대학 BI 분석을 위한 주제분류기의 구현

장서윤O, 장현영*, 차채원*
O금오공과대학교 컴퓨터공학과,
*금오공과대학교 컴퓨터공학과
e-mail: {seoy316, hung1224, 20171145}@kumoh.ac.krO*

Implementation of Topic Classifier for University News-based BI Analysis

Seo-Yoon JangO, Hyeon-Yeong Jang*, Chae-Won Cha*
ODepartment of Computer Engineering, Kumoh National Institute of Technology,
*Department of Computer Engineering, Kumoh National Institute of Technology

● 요 약 ●

본 논문에서는 대학별 홍보 전략, 발전에 기여하기 위한 서비스를 제안한다. 이 서비스는 데이터 수집에는 크롤링을 사용하고 사이킷 런을 사용하여 정확도를 최대화하고, 각 분류된 카테고리의 오류을 최소화한다.

이 서비스는 각 카테고리별로 특성이 높은 키워드를 사용하여 카테고리 별 학습 데이터셋을 생성한 후 이러한 학습 데이터셋을 바탕으로 각 기사들을 최적의 카테고리로 분류해주는 분류기를 구현한다. 이러한 분류기를 사용하여 분류된 기사들을 분석하여 막대 그래프 등의 시각화된 자료들로 볼 수 있도록 하여 기존의 대학 홍보 자료에 비해 누구든 쉽고 간단하게 접근이 가능하도록 한다.

키워드: 카테고리 분류(Classifiaction), 딥러닝(Deep learning), 시각화(Visualization)

I. Introduction

예부터 광고 시장은 우리의 생활에 크게 영향을 끼쳤으며 인터넷의 발달 이후 급속도로 성장하고 있다. 이에 따라 기업에서는 Business Intelligence (BI) 라는 홍보 전략을 사용하여 정보를 수집, 가공, 분석하여 사업의 방향성을 잡았다. 반면 대학에서는 교육의 질을 개선하여 입시 결과를 높이고, 재학생들에게 유익한 사업을 유치하는 등의 노력을 하고 있다. 하지만 기대효과에 미치지 못하는 곳이 대다수이다. 그렇기에 확실한 효과를 보장받기 위해서는 구체화한 방법이 필요하다. 구체화한 방법으로 위와 같이 기업에서 활용하고 있는 BI로 극대화된 홍보 효과를 모색하고자 한다.

서비스의 구현 초기 단계에서는 크롤링을 통해 대구 경북권의 주요한 세 대학의 최근 5년치 온라인 기사를 수집하였고, 총 50,000건에 달한다. 다음 단계에서는 수집한 기사들을 6가지 카테고리로 분류하는 분류기를 구현하였다. 이를 통해 분류된 각 카테고리의 데이터를 분석하여 결과를 시각화하는 단계까지 구현하였다.

II. Proposed Scheme

1. 작업 환경

파이썬 3.7과 3.8을 사용하였고 GPU 사용을 위해 구글의 코랩을 사용하였다.

2. 데이터 구축 및 주요 토픽 선정

크롤링에 사용한 라이브러리는 BeautifulSoup과 Selenium이며, 이를 사용하여 최종 목표의 대상이 될 기사들을 수집한다. 수집된 데이터의 형식은 날짜와 기사 본문으로 이루어져있다. 그후 re 라이브러리로 파이프라인을 구성하여 불필요하거나 정보성이 낮은 데이터를 제거하였다. 이러한 전처리 과정을 거치며 학습 모델의 데이터셋 균형을 맞추어주었다. 다음으로는 카테고리를 대표하는 '이벤트어'를 선정하였고, 이를 이용하여 약 20,000건의 기사를 연구, 교육/사업, 문화/복지, 학생, 교수진, 입시 총 6가지의 카테고리로 필터링하였다. 각각의 카테고리에는 1,000건의 데이터가 들어있다. 마지막으로 Mecab 라이브러리를 사용하여 필터링이 완료된 데이터셋을 형태소 단위로 토큰화하고, 명사만 추출한다. 명사를 추출하기 전 분석이 안 되는 단어들은 사용자 사전을 만든다. 이렇게 구축한 사전에는

학과명, 학교명, 교육사업명, 이벤트어 등이 있다.

3. 주제 분류기 구현

카테고리별로 필터링 된 학습 데이터셋은 TfidfVectorizer를 이용하여 벡터화하였다. 전체적으로 언급이 높은 단어를 제외함으로써 벡터 공간의 차원을 줄일 수 있고, 해당 샘플을 잘 표현하는 단어에 가중치를 두어 학습 정확도를 높이고자 하였다. 이후 scikit-learn의 SVM 알고리즘을 사용하였다. SVM은 벡터 공간에서 최대 분류 마진을 찾아내는 알고리즘으로 다중 분류 성능이 뛰어나다는 장점이 있다. 그러나 각각의 클래스를 정확히 분류하는 초평면을 찾아내는 것이 제한되어있다는 문제가 있다. 이를 해결하기 위해 더 높은 차원으로 사상시키고, 최적의 초평면을 찾을 수 있도록 SVM의 확장된 형태인 SVC 선형 분류법으로 학습하였다. 학습하는데 사용된 데이터셋의 양은 카테고리마다 1,000건씩 해서 총 6,000건이 사용되었다. 그중 70%는 train data로 사용하였고, 나머지 30%는 test data로 사용하였다. 정확도는 92%가 나온다. 다음으로 joblib을 사용하여 92% 성능을 끌어낸 모델을 피클로 저장하였고, 저장된 모듈을 로드해서 사용함으로써 여섯 가지의 카테고리를 분류할 수 있는 분류기가 생성되었다. 분류기에 CSV 파일로 기사를 넣으면 각 기사에 해당하는 카테고리로 분류되고, 그 안에서 새로운 CSV 파일을 생성하여 저장된다.

III. Analysis Results

1. 분석 방법

사용한 라이브러리는 matplotlib와 TfidfVectorizer이다. 2020년 9월, '학생' 카테고리의 데이터에서 연관 단어 상위 10개를 추출하여 막대그래프를 형성하도록 하였다. 연관성의 기준은 TF-IDF 가중치로 판단한다. 가중치가 높은 단어는 해당 단어의 빈도수를 함께 측정하여 시각화 하였다.

2. 분석 결과

2.1 경북대학교 결과

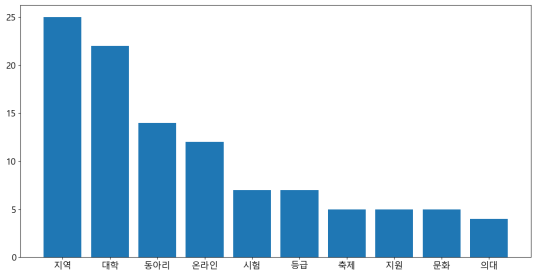

Fig. 1. 2020년 9월, '학생 카테고리 기준'

9월이라는 신학기 시기에 축제, 동아리 등의 키워드가 상위권에 있는 것을 볼 수 있다. 또한, 온라인이라는 키워드로 미루어 보아 비대면 강의와 관련된 기사가 많은 것을 알 수 있다.

2.2 영남대학교 결과

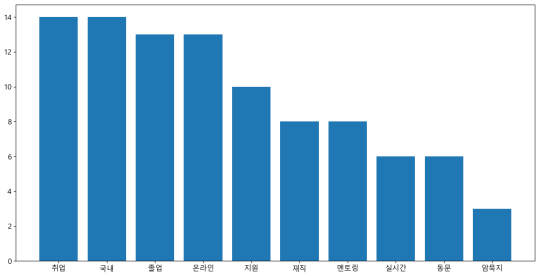

Fig. 2. 2020년 9월, '학생 카테고리 기준'

코로나로 인한 비대면 강의를 했음을 알 수 있고 온라인 강의뿐만 아니라 온라인 취업 프로그램도 활성화되어 있는 것 또한 알 수 있다. 그 밖에도 학생 지원 멘토링 등 학생 교육 활동이 활발하게 이루어지고 있는 것을 알 수 있다.

2.3 금오공과대학교 결과

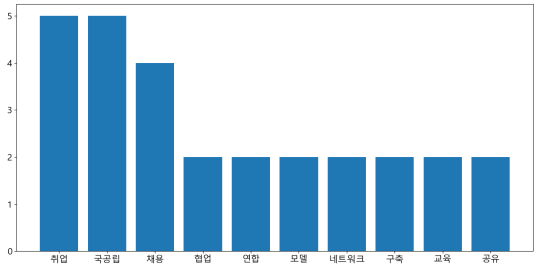

Fig. 3. 2020년 9월, '학생 카테고리 기준'

취업, 채용 등 졸업을 앞둔 학생들과 관련된 키워드가 두드러진다.

IV. Conclusions

본 논문에서는 대학별 홍보 전략, 발전에 기여하기 위한 서비스를 제안하였다. 기사 수집 후 딥러닝을 사용하여 대학교 뉴스 기사의 주제 분류기를 구현하였다. 구현된 분류기를 사용하여 수집한 기사를 6가지 카테고리로 자동분류를 하고 각 카테고리의 관련도가 높은 상위 10개 단어를 추출하여, 그래프로 시각화 시킬 수 있는 환경을 구현하였다. 이 논문에서 제공하는 서비스를 이용할 시 입시생, 재학생, 학교 등 사용자들에게 도움이 되는 정보의 폭을 넓힐 수 있고, 이를 시행해봄으로써 각자의 입장에서 좋은 기회를 제공받을 수 있다. 구현에서의 부족한 점을 살펴보면 현재 이 서비스는 명사만을 추출하여 그래프로 보여주기 때문에 모든 학교에서 자주 나타나는 단어가 표시될 수 있다. 따라서 향후에 개선한다면 하나의 이벤트어를 그대로 추출하는 방법을 모색하여 적용하고자 한다.

REFERENCES

[1] ”Naver News Crawling & Pre-Procession Exercise” NAVER blog, [Accessed: Dec. 24, 2020]. https://m.blog.naver.com/jeonghj66/222057942747

[2] "An introduction to machine learning with scikit-learn", scikit Learn, [Accessed: Dec. 24, 2020], https://scikit-learn.org/stable/tutorial/ basic/tutorial.html

[3] “Chapter 3 Machine Learning Classification Model Using Scikit-Learn”, Tistory, last modified 12.29.2020, [Accessed: Dec. 24, 2020], https://learningmachine.tistory.com/6

[4] “tfidf Extract new sentence unit keywords.py”, github, last modified 08.13.2019, [Accessed: Dec. 24, 2020], https://lab.hanium.or.kr/vallot8/han_project/blob/4dbc7667af08ced123dad312aa55bbbadbf888d3/Zn_text_summary/tfidf%EC%97%AC%EB%9F%AC%EB%AC%B8%EC%9E%A5%EB%8B%A8%EC%9C%84%ED%82%A4%EC%9B%8C%EB%93%9C%EC%B6%94%EC%B6%9C.py

[5] 김미지, 이재원, 장달원, 이종설. (2018). 키워드 가중치를 이용한 뉴스 기사에서의 이슈 키워드 자동 추출 시스템. 한국방송미디어공학회 학술발표대회 논문집, 146-148.

[6] 박미선, 신효주, 김우제. (2018). 텍스트마이닝을 활용한 블록체인 관련 기사 주제 분류. 한국정보과학회 학술발표논문집, 1003-1005.

[7] 전성해. (2007). 차분진화 기반의 Support Vector Clustering. 한국지능시스템학회 논문지, 17(5), 679-683.