

텍스트 마이닝을 통한 리터러시 융합 교육: 영화 <엑트오브킬링>을 중심으로

강보영* · 홍문표**(성균관대학교)

국문초록

본 연구는 자연어 처리의 텍스트 마이닝 기술과 데이터 시각화 기법을 사용하여 능동적 실험(learning by doing)을 해봄으로써 영화라는 미디어가 생산하는 메시지의 의미를 심층적이며 효과적으로 학습하는 방법을 제안한다. 미디어 환경은 21세기 들어 급격하게 변화해왔으며 그 변화의 기저에는 플랫폼으로의 전환과 AI, 빅데이터, 클라우드 등의 ICT기술의 획기적인 발전이 있었다. 따라서 미디어 리터러시란 시대와 기술의 변화에 맞물린 개념으로 다양한 미디어가 생산하는 메시지의 의미를 파악하고 비판적 참여를 할 수 있는 능력을 뜻한다. 본 연구에서는 자연어 처리기술 중의 하나인 텍스트 마이닝 기법을 이용해 다큐멘터리라는 시사적인 미디어에 대해 소셜미디어를 통해 표출된 대중의 의식과 함의를 분석해보고자하였다. 소셜미디어에 게재된 영화 리뷰는 비정형 데이터이나 텍스트 마이닝 기법을 활용해 정형 데이터로 분석되었을 때, 어떤 전문가들의 비평이나 평가와는 다른 소셜 데이터로서의 새롭고도 중요한 가치를 지닌다. 본 연구에 소개된 학습을 위해 조금 이상의 파이썬 코딩 실력과 머신러닝 툴을 사용한 데이터 분석을 할 수 있는 학습자 뿐 아니라 코딩을 할 수 없는 사람이라도 웹 사이트 및 구글 클라우드 오픈 API를 활용하여 워드 클라우드와 감성 분석을 체험해 볼 수 있다. 본 논문은 네이버 영화에서 <엑트 오브 킬링>영화관련 웹 크롤링 및 동적 크롤링, 2) 텍스트 데이터 형태소 분석, 3) 빈도 분석, 4) 맥락분석, 5) 감성분석의 절차로 연구 프로세스를 진행하였다. 연구 결과 텍스트 마이닝을 통해 광주 민주화 운동 등의 한국의 민주화 역사와 최근 미얀마 사태에 대한 생각을 읽을 수 있었다. 본 연구를 통해 문·이과를 넘어 컴퓨터 프로그램 활용 능력 수준별 미디어 리터러시 STEAM 융합 교육을 제안한다.

주요어 : 텍스트 마이닝, 미디어 리터러시, 엑트 오브 킬링, 자연어 처리, 데이터 시각화, 웹 크롤링, 빈도 분석, 감성 분석

I. 서론

미디어 환경은 21세기 들어 급격하게 변화해왔으며 그 변화의 기저에는 플랫폼으로의 전환과 AI, 빅데이터, 클라우드 등의 ICT기술의 획기적인 발전이 있었다. 따라서 미디어 리터러시란 시대와 기술의 변화에 맞물린 개념으로 다양한 미디어가 생산하는 메시지의 의미를 파악하고 비판적 참여를 할 수 있는 능력을 뜻한다. 동핀란드대 발토넨(Valtonen, 2019) 교수와 세계적인 기술지, 와이어드

* 성균관대학교 미래인문학소셜앙트레프렌쉽 융합 전공 박사과정 bykang2015@gmail.com

** 성균관대학교 미래인문학소셜앙트레프렌쉽 융합 전공 교수 skkhmp@skku.edu

(Wired)의 공동창업자인 켈리(Kelly, 2017)는 미디어 소비자가 사용하거나 생산한 대량의 비정형화된 영상, 텍스트, 음성, 위치 정보 등의 데이터가 가공되거나 조작되어 다시 소비자들을 대상으로 광고하거나 유인하는데 사용된다는 사실에 대해 강조한다. 뉴스를 포함한 모든 미디어는 현실 그대로를 전달하기 보다는, 의도를 가지고 만들어진 ‘재현(representation)’의 산물(Buckingham, 2003)이며 미디어 리터러시 교육의 핵심은 미디어가 재현하는 내용에 대한 비판적 분석과 이해, 그리고 이를 바탕으로 한 시민으로서의 사회 참여 역량을 기르는 것이다(정현선, 2019). 이미 핀란드에서는 K-12 초·중등 교육에서 컴퓨터와 관련된 ICT 능력을 리터러시 교육과 밀접하게 연관시켜 교육하여 왔는데 특히 국가주도의 소프트웨어 교육을 필수교육과정으로 명시한 KOODI 2016 교육과정을 개발하여 전면적으로 시행하고 있다. 핀란드의 코딩교육은 코딩의 원리에 대한 이해, 컴퓨팅 사고력 및 문제 해결력을 바탕으로 새로운 것을 창의적으로 표현할 수 있는 인재양성을 추구한다고 볼 수 있다. 발토넨(2019) 교수는 더 나아가서 미디어와 정보의 홍수시대에 학생들이 인공지능 기술 및 텍스트 마이닝 등의 기술을 활용하여 탐색적이며 능동적으로 미디어 대처하는 새로운 미디어 리터러시를 갖추는 준비를 해야 한다고 주장한다. 우리나라에서 미디어 리터러시 교육의 역사는 시민사회 영역에서 먼저 시작되었고 뒤늦게 학교에서도 도입되고 있으나 주로 디지털 환경의 미디어 리터러시의 측정에 대해 인지적 측면과 태도적 측면의 연구가 주를 이루고 있다(정현선, 2019). 그러나 실제적인 리터러시 역량을 발전시키는 학습 방법에 대한 연구는 드물며 미디어의 이해를 바탕으로 한 텍스트의 분석, 평가, 창조 능력을 평가하기 위한 방법에 대한 보다 구체적인 연구가 필요한 상황이다.

본 연구에서는 콜브(Kolb, 1984)의 경험학습(experimental learning)이론에 입각해서 텍스트 마이닝 실습 과정을 통해 실제적인 리터러시 계발을 유도하도록 하였으며 이러한 방식의 미디어 리터러시 학습은 최근 그 중요성이 강조되는 이해 중심의 교육과정 및 평가의 패러다임과 잘 부합한다. 본 연구의 분석 대상은 다큐멘터리 영화이다. 예술적이면서도 시사적인 미디어인 다큐멘터리의 의미와 메시지의 효과를 과학적인 방법인 텍스트 마이닝을 통해 분석해보는 것은 융합형 인재를 양성하기 위한 STEAM 교육의 한 방법이 될 수 있다. 예술이 다른 학문을 도와주는 도구적이며 보조적 역할을 하는 것이 아니라, 예술이 주축이 되면서도 미래 사회에 필요한 다양한 학문, 예를 들면 데이터 과학, 언어, 역사 등을 학습하면서 우리가 살고 있는 사회의 시사적인 이슈에 대한 고찰과 해결책을 만들어 낼 수 있는 중추적 역할을 할 수 있다. 본 연구가 과학기술이 예술과 상보적 관계로 나아가기 위한 접근 방법이자 융합적 학습방법으로 잘 활용되길 바란다.

II. 연구 배경

1. 리터러시와 소프트웨어 융합 교육

영국 옥스퍼드대 부설 로이터 저널리즘 연구소가 발간한 디지털 뉴스 리포트(2020)는 뉴스에 대한 대중의 이해와 신뢰도를 측정하는 통계 보고서를 발표했는데 한국은 조사 대상 40개국에서 언론에 대한 신뢰가 가장 낮은 국가로 집계되었다. 미국의 비영리교육단체인 커먼센스 미디어(Common Sense Media)에서는 미디어 리터러시를 다양한 미디어가 생산하는 메시지의 의미를 파악하는 능력으로 보았으며, 캐나다의 비슷한 성격의 비영리교육기관인 미디어 스마트(Media Smarts)에서는 ‘대

중매체에 대한 비판적 참여(critical engagement with mass media)'로 정의한다.

그런데 이러한 미디어 리터러시의 공통 분모에는 변화하고 있는 디지털 미디어에 대한 디지털 리터러시가 포함되어 있는데 디지털 리터러시란 디지털 사회를 살아가기 위한 개인적, 기술적, 지적 기능(personal, technological, and intellectual skills for living in digital society)을 말한다. 따라서 서론에서 언급한 ICT 능력 및 컴퓨팅 사고 능력과 문제 해결력을 갖추는 것이 미디어 리터러시를 배양하는데도 필요하다고 사료된다.

우리나라에서는 사회에서 필요로 하는 소프트웨어 인재 양성을 위해 2018년부터 초·중등학교에서 소프트웨어 교육을 필수 편성·운영하는 등의 변화가 있어왔다. 그러나 박현제(2020)¹⁾는 해외 주요국은 초등학교에서부터 매년 32~68시간 정도 수업을 실시하는 데 반해, 우리나라는 초등학교 총 17시간, 중학교 총 34시간 등 SW 관련 수업 시수 및 프로그래밍 경험이 부족하다고 설명한다. 특히 2015년 개정 교육 과정은 소프트웨어 교육을 통해 문·이과 구분을 없애고 과목간 융합교육 및 창의적 인재 양성을 주요 골자로 하는데, 그 배경에는 모든 산업에서 요구되는 데이터 과학 및 머신러닝은 학제간 융합 연구를 필요로 하기 때문이다. 이서빈(2018)의 연구에서 컴퓨터 프로그래밍 교육을 통해 단지 컴퓨터 활용 능력이 향상될 뿐 아니라 프로그래밍에 대한 가치와 실용성 및 프로그래밍 자기효능감이 향상되었다. 그러나 소프트웨어교육 선도학교 중에서도 소프트웨어교육 기반의 융합교육 과정을 운영하는 학교는 전체 선도학교의 절반 정도(50.8%)에 불과한 것으로 나타났다. 선도학교로 손꼽히는 포항제철중학교의 소프트웨어 융합교육 프로그램(2019)²⁾을 살펴보면 국어 교과 과목에서 정보 교육과 융합하여 엔트리나 스크래치 프로그램을 이용하여 표현하거나 간단한 프로그래밍이 있다. 본 연구에서 사용되는 기법인 텍스트 마이닝도 국어 교과와 융합하거나 실생활 문제 해결형 등 다양한 프로그램으로 개발 운영할 수 있다.

데이터의 증가 속도와 맞물려 점점 그 활용의 중요성도 강조되고 있는 오늘날의 데이터의 80% 이상이 비정형 데이터로 추산되고 있는데(Chakraborty, Pagolu, 2014), 텍스트 마이닝은 비정형 데이터의 분석 방법론으로서 각광을 받고 있다. 정형화된 데이터베이스에서 패턴을 추출하는 데이터 마이닝 작업과는 달리 텍스트 마이닝은 일반적으로 구조화되지 않은 형식으로 저장되는 개체간의 관계나 패턴을 추출하기 위해서 사용된다. 웨스트(Jason West, 2017)³⁾는 교사가 교과 과정을 개발할 때 텍스트 마이닝 기법을 이용하면 연구 주제의 동향을 파악할 수 있는 장점이 있다고 설명한다. 실제로 텍스트 마이닝 기법은 다양한 문헌의 주요 트렌드 동향 및 언론 미디어 분석의 주요 방법론으로 활용된다(감미아, 2012; 이동국, 2020).

1) “컴퓨팅 사고력 배양 위한 초기 정규교육 시급”, 정보통신신문, 2021.03.23.

2) 김하중(2019), “초·중등 소프트웨어교육 운영실태와 개선과제”, 입법·정책보고서, 제34호, p41

3) Jason West(2017), Validating curriculum development using text mining, THE CURRICULUM JOURNAL, VOL. 28, NO. 3, p390

표 1. 포항제철중학교의 소프트웨어 융합교육 프로그램⁴⁾

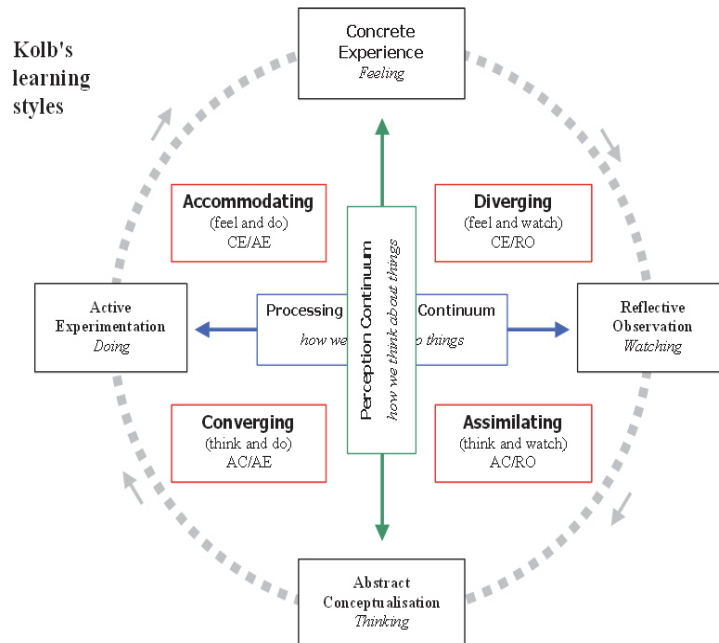
유형	관련교과	소프트웨어융합 교수 학습 활동	교구
교과 지식 융합형	정보(국어)	문학을 감상하고 느끼고 이해한 것을 소프트웨어를 이용하여 표현	엔트리
	정보(국어)	시와 스크래치 프로그램의 만남을 통한 창작 모방 시 재생 프로그래밍	스크래치
	정보사회	증강현실 어플을 이용하여 세계의 다양한 지리정보를 설명	스마트패드
	정보(과학) 미술음악	빛의 합성에 관한 과학적 지식, 코딩 지식을 활용한 뮤직 비디오 제작	스크래치, 로버티스트
	정보(과학)	액상에 따른 지시약의 색을 이해하고, 이를 활용한 신호등 만들기	스크래치, 3D펜
	정보사회 (기술)	등하굣길의 주요 안전 요소를 점검하고, 위험에 대비하기 위한 지도를 구상	엔트리
	정보(가정)	녹색 식단을 이해하여 가족의 영양과 건강을 고려한 식단 프로그래밍	엔트리
	(정보) 과학	지진계의 원리를 습득, 센서를 활용한 피지컬 컴퓨팅을 통해 알고리즘 을 설계하고 프로그램으로 구현	스크래치 코드이노 센서보드
	정보 (음악)	장구의 원리를 파악하고 프로그래밍을 통해 전자장구 실현	엔트리 코드이노 센서보드
	정보 (과학)	생물을 계로 구분짓는 특징을 알고, 각 특징을 바탕으로 제시된 생물 을 엔트리 활용하여 생물 분류 퀴즈 프로그램 제작	엔트리
실생활문제 해결형	정보 (영어)	코딩로봇을 통해 여행지에서 필요한 길찾기 표현에 대해 학습	오조봇

* ()안 과목은 중심교과이며 융합수업을 통해 학습목표 및 성취기준을 달성해야 하는 교과이고, 나머지는 중심교과의 학습을 돕기 위해 지식이나 기능을 제공하는 융합교과임

2. 경험 학습(Experimental leaning) 이론

콜브(Kolb)는 학습을 “경험의 변형을 통해 지식이 만들어지는 과정”(김진한, 2020) 이라고 정의하며 구체적 학습(learning by feeling), 관찰적 관찰(learning by watching), 추상적 개념화(learning by thinking), 능동적 실험(learning by doing)으로 순환하여 지식을 통합하는 과정으로서 경험, 학습, 지식의 관계를 설명하였다. 콜브의 경험 학습 모형의 특징은 학습을 결과가 아닌 과정으로 보았으며 지식을 경험에 의해 계속적으로 창조되는 변형의 과정으로 본다는 것이다. 경험 학습은 ‘경험에 의한 학습(learning by doing)’과 ‘경험으로부터의 학습(learning from experience)’을 포괄적으로 함의하는 개념으로 최근 많은 교육과정에서 도입되는 ‘러닝바이두잉(learning by doing)’은 학습할 내용을 실제적으로 능동적 실험을 통해 경험하면서 습득하게 되는 교육 방식에 초점을 두는 콜브의 이론을 모태로 한다고 할 수 있다.

4) 김하중(2019), “초·중등 소프트웨어교육 운영실태와 개선과제”, 입법·정책보고서, 제34호, p41 [표 13]재사용

그림 1. 콜브(Kolb)의 경험 학습 모형⁵⁾

앞서 언급한 소프트웨어 융합 교육은 물론이고 융합형 인재를 양성하기 위한 STEAM 교육에서도 능동적 실험 과정이 중요시 되고 있다(홍현미, 2017). STEAM 교육이란 과학(Science) · 기술(Technology) · 공학(Engineering) · 예술(Arts) · 수학(Mathematics)의 약칭으로, 융합인재교육에 국가 경쟁력을 높일 수 있는 창조적이고 융합적인 인재 양성을 위한 예술(Arts)을 통합한 교육을 의미한다. 따라서 21세기 들어 대두되고 있는 융합인재교육을 바로 STEAM 융합교육이라고 칭하기도 하며 급변하는 사회의 요구를 반영하여 창의와 인성을 지닌 미래 융합형 인재를 양성하는 것을 목표로 한다(교육과학기술부, 2010; 한국과학기술단체총연합회, 2011). 본 연구에서 텍스트 마이닝이라는 기술을 사용하여 다큐멘터리 영화를 분석하고 시사점을 도출하는 과정을 능동적 실험을 통해 단지 영화 예술이 보조적인 도구나 대상이 되는 것이 아니라 함의를 지닌 미디어로서의 기능을 발견하는 경험 학습이 이루어질 수 있다.

3. 데이터 시각화 교육

데이터 시각화 기술의 발달과 멀티모달 (multi-modal) 교재의 교육 효과성에 기인하여 데이터 시각화 기법은 빈번히 사용된다. 특히 초·중등 교육에서 데이터 시각화는 주로 텍스트를 부연 설명하기 위해서 사용하는데 가장 보편적인 기능인 정교화(elaborate)와 그 다음으로 꾸밈(embellish), 강화(reinforce) 및 요약(summarize) 등의 기능 순으로 사용된다.

5) "An Introduction to Kolb's Learning Styles", Businessballs.com, <https://www.businessballs.com/self-awareness/kolbs-learning-styles/> 재사용

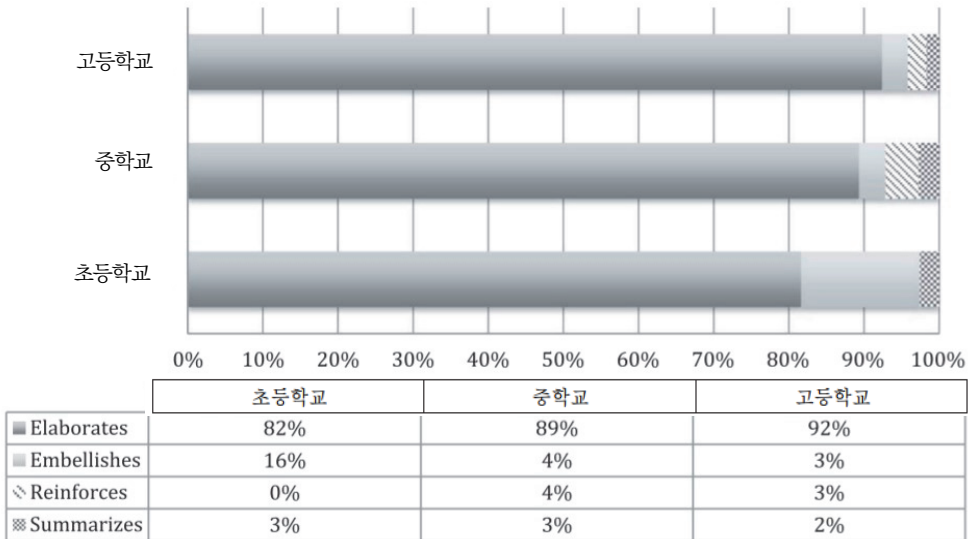


그림 2. K12 교재에서 데이터 시각화의 목적⁶⁾

교재에서 이미 만들어진 표와 차트 등의 의미를 정확하게 이해하고 활용하는 능력을 학습하는 데서 한 단계 더 나아가서 학생들이 직접 본인이 읽고 있는 교과서나 뉴스, 영화 등의 미디어 텍스트를 가지고 텍스트 마이닝을 통한 빈도 분석과 워드클라우드 등의 분석 기법을 사용하여 데이터 시각화 작업을 해보는 것은 정교화 등을 통한 인지적인 효과에 능동적 경험 학습의 효과성까지 배가될 수 있다. 본 연구에서 활용해 볼 데이터 시각화 학습은 크롤링된 텍스트 데이터를 빈도 분석을 실행함에 있어서 사이트에서 제공하는 워드 클라우드 생성 서비스와 머신러닝 라이브러리와 파이썬 코딩 프로그램을 통한 워드 클라우드 시각화 데이터를 직접 생성해내고 비교 분석해 보는 것이다. 또한 감성 분석에 있어서 구글 클라우드 API 호출과 출력을 경험하고 시각화하며 코딩을 통한 머신러닝 감성 분류 학습의 결과를 비교하는 것도 좋은 시각화 교육이 될 것이다.

III. 연구 방법

본 논문은 다큐멘터리 영화 <엑트 오브 킬링>에 대한 네이버 영화 리뷰와 블로그에 게재된 텍스트를 분석하였다. 조슈아 오펜하이머(Joshua Oppenheimer) 감독과 한국계 크리스틴 신 감독이 공동 연출한 엑트 오브 킬링(The Act of Killing)은 1965년 인도네시아에서 발생한 수하르토 대학살을 소재로 한 영화이다. 이 영화가 주목을 받게 된 가장 중요한 이유 중 하나는 영화 제목처럼 역사적인 ‘살인 행위’의 재현(再現) 뿐만 아니라 실제 가해자들이 저지른 범죄에 대한 ‘살인 연기’의 직접적인 재연(再演) 양식이 결합되어 있다는 것이다.

6) Tamara L. Shreiner (2018), “Data Literacy for Social Studies: Examining the Role of Data Visualizations in K-12 Textbooks” 에서 Figure 4. 재사용(일부 번역)

최초의 다큐멘터리 이론가인 존 그리어슨(John Grierson)은 다큐멘터리를 ‘실재의 창조적 방법(creative treatment of actuality)’ 이라고 정의하며, 미국의 영화 비평가이자 다큐멘터리 입문(Introduction to Documentary)의 저자인 빌 니콜스의 주장에 의하면 ‘우리가 응시하는 세계의 재현(representation)’이다. 따라서 실재의 재현이자 작가의 주관적인 해석이 병존하는 다큐멘터리의 의미와 효과는 시간이 흐른 뒤에 사람들의 내재된 의식의 표출 효과를 과학적으로 분석해 봄으로써 명확해질 수 있다.

본 연구에서는 5.18 광주 민주화 운동 41주년에 기념해서 민주화와 제노사이드라는 트라우마를 소재로 한 다큐멘터리 영화에 대해 소셜미디어를 통해 표출된 대중의 의식과 함의를 분석해 보고자 하였다. 영화진흥위원회에 따르면 2016년 현재 한해 극장에 개봉된 한국의 독립영화와 예술영화를 포함한 다양성영화는 대략 100여편에, 다큐멘터리는 30편가량 된다. 액트 오브 킬링은 국내에서는 2013년 17회 부천국제판타스틱영화제에서 최초 상영되었으며 2014년 11월 20일 개봉되었다. 전국 상영관 32개에서 총 8,683명이 관람했으나, 네이버 영화 리뷰에 직접 참여한 사람은 약 390명으로 집계된다. 본 연구에서는 영화의 관객 반응을 파악하기 위해, 네이버 리뷰 및 관련 블로그 텍스트를 2021년 5월 6일 기준 수집하였으며 1)웹 크롤링, 2)형태소 분석, 3)빈도 분석, 4)맥락분석, 5)감성분석의 절차로 연구 프로세스를 진행하였다. 프로그래밍 개발 환경은 구글 콜랩(Colab)과 아나콘다 주피터 노트북(Jupyter Notebook)에서 실행하였다. 콜랩은 설치 없이 브라우저(<https://colab.research.google.com>)를 통해 코드를 작성하고 실행하며 구글 드라이브로 공유하여 협업을 하기도 편리하나 무료로 사용하는 클라우드 상황에서 실행이 불안정한 경우가 발생한다. 반면 주피터 노트북은 아나콘다(Anaconda) 패키지를 통해 함께 설치해야 하며 라이브러리인 '넘파이(numpy)', '판다스(pandas)' 등을 한꺼번에 다운로드 받을 수 있는 장점이 존재한다. 두 가지 개발 인터페이스 모두 프로그래밍 언어로서 파이썬(Python)을 사용한다. 파이썬은 간결하고 직관적인 문법 체계이기 때문에 초·중등 교육 과정에서도 많이 사용된다.

1. 웹 크롤링

소셜미디어에 게재된 영화 리뷰는 수집 및 분석하기 용이하며 텍스트 마이닝을 통해 평가되었을 때 어떤 전문가들의 비평이나 평가와는 다른 소셜 데이터로서의 새롭고도 중요한 가치를 지닌다. 본 연구에서는 영화 <액트 오브 킬링>에 대한 관객들의 반응을 파악하기 위해, 네이버 영화 리뷰를 웹 크롤링을 통해 수집하였다. 웹 크롤링이란 자동화된 프로그래밍으로 웹 상에 존재하는 콘텐츠를 수집하는 기법을 말하며 정적 크롤링과 동적 크롤링이 있다. 본 연구에서 영화 리뷰 텍스트 크롤링을 위해서 정적 크롤링을 사용하였으며 영화 관련 블로그 텍스트의 크롤링을 위해서는 동적 크롤링의 방법을 사용하였다. 정적 크롤링이란 주소를 통해 단발적으로 접근하는 것을 말하는데, 속도는 빠르지만 수집하는 양에는 한계가 있다. 본 연구에서는 파이썬 3.7 버전의 BeautifulSoup 라이브러리와 한글 형태소 분석기 KoNLPy를 이용하였다. requests 모듈로 평점(score), 리뷰글(review), 찬성(sympathy), 반대(notSympathy) 속성을 수집하였다. 반면 동적 크롤링은 브라우저를 통해 연속적으로 접근하는 방식으로 방대한 양을 수집할 수 있으나 속도가 느리다. 본 연구에서는 동적 크롤링을 위해 크롬 드라이버와 selenium 라이브러리를 이용하였다. 정적 웹 크롤링 결과 총 390명의 13,586 문자(공백제외)의 말뭉치(corpus)가 수집되었으며 찬성, 반대 포함 누적 합산 2,922명이 의견이 반영된 결과가 표 2에 나타나 있다. 크롤링 결과 높은 점수대에서 매우 높은 찬반비를 나타냈으며 이는 예를 들어 표에서 10점 평점을 준 리뷰어는 204명이나 리뷰에 찬성함으로써 간접 참여한 사람들의 의견도 고려해서 가중치를 부여한다면 훨씬 더 높게 집계할 수 있음을 의미한다. 한편 블로그 텍스트의 동적 웹 크롤

링 결과 총155개의 블로그와 242,891 문자의 텍스트가 크롤링되었다.

표 2. 영화 액트 오브 킬링 평점 및 찬반비 요약⁷⁾

평점	리뷰수(비율,%)	찬성수	반대수	찬반비
10	204(52.3)	1995	262	7.6
9	81(20.8)	373	40	9.3
8	57(14.6)	97	11	8.8
7	22(5.6)	27	8	3.4
6	12(3.1)	14	20	0.7
5	3(0.8)	6	6	1.0
4	3(0.8)	5	17	0.3
3	0(0.0)	0	0	0
2	2(0.5)	4	9	0.4
1	6(1.5)	7	21	0.3
합계	390(100)	2528	394	-

2. 텍스트 마이닝

텍스트 마이닝 혹은 텍스트 데이터 마이닝이란 텍스트에서부터 정보를 자동으로 추출하여 이전에 알려지지 않았던 새로운 정보를 발견하는 것을 말한다. 텍스트 마이닝을 통해 표층적인 영화 선호도의 평가 뿐 아니라 텍스트 데이터 기반의 심층적인 의식 분석이 가능하며 분석의 신뢰도를 높일 수 있다. 텍스트 마이닝 기법에는 빈도 분석, 토픽 모델링, 감성 분석, 문서 자동 분류, 문장 자동 생성 등이 있다. 본 연구에서는 크롤링으로 수집한 텍스트 단어를 하나의 숫자에 매핑하는 방식인 BoW(Bag of Words)을 이용하여 딕셔너리 형태로 출력하는 함수를 만든 후 한국어 형태소 분석기, Okt(Open Korean Text)를 이용하여 명사를 추출하였다.

1) 형태소 분석

형태소 분석은 입력된 문장을 형태소 단위로 쪼개는 과정이다. 형태소란 언어학에서 일정한 의미가 있는 따로 떼어낼 수 가장 작은 단위로 한국어는 교착어(膠着語)로서 어근과 접사에 의해 단어의 기능이 결정되는 특성이 있기 때문에 자연어 처리에서 형태소 분석은 필수 과정이다. 네이버 블로그에서 크롤링한 아래의 예문을 보면 다음과 같이 하나의 문장, “조슈아 오픈하이머의 <엑트오브킬링>과 <침묵의 시선>이 그렇듯 <두 개의 문>과 이 <공동 정범> 역시 한 몸으로 다른 방향을 바라보고 있는 삼쌍둥이 같은 영화입니다”에 태깅된 각 어근과 접사의 형태소 구성을 파악할 수 있다.

(‘조슈아’, ‘Noun’), (‘오픈하이머’, ‘Noun’), (‘의’, ‘Noun’), (‘엑트’, ‘Noun’), (‘오브’, ‘Noun’), (‘킬링’, ‘Noun’), (‘과’, ‘Noun’), (‘침묵’, ‘Noun’), (‘의’, ‘Josa’), (‘시선’, ‘Noun’), (‘이’, ‘Noun’), (‘그렇듯’, ‘Adjective’), (‘두’, ‘Noun’), (‘개’, ‘Noun’), (‘의’, ‘Josa’), (‘문’, ‘Noun’), (‘과’, ‘Noun’), (‘이’, ‘Noun’), (‘공동’, ‘Noun’), (‘정범’, ‘Noun’), (‘역시’, ‘Noun’), (‘한’, ‘Verb’), (‘몸’, ‘Noun’), (‘으로’, ‘Josa’), (‘서로’, ‘Noun’), (‘다른’, ‘Noun’), (‘방향’, ‘Noun’), (‘을’, ‘Josa’), (‘바라보고’, ‘Verb’), (‘있’, ‘Noun’), (‘삼쌍둥이’, ‘Noun’), (‘같은’, ‘Adjective’), (‘영화’, ‘Noun’), (‘입니다’, ‘Adjective’)

7) <https://colab.research.google.com/drive/1q1ARhtqn9N0uoLSd8XN2b7UGGxuVq2gG?usp=sharing>참조

형태소 분석은 각 단어의 빈도 분석을 위한 토큰라이저(tokenizer)가 작동하기 위해서도 필요하지만 그 다음 단계의 맥락 분석을 위한 벡터화(feature vectorization)를 위해서도 선행되어야 한다.

2) 빈도 분석

형태소 분석 후, 불용어(stopwords) 및 중요도가 낮은 단어를 제거하는 전처리 과정을 거쳐 빈도 분석을 하였다. 불용어란 텍스트 데이터에서 빈번하게 출현하지만 의미 구성에는 큰 영향을 미치지 않는 불필요한 어휘를 말하며 자연어 처리 기술에서 유의미한 단어 만을 선별하기 위해서는 큰 의미가 없는 단어 토큰을 제거하는 데이터 전처리 작업에 해당된다. ‘영화’, ‘다큐멘터리’ 약 100여개의 불용어를 제거한 후 빈도 분석을 행하였다. 앞의 형태소 분석에서 가장 빈번히 등장하는 20개 키워드 명사를 출력하는 파이썬 코딩⁸⁾과 결과물은 다음과 같다.

```
noun_list = []
for sentence1 in sentences_tag:
    for word, tag in sentence1:
        if tag in ['Noun']:
            noun_list.append(word)
counts = Counter(noun_list)
print(counts.most_common(20))
data = nltk.Text(data, name='엑트오브킬링')
plt.figure(figsize=(12,8))
data.plot(20)
plt.show()
```

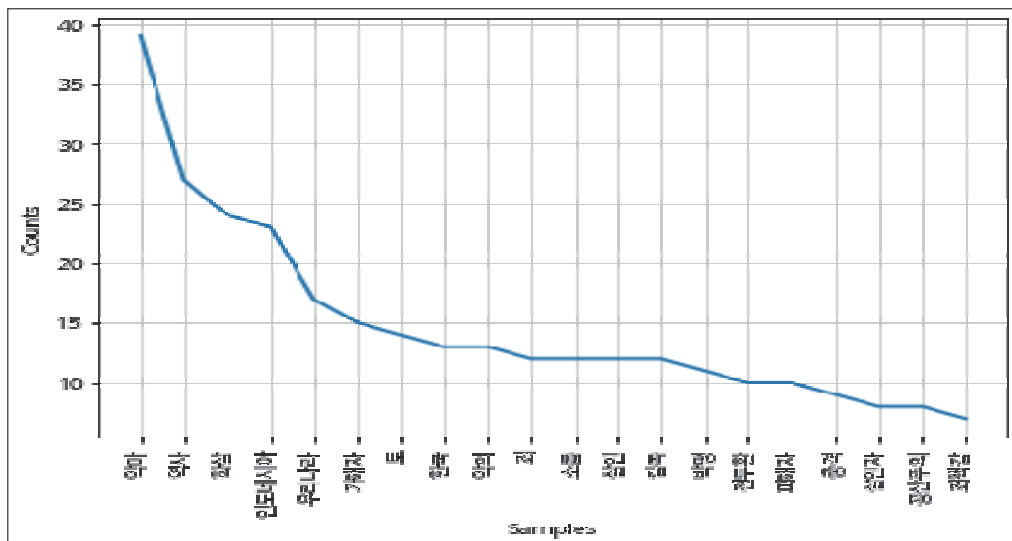


그림 3. 엑트 오브킬링 리뷰 텍스트 빈도 분석

8) <https://colab.research.google.com/drive/1VP56-d0iTpFFYbsvcr4vr1gd0caWvkl2?usp=sharing> 참조

코딩 프로그램을 접해보지 않은 학생들은 브라우저 방식으로 워드클라우드 생성 서비스를 제공하는 워드클라우드닷컴(www.wordclouds.com)과 같은 사이트에서 워드클라우드를 경험해보는 것으로 시작해 볼 수 있다. 단 무료로 제공하는 이 서비스는 데이터의 용량이 200KB대 이하만이 가능하며 형태소 혹은 품사별 단어의 취사선택의 한계가 있다. 따라서 다음 단계의 학습으로 코딩 프로그램을 통한 워드클라우드 분석으로 유도해 나가는 방식으로 프로그램을 기획할 수 있을 것이다.



그림 4. wordclouds.com의 워드클라우드



그림 5. 파이썬 코드로 실행한 워드클라우드

3) 맥락 분석

concordance(맥락) 모듈을 이용하면, 특정 단어가 사용된 문맥(context)이 어떤지 파악해 볼 수 있다. 여기에서 문맥은 해당 단어가 등장하는 같은 문장에서 사용된 단어들을 뜻한다. 본 연구에서는 형태소 분석한 블로그 텍스트에서 '민주화'라는 단어가 사용된 문장의 맥락(concordance)분석을 수행하였다. 코딩 결과, 토큰나이저로 쪼개진 52671개의 단어, 이 중 중복이 없는 8676개의 단어로 구성된 코퍼스라는 사실이 요약되어 출력되었으며 민주화라는 단어가 17번 문장에서 사용되었다는 의미이다.

```
data = nltk.Text(tokens_data, name = '엑트오브킬링블로그롤링')
print(len(data.tokens))
print(len(set(data.tokens)))
data.count('민주화')
data.concordance('민주화')
Displaying 17 of 17 matches
```

52671

8676

Displaying 17 of 17 matches:

실질 지도자 국가 고문 겸 외무 장관 아웅산 쯔 미얀마 독재 정부 민주화 운동 진행 미얀마 정치인 노벨평화상 미얀마 민주화 운동 상징 가택
미얀마 독재 정부 민주화 운동 진행 미얀마 정치인 노벨평화상 미얀마 민주화 운동 상징 가택 연금 석방 정치 활동 재개 그녀 승리 국가 자문 역
이 군부 지난 총선 부정선거 미로 민주 세력 정치인 대거 구금 사실 민주화 결과 미얀마 정치 상황 최근 다류멘터리 엑트 오브 킬링 수 작품 배
오브 킬링 수 작품 배경인 인도네시아 역시 군부독재 흔적 못 반쪽 민주화 그 이후 청산 못 과거 독재 부역 세력 견제 할 등 미얀마 정치 상
사회 질서 정의 실현 위 구국 결단 뿐 인도네시아 미얀마 마찬가지로 민주화 해도 그 사회 체제 헤게모니 군부독재 시절 별반 정권 교체 과거 과
순간 머리 속 사람 얼굴 아흔 나이 필드 골프채 통장 추정 금 납부 민주화 운동 학살 전두환 말 남 나라 볼 것 우리 과거 가해자 단죄 피해자
러 이권 불법 사업 개입 돈 세 어 두운 과거 현재 진행형 군부독재 민주화 과정 계속 우리 역사 대조 영화 볼 수 김재규 그 날 총 김영삼 전

이는 ‘민주화’라는 단어와 유사한 맥락인 ‘민주주의’, ‘민주’ 등의 단어 등은 제외한 숫자이다. 민주화를 언급한 문맥에서 광주 민주화 운동을 언급하는 사례를 찾아볼 수 있었으며 최근에 작성된 블로그에서는 미얀마 사태를 연상하는 글들을 볼 수 있었다. 결과물은 표 3과 같다.

표 3. ‘민주화’ 단어가 들어가는 맥락 분석

수록 일자	내용
21년 3월	"미얀마 독재 정부에 맞서 민주화운동을 진행했던 미얀마의 정치인", "이번 미얀마 사태를 보고 <엑트 오브 킬링>이 생각났는데"
21년 2월	"미얀마의 정치상황을 보며 최근 본 다큐멘터리 <엑트오브킬링>을 떠올리지 않을 수 없었다. 작품의 배경인 인도네시아 역시 군부독재의 흔적을 완전히"
20년 7월	"나는 그 이유를 우리네와 비슷한 처지에 있었고", "지금은 민주화가 되어있는 인도네시아이지만"
20년 5월	"현재 광주 민주화 운동을 바라보며 우리가 민주화 운동을 대하는 방식을 보고 배우고 있다고 전하고 있습니다."
19년 11월	"이 다큐멘터리가 다루는 인도네시아 대학살 뿐 만 아니라 홀로코스트, 5.18 광주 민주화 운동 등 인류의 역사에 끔찍한 결과를 낳은 사건들을 다룬"
19년 5월	"안와르 콩고는 자신의 다양한 학살방법을 자신이 향유했던..." "최근 5.18을 기록한 영화 <김군>이 개봉하였다." "<엑트 오브 킬링>이 승자들의 역사 속에서 학살자의 모습을 담아냈다면, <김군>은 피해자들의 39년..."

4) 감성 분석

감성 분석은 단어와 문맥을 분석하여 텍스트의 감정을 파악하는 기술로서 자연어 처리 및 텍스트 마이닝 기술을 활용하여 사용자의 리뷰 및 설문 조사 응답, 미디어 분석 등 다양한 용도로 활용된다. 감성 분석은 상품 및 서비스의 사용자의 의견을 도출해내는 것이기 때문에 오피니언 마이닝(opinion mining)이라 하기도 한다. 본 연구에서 영화 리뷰에 대한 감성 분석은 크롤링한 네이버 영화리뷰 텍스트를 두 가지 방법으로 행하였는데 첫번째는 구글 클라우드 API의 감성분석기⁹⁾를 이용한 방법과 두번째는 지도학습 로지스틱 회귀분석의 방법으로 행하였다.

구글 클라우드 API 감성분석기는 호출자가 별도로 언어 지정을 하지 않아도 한국어를 감지하고 반환한다. 앞서 웹사이트에서 접속하여 자동적으로 출력된 워드클라우드처럼 코딩을 전혀 하지 않고도 분석할 텍스트를 구글 클라우드 API 사이트에 접속해서 입력창에 입력하기만 하면 자동적으로 분석값이 출력된다. 이 분석기의 또 하나의 장점은 텍스트 내의 감성을 긍정, 부정, 중립의 세 가지 분류 뿐 아니라 숫자 스코어(score) 및 크기(magnitude) 값으로 표현한다는 것이다. 감정 스코어의 범위는 -1.0(부정적)에서 1.0(긍정적)까지이며 텍스트의 전반적인 정서 성향을 나타내는데 긍정(0.25~1.0)은 초록색, 중립(-0.25~0.25)은 노란색, 부정(-0.25~-1.0)은 빨간 색으로 표시된다.

9) Natural Language AI, <https://cloud.google.com/natural-language>

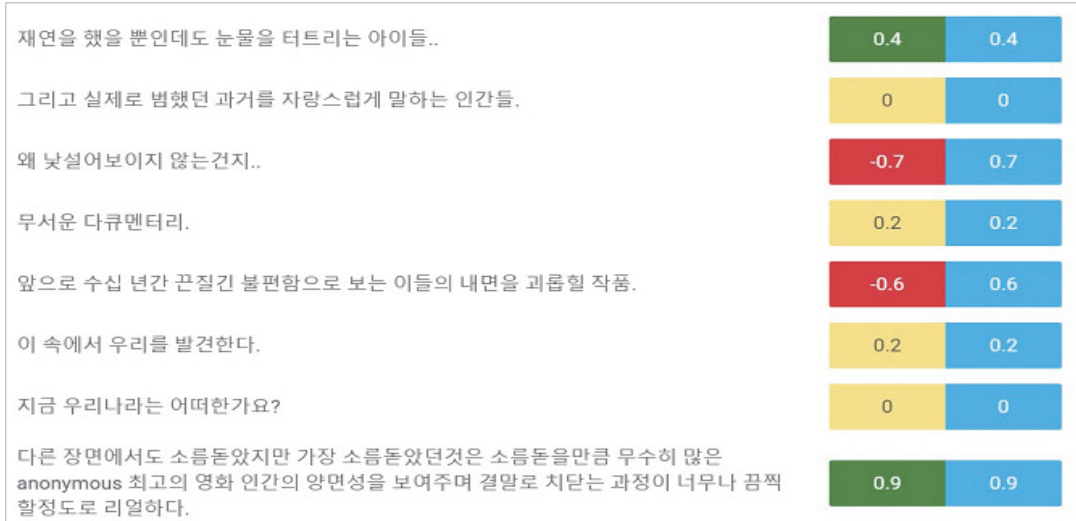


그림 6. 구글 클라우드 API 감성분석기를 활용한 감성 분석

문서의 감성 스코어는 문서의 전반적인 정서를 나타내는 +-의 기호를 가지는 반면 크기(magnitude)는 텍스트에 얼마나 많은 감정이 담겨 있는지 양(quantity)을 나타내기 때문에 절대값을 가지며 이번 영화리뷰 텍스트 분석에서는 다음과 같이 스코어의 절대값이 크기의 결과값으로 나왔다.

그림 6과 같이 구글 감성분석기는 영화 리뷰 용도로 만들어지지 않았기에 긍정적인 어휘 혹은 부정적인 어휘에 따른 분류가 되며, 대체로 수궁이 가는 스코어 결과값이 나왔으나 한국어 중 은어를 활용한 표현, 예를 들면 그림 7과 같이 ‘핵노잼’은 긍정(0.3)으로 분석하는 오류를 범하고 있음을 알 수 있다.

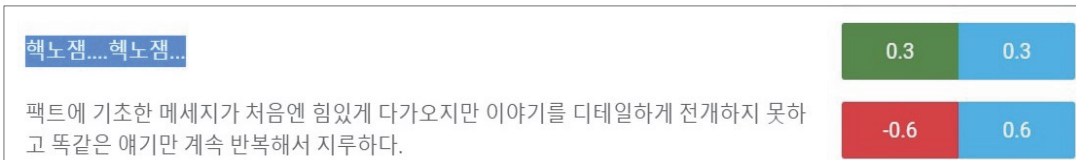


그림 7. 구글 클라우드 API 감성분석기를 활용한 감성 분석, ‘핵노잼’

구글 API 감성분석기는 위와 같이 스코어가 표시되므로 스코어 데이터를 이용한 데이터 시각화 학습을 유도할 수 있다. 그림 8은 구글 클라우드 API 감성분석기를 이용한 트리맵 그래프이며 그림 9는 스코어를 8단계로 좀 더 세분화하여 트리맵 그래프를 그려본 결과이다. 트리맵 그래프는 분류의 세부 단계가 아주 많지 않을 때 유용하며, 유명한 화가 몬드리안의 그림처럼 몇 가지 특징적 컬러로 나타낼 수 있어 학생들에게 흥미를 유도할 수 있다.

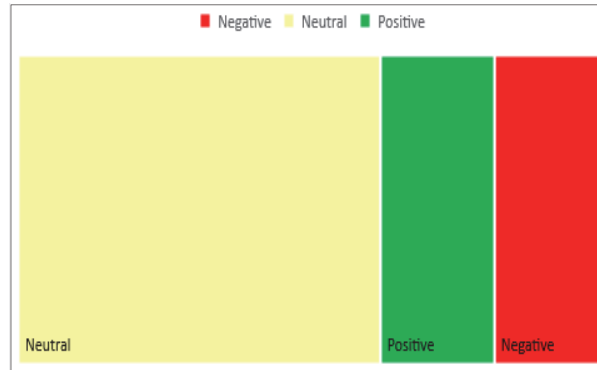


그림 8. 구글 감성분석 수치를 활용한 데이터 시각화(3단계)

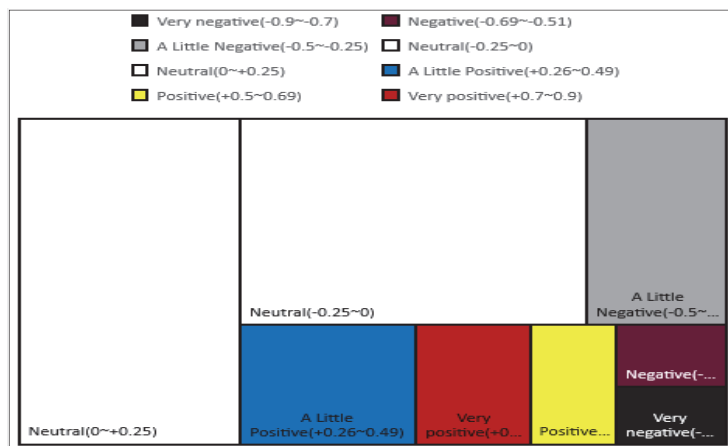


그림 9. 구글 감성분석 수치를 활용한 데이터 시각화(8단계)

두 번째 방법의 감성 분석은 머신러닝 지도학습으로 행하였는데 초급 이상의 파이썬 코딩을 할 수 있거나 모듈을 활용하고 이해할 수 있는 학습자라면 가능한 방법이다. 데이터 분석방법은 로지스틱 회귀분석(logistic regression)의 방법으로 행하였으며 사이킷런(scikit-learn) 라이브러리를 이용하였다. 로지스틱 회귀분석은 0과 1로 라벨링한 감성 분류와 같이 결과가 범주형일 때 분류 알고리즘으로 사용되며 사이킷런은 파이썬용 머신러닝 라이브러리로서 지도 학습 및 비지도 학습 모듈 등이 있는데 본 연구에서는 지도학습 모듈을 사용하였다. 데이터는 네이버 영화 리뷰 코퍼스 약 20만개의 감성 라벨링 데이터 중 10,000개를 취하여 8,000개는 학습용 데이터로 2,000개는 테스트용 데이터로 사용하였다. 비정형 텍스트 데이터를 숫자형 피쳐(feature)기반으로 추출하여 의미 있는 값을 부여하기 위해서 tf-idf를 이용한 피쳐 벡터화(feature vectorization)와 피쳐 추출(feature extraction)을 진행한다. tf-idf 기법은 개별 문서에서 자주 나타나는 단어에 높은 가중치를 주되, 전체 문서에서 자주 나타나는 단어에는 페널티를 부여하는 방식으로 피쳐를 벡터화한다. 최종적으로 최적화된 하이퍼파라미터(hyperparameter)를 찾는 것이 목표이며, 이후 입력된 벡터 값의 좌표의 위치에 따라 해당 클래스를 반환한다. 다음은 사이킷런 라이브러리에서 로지스틱 회귀분석 함수를 import한 후 tf-idf를 이용한 피쳐를 벡터화해서 최적화하는 과정의 알고리즘이다.

```

from sklearn.linear_model import LogisticRegression
SA_lr = LogisticRegression(random_state = 0)
SA_lr.fit(nsmc_train_tfidf, nsmc_train_dff['label'])
from sklearn.model_selection import GridSearchCV
params = {'C': [1, 3, 3.5, 4, 4.5, 5]}
SA_lr_grid_cv = GridSearchCV(SA_lr, param_grid=params, cv=3, scoring='accuracy',
verbose=1)
SA_lr_grid_cv.fit(nsmc_train_tfidf, nsmc_train_dff['label'])
SA_lr_best = SA_lr_grid_cv.best_estimator_

```

본 연구에서는 네이버 영화 감성 분류 데이터를 가지고 머신러닝을 학습한 후 크롤링한 영화 <엑트 오브 킬링> 리뷰에 대한 감성 분석을 예측해 보는 것까지 실행하였다. 구글 감성분석기의 결과와는 달리 ‘핵노잼’과 같은 은어적 표현을 정확하게 부정으로 분류한 결과를 볼 수 있다.

```

st = input("안녕하세요? 저는 무비센스봇입니다. 영화 <엑트오브킬링>에 대한 감상평을 입력해주세요~~")
# 0) text pre-processing
st = re.compile(r'[ㄱ-힣]+').findall(st); print(st)
st = [" ".join(st)] ; print(st)
# 1) new text feature vectorization
st_tfidf = tfidf.transform(st)
# 2) optimal sentiment analysis model evaluation
st_predict = SA_lr_best.predict(st_tfidf)
# 3) print sentiment prediction
if(st_predict== 0):
    print(st, "->> 부정감성 😞")
else:
    print(st, "->> 긍정 감성 😊")
sentiment_analysis_accuracy : 0.815
안녕하세요? 저는 무비센스봇입니다. 영화 <엑트오브킬링>에 대한 감상평을 입력해주세요~~핵노잼
['핵노잼']
['핵노잼'] ->> 부정감성 😞

```

그림 9. 머신러닝 분류 학습을 통한 감성 분석

III. 연구 결과

본 연구를 통해 <엑트 오브 킬링> 네이버 영화 리뷰 코퍼스에서 가장 빈번하게 등장하는 유의미한 키워드는 ('악마', 39), ('역사', 27), ('학살', 24), ('인도네시아', 23), ('우리나라', 17), ('가해자', 15), ('토', 14), ('한국', 13), ('악의', 13), ('죄', 12), ('소름', 12), ('살인', 12), ('감독', 12), ('박평', 11), ('전환환', 10), ('피해자', 10), ('충격', 9), ('살인자', 8), ('공산주의', 8), ('죄책감', 7) 순으로 나타났다. 학생들은 웹사이트에 접속해서 코딩 없이 워드클라우드를 경험해 볼 수 있으며 코딩 프로그램을 통한 형태소 분석을 거친 명사의 빈도를 시각화한 워드클라우드와 차이를 비교 분석해 볼 수 있다. 빈도가 높은 키워드를 살펴 보면 영화에서 다루어진 인도네시아의 잔인하고 충격적인 학살 사건을 통해 한국인들이 한국의 민주화 운동 및 역사적 사건을 연상한다는 사실을 알 수 있다.

맥락(concordance) 분석은 네이버 영화 리뷰에 비해 자유로운 글쓰기가 가능한 블로그에서 크롤링된 텍스트를 분석했다. 관찰하고자 하는 핵심 키워드 중의 하나인 ‘민주화’가 들어간 문장 맥락을 통해 최근의 미얀마 사태와 역시 광주 민주화 운동 등이 함께 기술되어 있음을 확인하였다.

감성 분석은 구글 클라우드 API의 감성분석기를 활용하여 출력된 스코어 값을 이용하여 트리맵 차트를 비롯한 다양한 데이터 시각화 연습을 할 수 있다. 또한 머신러닝의 로지스틱 리그레션 알고리즘을 사용하였으며 연민과 부정 감성으로 라벨링한 학습용 데이터와 테스트 데이터로 지도학습으로 실행하였으며 81.5%의 정확도를 보였다. 학생들은 최종적으로 영화 리뷰 감성분석 로봇, 무비센스봇이 “저는 무비센스봇입니다. 무엇을 도와드릴까요?”라고 질문하고 챗봇 사용자가 대답하면 감성 분석 예측 결과가 출력되었다.

IV. 결론

본 연구는 머신러닝 기술 중의 하나인 텍스트 마이닝 기법과 데이터 시각화를 통해 다큐멘터리라는 시사적인 영상 미디어를 분석하고 이해하고자 하였다. 소셜미디어에 게재된 영화 관련 리뷰 글들은 사람들의 영화에 관한 생각과 느낌, 시간적·사회적 맥락의 이야기를 담고 있는 비정형 데이터로서, 텍스트 마이닝 기법은 이러한 특성을 가진 비정형 데이터의 분석 방법론으로서 여전히 효과적인 도구라는 사실을 확인하였다. 네이버 영화 섹션과 블로그에 수록된 글들은 데이터 접근성이 높기 때문에 데이터 분석 경험이 많지 않은 학생들도 데이터를 수월하게 수집할 수 있으며 텍스트 마이닝 기법 또한 비교적 활용하기 용이한 데이터 분석 방법론이다. 본 연구에서는 콜브(Kolb)의 경험 학습 이론을 소개하고, 일련의 데이터 분석 및 시각화 과정을 설명하였는데 이는 바로 능동적 실험(learning by doing)을 하기 위한 실험 절차에 해당된다. 텍스트 데이터를 수집하고 워드 클라우드를 통해 도출된 데이터를 관찰하고 그 함의를 분석하고 개념화하며 맥락 및 감성을 분석하는 과정이 그것이다. 단지 영화를 관람하고 감상평을 토론하는 학습과 영화 관람 후 텍스트 마이닝 기법을 이용해서 미디어 리뷰 텍스트를 분석하고 데이터 시각화까지 실습한 학생들의 경험 지식은 다를 수밖에 없다. 파이썬 코딩과 머신러닝 데이터 분석 툴에 대한 사전 기초지식이 없는 학생들은 본 연구에 소개된 오픈 소스 웹 사이트 기반의 실습만으로는 STEAM 융합 학습의 효과가 제한될 것이다. 따라서 다음 단계인 파이썬 코딩을 통한 텍스트 마이닝 기법 학습으로 진행되도록 교사의 지도가 필요할 것으로 사료된다. 또한 여러 다큐멘터리 영화들의 텍스트마이닝을 통해서 국내외 역사적인 사건을 비교한다든가 혹은 같은 영화에 대해 국내외 리뷰어들의 반응을 비교 분석해보는 학습 방법도 고려될 수 있을 것이다. 이러한 능동적인 경험 학습을 통해 다큐멘터리라는 대중매체에 대한 영화 평점이나 전문가 비평과 같은 다소 피상적인 검색에서 한 단계 나아가서 텍스트 마이닝이라는 분석기법과 친숙해지고 미디어 리터러시 능력이 제고되기를 바란다.

참고문헌

- 김진한 (2020). “콜브(D. Kolb)의 경험학습이론에 대한 해석학적 연구”, 서울대 교육학석사학위논문.
- 김하연, 서대호 (2020). “텍스트 마이닝을 활용한 국외 현대무용 연구 트렌드 변화 분석”, *예술교육연구*, 18(1), 171 ~ 192.
- 김하중(2019), “초·중등 소프트웨어교육 운영실태와 개선과제”, *입법·정책보고서*, (34).
- 노들, 옥현진 (2020). “텍스트 마이닝 기법을 통한 미디어 리터러시와 디지털 리터러시 개념의 비교 분석 - 신문 기사를 중심으로”, *한국리터러시학회*, (5), 103-129.
- 여신, 김바로 (2020). “우리는 서발틴의 목소리를 들을 준비가 되었는가? : 텍스트 마이닝을 통한 영화 <82년생 김지영>리뷰 분석”, *인문콘텐츠학회*, (58), 77-103.
- 이서빈 (2018). “파이썬을 활용한 수학교과 코딩수업(DM³)의 효과”, *대한수학교육학회지 수학교육학연구*, 479 -499.
- 이송하, 유재봉 (2021). “문화예술교육에서의 예술교육자의 미적경험에 대한 질적 연구: 듀이의 경험론을 중심으로”, *한국예술교육학회*, *예술교육연구*, 19(2), 1 ~ 17.
- 임정택 (2017). “제노사이드 영화와 사회치유 -조슈아 오펜하이머의 다큐멘터리를 중심으로”, *탐라문화*, 105-141.
- Jason West(2017), Validating curriculum development using text mining, *The Curriculum Journal*, 28(3), 389-402.
- Li Zhuang et al (2006). “Movie Review Mining and Summarization”, *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, 43-50.
- Teemu Valtonen et al (2019). "Media Literacy Education in the Age of Machine Learning", *Journal of Media Literacy Education*, (2), 20-36.

ABSTRACT

Media Literacy as a Pathway to STEAM Education Using Text Mining Methodology: a Focus on < The Act of Killing >

Kang, Boyoung · Hong, Munpyo(Sungkyunkwan University)

This study proposes one of the experiential learning (learning by doing) methodologies using text mining technology of natural language processing and data visualization. The media environment has changed rapidly in the 21st century, and the basis of the change was the transition to a platform and breakthrough development of ICT technologies such as artificial intelligence, big data, and cloud. Therefore, media literacy is intertwined with the changing times and it refers to the ability to grasp the meaning of messages produced by the various media and to participate critically in the media and society. In this study, I tried to analyze the public consciousness and implications expressed through social media for documentary film using text mining technique, which is one of the natural language processing technologies. Movie reviews posted on social media are easy to collect and analyze, and when evaluated through text mining, they have a new and important value, which is different from criticism or evaluation by some experts. In this study, people who can code Python above the beginner level and have the knowledge of machine learning can follow the step by step approaches. However those who cannot code also experience text mining by the live word cloud generator and Google Cloud open API. In this paper, the research process was conducted in the following steps: 1) web crawling and dynamic crawling related to the <Act of Killing> movie in Naver, Korea's number one search engine, 2) morphological analysis, 3) word frequency analysis, 4) context analysis, and 5) sentiment analysis. The comments about democratization in Korea, such as the Gwangju Democratization Movement, and the recent incident in Myanmar was uncovered through the text mining of this research. This research suggests media literacy based STEAM education that transcends the boundaries of arts and any specific disciplines through text mining and data science technology.

Key words : text mining, media literacy, the act of killing, STEAM education, data visualization, web crawling, sentiment analysis

접수일 : 2021년 8월 1일

심사완료일 : 2021년 8월 23일

게재확정일 : 2021년 9월 1일

