이다.

3.5 대학 종합 평판도 순위 측정

이 절에서는 대학 종합 평판도 순위를 측정하는 기법에 대하여 기술한다. 대학 종합 평판도 순위 측정을 통해 각 대학의 종합적인 평판도 순위를 쉽게 알 수 있다. 대학의 종합 평판도 순위를 측정하는 알고리즘은 그림 3과 같다. 줄번호 6~13은 각 대학의 긍정·부정 평판 스코어와 문단의 개수를 추출하는 과정이다. 줄번호 14~15는 각 대학의 종합 평판 스코어를 구하는 과정으로 대학의 긍정·부정 평판 스코어를 긍정·부정 문단의 개수로 나누어 구한다.

algorithm1: Measurement of the ranking of the university's comprehensive reputation

Input : S_{ui}, $S_{ui_{score}}$
Output : S_{up}, S_{un}, S_u

1: S_{ui}: the ith sentiment paragraph of u University
2: $S_{ui_{score}}$: the score of the ith sentiment paragraph of u university
3: S_{up}: u university's positive score
4: S_{un}: u university's negative score
5: S_u: u university's comprehensive reputation score
6: for u in U
7: for i in S:
8: if $S_{ui_{score}} > 0$:
9: $pos_{score} = pos_{score} + S_{ui_{score}}$
10: pos_{num}++
11: else if $S_{ui_{score}} < 0$:
12: $neg_{score} = neg_{score} + S_{ui_{score}}$
13: neg_{num}++
14: $S_{up}=pos_{score} \div pos_{num}$; $S_{un}=neg_{score} \div neg_{num}$;
15: $S_u = S_{up}-S_{un}$
16: Descending sort to S_u

그림 3. 대학 종합 평판도 순위 측정 순서도
Fig. 3. Flowchart of ranking measurement of comprehensive university reputation

IV. 실험 및 평가

4.1 실험 환경

이 논문에서는 네이버의 사회-교육면의 2016년 10월부터 2017년 9월 사이의 뉴스 기사를 대상으로 실험을 수행한다. 뉴스기사는 23개의 대학 키워드를 통해 분류하고, 파이썬 패키지인 Konlpy[11]의 트위터 형태소 분석기를 활용하여 형태소 분석을 수행한다. 감성분석은 ubuntu 16.04에서 파이썬 프로그래밍 언어를 통해 수행한다, 클러스터링을 수행하기 위해서는 오픈 소스인 Weka 3.8[12]을 사용하며 EM 알고리즘을 통해 데이터를 분류한다. 다중클래스 신경망은 azure ml studio[13]에서 수행한다. 또한 fasttext[14]의 skipgram을 활용하여 워드 임베딩(Word Embedding)을 수행하여 하나의 문단을 100개의 벡터 값으로 표현한다.

4.2 실험결과

4.2.1 주제 추출 결과

추출된 감성 문단의 개수가 많은 대학은 다양한 이슈와 주제가 포함되어 있을 것이다. 이 절에서는 대학이 다룰 수 있는 다양한 주제를 추출하기 위해 3.1절에서 기술한 AR-KNU 기반의 감성 분석을 통해 추출된 각 대학의 모든 감성 문단을 추출하고 감성 문단의 개수가 많은 상위 3개의 대학인 U1, U2, U3에 대하여 군집화를 수행하여 주제를 추출한다.

U1, U2, U3를 통해 추출된 후보 문단에 대하여 군집화 수행 결과, 각 대학별로 15, 13, 11개의 군집이 추출되었다. 추출된 군집에 대한 합병을 수행한 결과 공통적으로 3개의 군집이 추출되었다. 공통적으로 추출된 3개의 군집에 대하여 각 군집이 무엇을 대표하는지에 대해 상위 k개 단어를 통해 레이블링을 수행하였다. 표 2는 각 군집을 레이블링 한 결과를 보여준다. 이 논문에서는 각 군집의 레이블을 그 군집을 대표하는 주제로 사용하며, 그 결과 대학에 관한 뉴스 기사에서는 대표적으로 '연구·교육', '입학' 그리고 '학생 활동'이라는 3가지의 주제가 다뤄지는 것으로 알 수 있다.

표 2. 주제 레이블링 결과
Table 2. Topic labeling result

Topic 1	Topic 2	Topic 3
Research education	Entrance	Student activities

표 3은 U1, U2, U3에 대한 군집화 결과를 보여준다. 이를 통해, 각 대학이 3가지 주제에 대한 긍정·부정 양상을 알 수 있다. 특히 U2와 U3의 입학 부분에 대하여 부정이 긍정보다 현저히 높은 것을 볼 수 있으며, 이는 해당 입학이라는 주제를 확인한 결과, 입학이라는 주제는 단순히 입학만을 나타내는 것이 아니라 등록금이라는 세부 분야도 포함되어 있으며 U2와 U3는 등록금이라는 세부 주제에 있어 부정적인 양상이 높은 것을 데이터를 통해 알 수 있었다. 표 4는 U2와 U3에 각각 대한 입학에 관한 부정적인 견해에 대한 대표적인 예를 보여준다.

표 3. U1, U2, U3 주제 추출 결과
Table 3. Topic labeling result of U1, U2, U3

Topic	Univ	Num	Pos	Neg
Research education	U1	890	69%	31%
	U2	538	72%	28%
	U3	504	67%	33%
Entrance	U1	248	29%	71%
	U2	59	31%	69%
	U3	99	25%	75%
Student activities	U1	368	72%	28%
	U2	173	76%	24%
	U3	174	72%	28%

표 4. 입학-등록금에 관한 부정적인 견해
Table 4. Negative opinion about entrance into a school-tuition

Univ	Content
U2	Which university has the most expensive tuition?...(omission)...In engineering, U2 schools cost 9.68 million won
U3	The average four-year college tuition is 6.68 million won... U3 9.01 million won is the most expensive.

4.2.2 주제 분류

(1) 분류 모델 검증

학습된 모델의 정확도를 평가하기 위해서 학습 데이터의 10%에 대하여 검증 데이터(Validation Set)를 구축하여 모델 평가에 사용한다. 표 5는 학습된 정확도(Accuracy), 정밀도(Precision), 재현율(Recall)을 나타낸다. 이를 통해 학습된 모델이 문서를 약 89% 정확도로 분류한다는 것을 알 수 있다.

표 5. 모델 검증 결과
Table 5. Model verification results

Accuracy	Precision	Recall
0.895981	0.843972	0.84972

(2) 주제 분류 및 긍정·부정 비율

표 6은 23개 대학 중 U1, U2, U3를 제외한 감성 문단의 개수가 많은 상위 10개 대학의 연구 및 교육, 입학, 학생 활동 대한 주제 별 및 긍정·부정 비율을 나타낸 결과이다.

표 6. 상위 10개 대학 주제별 분석 결과
Table 6. Top-10 university analysis result by topic

Topic	Univ	Num	Pos	Neg
Research education	U5	335	84%	16%
	U9	158	65%	35%
	U4	230	77%	23%
	U14	105	92%	8%
	U16	82	82%	18%
	U15	230	77%	23%
	U8	248	77%	23%
	U6	313	84%	16%
	U22	124	75%	25%
	U13	139	82%	18%
Entrance	335	84%	16%	71%
	U9	158	65%	35%
	U4	230	77%	23%
	U14	105	92%	8%
	U16	82	82%	18%
	U15	230	77%	23%
	U8	248	77%	23%
	U6	313	84%	16%
	U22	124	75%	25%
	U13	139	82%	18%
Student activities	U5	28	29%	71%
	U9	24	33%	67%
	U4	20	25%	75%
	U14	4	75%	25%
	U16	3	33%	67%
	U15	20	25%	75%
	U8	18	56%	44%
	U6	16	56%	44%
	U22	3	0%	100%
	U13	17	71%	29%

긍정·부정에 대한 비율은 3.1절 감성 분석에 의해 긍정·부정으로 분류된 문단의 개수를 통해 표현되었다. 이를 통해 각 대학이 특정 주제에 있어 어떤 긍정·부정의 양상을 보이는지 자세히 알 수 있다. 이와 같은 주제별 대학 평판도 조사를 통해 나온 결과는 각 대학에서 대학의 평판도 개선 방향을 잡을 수 있는 기초 자료로 사용 가능할 것이다.

4.3 정성 평가

표 7은 기존 연구와 이 논문에서 제안한 방법에 대한 정성 비교 평가를 보여준다. 기존 연구[5]는 대학 평판도 평가를 위해 정량 데이터와 정성 데이터를 혼합하여 활용하였다. 하지만 평가 대상 대학

의 정량 데이터와 정성 데이터의 기간 불일치성이 존재한다. 이는 분석 결과를 신뢰할 수 없다는 문제점을 발생시킨다. 기존 연구와 중앙일보는 대학 평판도 평가에 대하여 하나의 결과만을 보여주며 각 대학에 대한 긍정·부정 비율은 보여주지 않는다. 그렇기 때문에 각 대학이 어떤 주제에 대하여 어느 정도의 평판도를 보이는지 알 수 없다. 중앙일보는 대학 평판도를 평가하기 위해 단순하게 설문조사를 수행한다. 이는 서론에서 언급한 설문조사의 문제점을 발생시킬 수 있다.

제안 방안은 각 대학의 뉴스 기사를 유사한 문서들로 군집화 하여 주제를 추출하였다. 추출된 주제를 토대로 각 대학이 특정 주제에 대해 어느 정도의 여론을 가지고 있는지 추출하는 선호도 조사를 수행하였다. 이는 각 대학의 특정 주제에 대한 평판의 선호도를 확인할 수 있으며 다양한 측면으로 대학을 평가할 수 있다.

표 7. 기존 방안과 제안 방안 정성적 비교
Table 7. Qualitative comparison of based method and proposed method

Item	Baseline[5]	Joongang	Ours
Period consistency	Discord	Accord	Accord
Topic detection	Unsupport	Support (manual)	Support (semi automatic)
Topic sentiment rate	Unsupport	Unsupport	Support

4.4 정량 평가

표 8은 이 논문에서 제안한 방안을 통해 추출된 각 대학의 종합 평판도 순위를 군산대학교 50명의 대학생의 설문을 통해 추출된 각 대학의 종합 평판도 순위, 중앙일보(기존연구1)의 2016년도 대학평가 평판도 항목 순위 그리고 기존 연구(기존연구2)에서 추출된 대학의 평판도 순위를 비교한 것이다. 이 논문에서 실시한 설문은 기존연구(기존연구2)의 23개 대학을 대상으로 평판도 순위를 측정하였다. 중앙일보 평가 결과의 경우, 2017년도 평판도 순위 데이터를 구할 수 없어 2016년도의 데이터를 활용하였으며 이 연구에서 활용된 대학과 중복된 대학만을 순위에 나열하였다.

표 8. 제안 방안과 설문조사에 대한 대학 평판도 추출
Table 8. Quantitative comparison of based method and proposed method

Rank	Ours	Survey	Baseline1	Baseline2
1	U20	U1	U1	U1
2	U12	U2	U3	U6
3	U23	U3	U4	U20
4	U10	U4	U2	U21
5	U3	U5	U5	U4
6	U7	U6	U9	U12
7	U15	U7	U16	U13
8	U8	U8	U6	U10
9	U4	U9	U7	U16
10	U6	U10	U12	U8
11	U14	U11	U8	U7
12	U2	U12	U13	U5
13	U22	U13	U20	U3
14	U19	U14	U17	U14
15	U5	U15	U21	U2
16	U1	U16	U22	U15
17	U11	U17	U23	U19
18	U21	U18	-	U11
19	U9	U19	-	U17
20	U17	U20	-	U18
21	U18	U21	-	U22
22	U13	U22	-	U9
23	U16	U23	-	U23

제안 방안과 설문조사를 통해 추출된 대학 평판도 순위를 확인해본 결과 많은 차이가 있음을 알 수 있다. 이와 같은 이유를 분석해본 결과 다음과 같은 분석 결과를 도출할 수 있었다. 첫째, 대학 평판도 평가를 위해 수집된 각 대학 뉴스 기사의 양에 대한 편차가 크다. 분석 결과는 데이터의 크기에 많은 영향을 받는다. 데이터의 크기가 많은 경우는 분석 결과에 있어 다양한 결과가 추출될 수 있음에 반하여, 데이터의 크기가 작은 경우에는 다양한 결과가 추출될 수 없다. 이처럼 뉴스 기사의 양이 많은 대학은 많은 이슈를 포함하고 있으며 그 안에서도 다양한 긍정과 부정에 대한 내용이 존재하는 특징이 있다. 이와 반대로 뉴스 기사의 양이 적은 대학은 적은 이슈를 지니며 그 안에서도 긍정과 부정이 한쪽에 치우치는 특징이 있다.

대표적으로 표 6에서 '개수'가 적은 대학의 주제에 대한 선호도를 보면 대체적으로 선호도가 한쪽으로 치우쳐 있는 것을 볼 수 있다. 이처럼 수집된 데이터의 특성상 적은 대학의 평판도에 대해서는

신뢰하기 힘들다는 것을 알 수 있다. 둘째, 설문을 수행한 설문 응답자를 인터뷰한 결과, 거의 모든 응답자는 유명한 상위 3개의 대학(U1, U2, U3)을 제외한 나머지 대학의 평판도에 대해서는 잘 알고 있지 못하였다고 하였다. 그렇기 때문에 상위 3개의 대학 이외 나머지 대학에 대한 평판도의 신뢰도는 떨어진다고 판단할 수 있다. 이는 중앙일보에서 평판도 평가에 사용되는 방식인 설문 조사에 대한 문제점을 보여주는 하나의 예로 사용될 수 있다.

실험 결과에 의하면 제안 방법과 기존 연구의 평판도 분석 결과 차이를 보이는데, 이는 설문 조사에 의해 수행된 기존 연구(기존 연구1)가 지니는 문제점과 기존 연구(기존 연구2)가 지니는 공공 데이터와 텍스트 데이터의 기간적 불일치라는 문제점을 해결하기 위해 이 연구의 제안 방안을 활용하였기 때문에 실험 결과 대부분이 불일치하는 결과를 보인다.

Ⅴ. 결론 및 향후 연구

이 논문에서는 정성적인 데이터를 통해 각 대학을 주제별로 세분화하여 평판도를 평가하는 방안을 제안하였다. 네이버의 '사회-교육'면의 뉴스로부터 23개의 대학 문서들을 수집하고 KNU 한국어 감성사전을 활용하여 대학이라는 도메인에 적용할 수 있는 AR-KNU 감성사전을 구축하였다. 구축한 감성사전을 통해 문단 단위로 감성분석을 수행하여 긍정·부정 문단으로 분류하였다. 대학의 공통적인 주제를 추출하기 위해 감성 문단이 가장 많은 상위 3개의 대학(U1, U2, U3)에 대하여 클러스터링과 클러스터 합병을 통해 추출된 클러스터에 대하여 레이블링을 수행하였다. 추출된 주제를 통하여 나머지 20개 대학에 대한 문단을 다중클래스 신경망을 통해 분류하였다. 주제별로 분류된 각 대학의 문단은 긍정·부정을 비율로 표현하여 주제에 대한 각 대학의 평판도에 대한 양상을 보여주었다.

제안한 알고리즘은 기존 연구들과는 다르게 대학의 종합 평판도만을 보여준 것이 아니라 대학에서 공통적으로 나타나는 주제를 자동으로 추출하여 각 주제에 대한 평판의 양상을 긍정·부정으로 보여주었다. 이와 같은 결과는 각 대학의 평판도를 하나의 종합적인 기준에서만 보여준 기존의 연구들과 다르게, 대학의 평판도를 3가지 주제에 따라 긍정·부정의 비율로 보여줌으로써 사용자에게 보다 유익한 결과를 제공하였다.

이 논문의 제안 방안은 다양한 도메인에 대한 여론 조사 분석에 사용이 가능할 것으로 예상된다. 대표적으로 각 대학을 개선하는데 있어 개선 방향 설정을 위한 기초자료로 활용할 수 있다. 또한 대학에 대하여 관심이 있는 사람들에게 3가지 주제에 따른 평판도를 보여줄 수 있는 자료로 활용될 수 있다.

이 논문의 한계점은 다음과 같다. 첫째, 제안 방안을 위해 사용된 각 대학 데이터의 양이 큰 편차를 보였다. 이는 각 대학 데이터에서 이슈가 한 쪽으로 편향되는 문제점을 발생시킬 수 있으며, 이로 인해 이슈가 한 쪽으로 치우치는 부정확한 결과를 보일 수 있다. 둘째, 특정 대학에 대한 하나의 기사 안에는 특정 대학만 언급하는 것이 아니라 다양한 대학이 함께 언급된다. 이는 특정 대학을 분석하는데 있어서 다른 대학이 영향을 끼칠 수 있다는 문제점을 발생시킬 수 있다. 또한 이러한 문제점은 데이터 마이닝의 군집화 알고리즘을 적용하는데 있어 다른 대학의 정보로 인해 군집하고자 하는 대학의 군집화 결과가 잘 수행되지 않을 수 있다는 한계점이 존재한다.

이 논문의 향후 연구로는 제안 방안의 한계점을 보완하기 위해 실험 데이터의 크기를 늘려 대학 평판도 평가를 분석할 것이다. 특정 대학에 대한 하나의 기사에서 특정 대학에 대한 문장, 문단만을 분류하는 분류 기법 또한 제안하여 정확한 분석 결과를 추출할 것이다. 마지막으로 주제를 보다 세부적으로 나누고 정확하게 추출할 수 있는 기법을 제안하여 대학 평판도를 평가할 예정이다.

References

[1] CIO, "Expansion of BI and analytical role", http://www.ciokorea.com/t/544/9118/15551. [accessed: Sep. 22, 2018]

[2] Wikipedia, "Text minig", https://en.wikipedia.org

/wiki/Text_ mining. [accessed: Sep. 22. 2018]
[3] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, Vol. 5, No. 4, pp. 1093-1113, Apr. 2014.
[4] Joongang, "University reputation evaluation", http://univ.joongang.co.kr/. [accessed: Sep. 22. 2018]
[5] 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.
[6] M. H. Yang, I. S. Jung, Y. T. Kim, and W. S. Cho, "Analysis and recognition of domestic universities using SNS data", Journal of Big Data Applications and Services, Vol. 1, No. 1, pp. 1-13, Dec. 2014.
[7] B. W. On, S. M. Park, and C. W. Na, "KNU sentiment lexicon", https://github.com/park1200656/KnuSentiLex. [accessed: Aug. 01, 2018]
[8] W. J. Lee, "An empirical approach to evaluate college image perception", Journal of Multimedia Services Convergent with Art, Jumanities, and Sociology, Vol. 1, No. 5, pp. 57-66, Feb. 2015.
[9] R. Xu and D. Wunsch, "Survey of clustering alogrithms", IEEE Trans. on Neural Networks, Vol. 16, No. 3, pp. 645-678, May 2005.
[10] Wikipedia, "Neural network", https://en.wikipedia.org/wiki/Neural_network. [accessed: Sep. 22, 2018]
[11] E. J. Park and S. J. Cho, "KoNLPy: Korean natural language processing in python", Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology, pp. 133-136, Oct. 2014.
[12] Waikato university, "weka 3.8", https://www.cs.waikato.ac.nz/ml/weka/. [accessed: Aug. 01. 2018]
[13] Microsoft, "azure ml studio", https://studio.azureml.net/. [accessed: Aug. 01, 2018]
[14] Facebook, "fasttext", https://fasttext.cc/. [accessed: Aug. 01, 2018]

저자소개

박 상 민 (Sang-Min Park)

2018년 : 군산대학교 소프트웨어융합공학과(학사)
2018년 ~ 현재 : 군산대학교 소프트웨어융합공학과(석사)
관심분야 : 텍스트 마이닝, 데이터 마이닝, 인공지능, 감성사전, 감성분석

엄 창 민 (Chang-Min Eom)

2013년 ~ 현재 : 군산대학교 소프트웨어융합공학과(학부생)
관심분야 : 프로그래밍, IoT, 소프트웨어공학

온 병 원 (Byung-Won On)

2007년 : The pennsylvania state U. 컴퓨터공학과(박사)
2008년 ~ 2009년 : 캐나다 U. of British Columbia 박사 후 연구원
2010년 ~ 2011년 : U. of Illinois at Urbana-Champaign ADSC 연구소 선임연구원
2011년 ~ 2014년 : 차세대융합기술연구원 공공데이터 연구센터 센터장
2014년 ~ 현재 : 군산대학교 소프트웨어융합공학과 교수
관심분야 : 데이터 마이닝, 정보검색, 빅데이터, 인공지능

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