학사학위논문

자연어 처리를 이용한 텍스트 감성 분석

영화 리뷰 긍정, 부정 판단 알고리즘

2019. .

忠 北 大 學 校

經營情報 專攻

박 지 수 (朴智秀)

학사학위논문

자연어 처리를 이용한 텍스트 감성 분석

영화 리뷰 긍정, 부정 판단 알고리즘

2019. .

忠 北 大 學 校

經營情報 專攻

박 지 수 (朴智秀)

자연어 처리를 이용한 텍스트 감성 분석

지도교수 조 완 섭

이 논문을 학사학위 논문으로 제출함

20 年 月

충 북 대 학 교

빅데이터 연계 전공

박 지 수

홍길동의 학사학위논문을 인준함.

지도교수 조 완 섭 (인)

20 年 月

충북대학교 경영정보학과

목 차

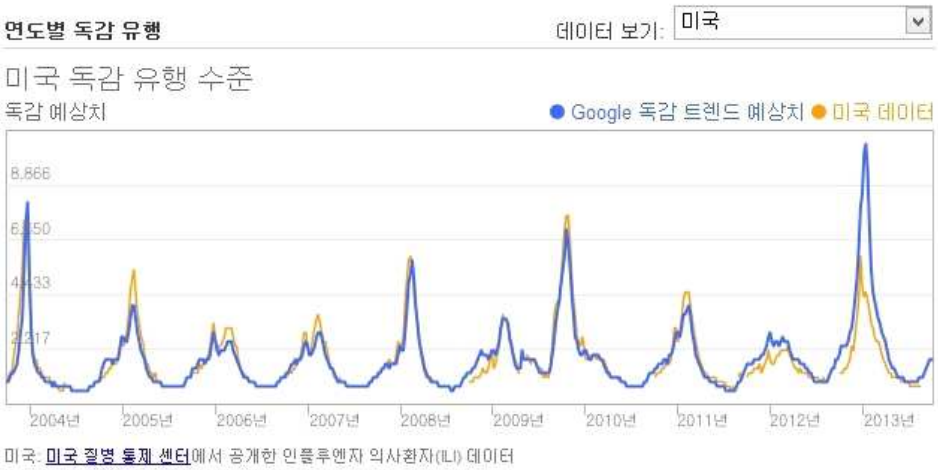
표 목 차

그 림 목 차

1. 서론

1.1 연구 배경

현재 우리는 1분 1초마다 새로운 정보가 쏟아지는 정보의 홍수 속에서 살아가고 있다. 각종 SNS와 뉴스를 통해 텍스트 데이터가 쏟아지고 이를 분석한다면 사회현상이나 이슈 등을 파악할 수 있으며, 이를 다방면에 사용할 수 있다. 대표적으로 구글 검색되는 '독감' 키워드로 독감 발생 지역과 경로를 파악했던 사례가 있다.



<그림1-1> 구글의 독감 예측

이처럼 개인이 인터넷상에 남긴 흔적들이 의미 없이 저장 공간을 차지하고 있는 글자가 아닌 현실 세사의 흐름과 이슈를 읽을 수 있는 지표로서 활용되기를 바란다. 이렇게 정제하고 처리한 유의미한 데이터를 사회적, 경제적으로 이용할 수 있는 방법을 강구하는 것도 연구를 진행한 배경이다. 또한 그저 텍스트의 등장 빈도나, 기간 등의 정형적인 통계 자료가 아닌 인간의 심리와 감정에 대한 단어를 추출하여 텍스트 안의 인간의 감정을 인식할 수 있는지에 대한 가능성을 확인해보고자 한다.

1.2 연구 목적

현재 감성 분석은 감성 사전을 미리 제작하여 해당 단어가 어느 감정을 뜻하는지 매치하는 방법을 주로 이용한다. 하지만 이는 새로운 언어나, 함축어, 은어를 포함하기 어렵고 모든 한글 단어를 감정 단어로 분류해야하는 번거로운 작업이 필요하다. 따라서 본 연구

를 통해 영화 리뷰와 같이 텍스트의 감정을 라벨링 할 수 있는 다양한 데이터를 수집하여 보다 정확한 판단 모델을 만들기 위한 발판으로 삼으려고 한다. 이후 인간이 느끼는 감정을 크게 6가지로 분류한 폴 에크만의 일차감정(Big six)으로 판단 범위를 넓혀 사회적 현상을 텍스트 데이터로 이해하고, 예측하기 위함을 목적으로 한다. 가시적인 목표로는 영화 리뷰를 분석하여 영화 평가에 미치는 요인을 분석하고 향후 영화 제작과 마케팅 분야에서 이를 이용하여 최대한의 이윤을 창출 하고자 한다. 더 나아가 영화뿐만 아니라 사회적, 정치적인 주제에 대한 사람들의 생각을 즉각적으로 파악할 수 있는 방법으로 확장시키려고 한다.

1.3 연구 필요성

특정 현상이나 사회적 상황에 대한 사람들의 의견을 듣기 위해 통계조사, 설문조사를 진행한다. 하지만 이는 현실시대에는 뒤떨어진 비효율적이며, 비용이 많이 드는 방법으로 이를 해결한다면 사회적 비용을 줄일 수 있을 것이라고 생각한다. 텍스트 분석과 감성분석을 통해 실시간으로 수집된 데이터를 이용해 현 정부의 정책, 혹은 특정 상품에 대한 개인의 선호를 파악하여 피드백으로 삼거나, 앞으로 발생할 국민들의 반응을 예측하여 결과를 수정할 수 있다. 또한 인터넷 속에서 존재하는 방대한 양의 텍스트들이 아무 없이 그저 저장되어 있는 것은 죽은 것과 다름이 없고, 앞으로 이런 데이터를 저장하기 위한 공간과 인력이 필요하게 될 것이다. 그렇기에 이것이 저장되어야 하는 경제적이고 사회적인 이유를 찾기 위해 텍스트 분석과 자연어처리가 진행될 필요가 있다.

2. 관련 연구

2.1 텍스트 분석, 자연어 처리

텍스트 분석이라는 용어 는 비즈니스 인텔리전스, 탐색 적 데이터 분석, 연구 또는 조사를 위해 텍스트 소스의 정보 내용을 모델링하고 구성하는 언어, 통계 및 기계 학습 기술 세트를 설명한다. 이 용어는 텍스트 마이닝과 대략 동의어이다. 실제로 Ronen Feldman은 "텍스트 분석"을 설명하기 위해 2004 년 "텍스트 마이닝" 에 대한 2000 개의 설명을 수정했다. "텍스트 마이닝은" 1980 년대로 거슬러 올라가는 초기 응용 분야의 일부에 사용되는 반면 후자의 용어는 이제 비즈니스 설정에서 더 자주 사용된다. 특히 생명 과학

연구와 정부의 정보에 사용된다.

Hotho et al.에 따르면 (2005) 우리는 텍스트 마이닝 의 세 가지 관점, 즉 정보 추출로서의 텍스트 마이닝, 텍스트 데이터 마이닝으로서의 텍스트 마이닝 및 KDD (데이터베이스의 지식 발견) 프로세스 인 텍스트 마이닝을 다르게 할 수 있다. 텍스트 마이닝은 “다른 서면 리소스에서 정보를 자동으로 추출하여 이전에 알려지지 않은 새로운 정보를 컴퓨터에서 발견 한 것”이다. 서면 자료는 웹 사이트, 서적, 이메일, 리뷰, 기사 일 수 있다. 고품질 정보는 일반적으로 통계적 패턴 학습과 같은 수단을 통한 패턴 및 경향의 고안을 통해 도출된다. 텍스트 마이닝은 일반적으로 입력 텍스트 구조화를 포함한다. 일반적인 텍스트 마이닝 작업에는 텍스트 분류, 텍스트 클러스터링, 개념 및 엔티티의 추출, 세분화된 분류 체계 생성, 감성 분석, 문서 요약 및 엔터티 관계 모델링이 포함된다.

텍스트 분석에는 정보 검색, 단어 빈도 분포, 패턴 인식, 태깅/주석, 정보 추출, 링크 및 연관 분석, 시각화 및 예측 분석을 포함한 데이터 마이닝 기술 을 연구하는 어휘 분석 이 포함된다. 가장 중요한 목표는 기본적으로 자연어 처리(NLP), 다양한 유형의 알고리즘 및 분석 방법 을 적용하여 텍스트를 분석 할 데이터로 변환 하는 것이다. 이 프로세스의 중요한 단계는 수집 된 정보의 해석이다.

자연어 처리 또는 자연 언어 처리는 인간의 언어 현상을 컴퓨터와 같은 기계를 이용해서 모사 할 수 있도록 연구하고 이를 구현하는 인공지능의 주요 분야 중 하나다. 자연 언어 처리는 연구 대상이 언어이기 때문에 당연하게도 언어 자체를 연구하는 언어학과 언어 현상의 내적 기재를 탐구하는 언어 인지 과학과 연관이 깊다. 구현을 위해 수학적 통계적 도구를 많이 활용하며 특히 기계학습 도구를 많이 사용하는 대표적인 분야이다. 정보검색, QA 시스템, 문서 자동 분류, 신문기사 클러스터링, 대화형 Agent 등 다양한 응용이 이루어지고 있다.

자연 언어 처리에서 말하는 형태소 분석이란 어떤 대상 어절을 최소의 의미 단위인 '형태소'로 분석하는 것을 의미한다. 정보 검색 엔진에서 한국어의 색인어 추출에 많이 사용한다. 형태소 분석 단계에서 문제가 되는 부분은 미등록어, 오탈자, 띄어쓰기 오류 등에

의한 형태소 분석의 오류, 중의성이나 신조어 처리 등이 있는데, 이들은 형태소 분석에 치명적인 약점이라 할 수 있다. 복합 명사 분해도 형태소 분석의 어려운 문제 중 하나이다. 복합 명사란 하나 이상의 단어가 합쳐서 새로운 의미를 생성해 낸 단어로 '봄바람' '정보검색' '종합정보시스템' 등을 그 예로 들 수 있다. 이러한 단어는 한국어에서 띄어쓰기에 따른 형식도 불분명할 뿐만 아니라 다양한 복합 유형 등에 따라 의미의 통합이나 분해가 다양한 양상을 보이기 때문에 이들 형태소를 분석하는 것은 매우 어려운 문제이다. 일반적으로, 다양하게 쪼개지는 분석 결과들 중에서 적합한 결과를 선택하기 위해, 테이블 파싱이라는 동적 프로그래밍 방법을 사용한다.

2.2 관련 패키지 소개

2.2.1 TF-IDF

TF-IDF(Term Frequency - Inverse Document Frequency)는 정보 검색과 텍스트 마이닝에서 이용하는 가중치로, 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다. 문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정하거나, 문서들 사이의 비슷한 정도를 구하는 등의 용도로 사용할 수 있다.

TF(단어 빈도, term frequency)는 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값으로, 이 값이 높을수록 문서에서 중요하다고 생각할 수 있다. 하지만 단어 자체가 문서군 내에서 자주 사용되는 경우, 이것은 그 단어가 흔하게 등장한다는 것을 의미한다. 이것을 DF(문서 빈도, document frequency)라고 하며, 이 값의 역수를 IDF(역문서 빈도, inverse document frequency)라고 한다. TF-IDF는 TF와 IDF를 곱한 값이다.

IDF 값은 문서군의 성격에 따라 결정된다. 예를 들어 '원자'라는 낱말은 일반적인 문서들 사이에서는 잘 나오지 않기 때문에 IDF 값이 높아지고 문서의 핵심어가 될 수 있지만, 원자에 대한 문서를 모아놓은 문서군의 경우 이 낱말은 상투어가 되어 각 문서들을 세분화하여 구분할 수 있는 다른 낱말들이 높은 가중치를 얻게 된다.

TF-IDF는 단어 빈도와 역문서 빈도의 곱이다. 두 값을 산출하는 방식에는 여러 가지가

있다. 단어 빈도 tf(t,d)의 경우, 이 값을 산출하는 가장 간단한 방법은 단순히 문서 내에 나타나는 해당 단어의 총 빈도수를 사용하는 것이다. 문서 d 내에서 단어 t의 총 빈도를 f(t,d)라 할 경우, 가장 단순한 tf 산출 방식은 tf(t,d) = f(t,d)로 표현된다. 그 밖에 TF값을 산출하는 방식에는 다음과 같은 것들이 있다.

불린 빈도: tf(t,d) = t가 d에 한 번이라도 나타나면 1, 아니면 0
로그 스케일 빈도: tf(t,d) = log (f(t,d) + 1)
증가 빈도: 최빈 단어를 분모로 target 단어의 TF를 나눈 값으로, 일반적으로는 문서의 길이가 상대적으로 길 경우, 단어 빈도값을 조절하기 위해 사용한다.

$$\mathrm{tf}(t,d) = 0.5 + \frac{0.5 \times \mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

역문서 빈도는 한 단어가 문서 집합 전체에서 얼마나 공통적으로 나타나는지를 나타내는 값이다. 전체 문서의 수를 해당 단어를 포함한 문서의 수로 나눈 뒤 로그를 취하여 얻을 수 있다.

$$\mathrm{idf}(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

|D|: 문서 집합 D의 크기, 또는 전체 문서의 수
$|\{d \in D : t \in d\}|$: 단어 t가 포함된 문서의 수.(즉, $tf(t,d) \neq 0$). 단어가 전체 말뭉치 안에 존재하지 않을 경우 이는 분모가 0이 되는 결과를 가져온다. 이를 방지하기 위해 $1+|\{d \in D : t \in d\}|$로 쓰는 것이 일반적이다.

TF-IDF는 다음과 같이 표현된다.

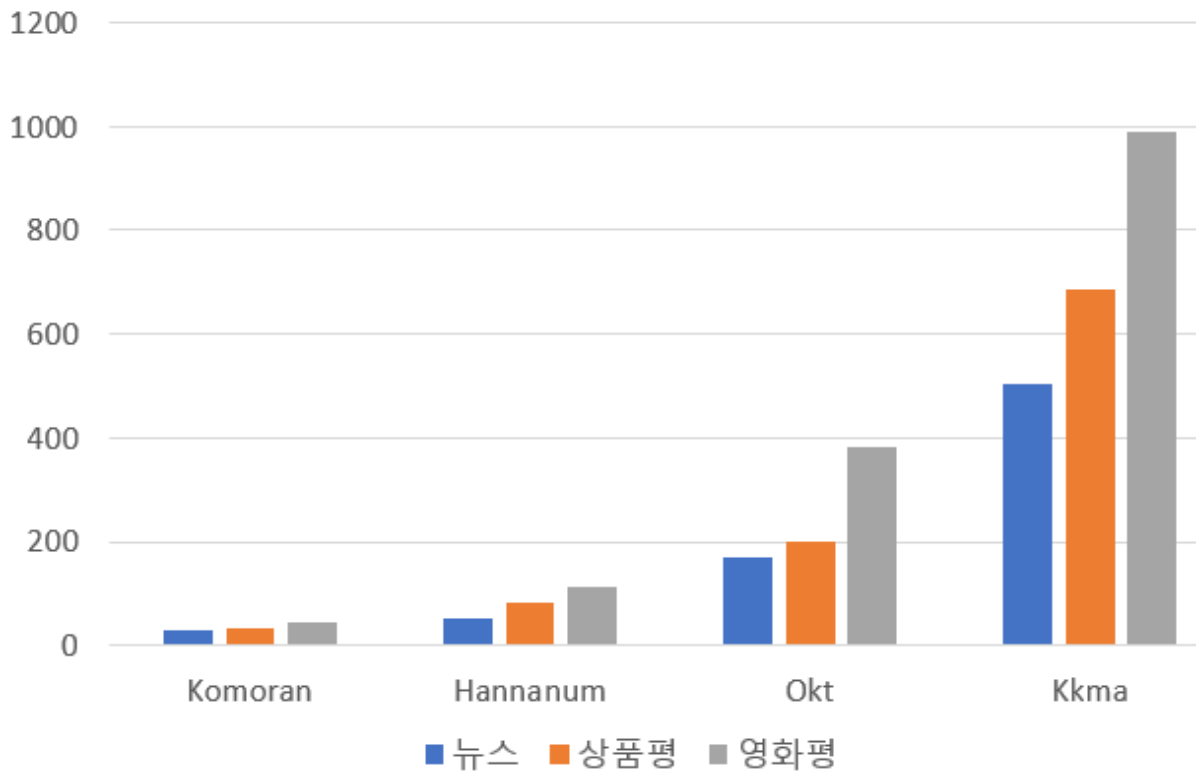
$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

특정 문서 내에서 단어 빈도가 높을수록, 그리고 전체 문서들 중 그 단어를 포함한 문서

가 적을수록 TF-IDF값이 높아진다. 따라서 이 값을 이용하면 모든 문서에 흔하게 나타나는 단어를 걸러내는 효과를 얻을 수 있다. IDF의 로그 함수값은 상 1 이상이므로, IDF값과 TF-IDF값은 항상 0 이상이 된다. 특정 단어를 포함하는 문서들이 많을수록 로그 함수 안의 값이 1에 가까워지게 되고, 이 경우 IDF값과 TF-IDF값은 0에 가까워지게 된다.

2.2.2 Konlpy (Kkma, Okt)

NLP (Natural Language Processing, 자연어처리)는 텍스트에서 의미 있는 정보를 분석, 추출하고 이해하는 일련의 기술 집합이며, KoNLPy는 한국어 정보처리를 위한 파이썬 패키지이다. Kkma는 SNU 의 지능형 데이터 시스템 (IDS) 연구소에서 개발 한 Java로 작성된 형태소 분석기 및 자연어 처리 시스템이다. Okt(구 Twitter Korean Text)는 Will Hohyon Ryu가 개발 한 스칼라로 작성된 오픈 소스 한국어 토크 나이저이다. 본 연구는 Okt를 이용해 분석을 진행했고, 이유는 아래의 그래프가 나타내듯, Okt의 처리시간이 상대적으로 빠르기 때문이다.



<그림2-1> 형태소 분석기 처리 시간 비교

<사진2>의 x축은 각 형태소 분석기의 이름이며, y축은 뉴스, 상품평, 영화평 10,000건에 대한 처리 시간(단위: 초)이다.

3. 연구 설계 및 방법

3.1 데이터 수집과 전처리

3.1.1 데이터 수집, 크롤링

본 연구에서는 감성분석 모델 구축을 위해 지도학습 방법을 이용하기 때문에 학습용 데이터와 실험 데이터가 필요하다. 이에 국내 최대 포털 네이버에서 운영 중인 사이트 '네이버 영화(movie.naver.com)'에서 데이터를 수집하였다. 데이터는 새로운 영화가 개봉하거나, 순위 변동 시 코드를 수정할 필요 없이 즉각적으로 분석에 이용하기 위해 크롤러를 사용하였다. 크롤링을 위해 사용한 파이썬 패키지는 BeautifulSoup, requests, parse 등이다. 먼저 Beautiful Soup 은 HTML 및 XML 문서를 구문 분석하기 위한 Python 패키지이다. 웹 스크래핑에 유용한 HTML에서 데이터를 추출하는 데 사용할 수 있는 구문 분석 된 페이지에 대한 구문 분석 트리를 작성한다. requests는 Apache2 라이센스에 따라 릴리스 된 Python HTTP 라이브러리다. 이 프로젝트의 목표는 HTTP 요청을 더 단순하고 인간 친화적으로 만드는 것이다.

본 연구에서는 입력된 네이버 현재 상영작 페이지 url로 이동 후, 첫 번째로 상위 10개 영화의 코드번호를 크롤링하여 테이블을 생성한다.

	code
1	136873
2	189053
3	175324
4	183803
5	190244
6	179482
7	188473
8	179159
9	182387
10	188056

<표3-1> 영화 코드

이후 영화의 리뷰 페이지 url이 영화의 코드 번호를 제외한 부분은 동일하다는 규칙을 이용해 코드 번호를 변경한 10개의 영화 리뷰 페이지 url을 테이블로 저장한다.

	url
1	https://movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=136873&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false
2	https://movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=189053&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false
3	https://movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=175324&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false
4	https://movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=183803&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false
5	https://movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=190244&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false
6	https://movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=179482&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false
7	https://movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=188473&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false
8	https://movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=179159&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false
9	https://movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=182387&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false
10	https://movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=188056&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false

<표3-2> 영화 리뷰 페이지 url

마지막으로 생성된 영화 리뷰 url로 이동하여 첫 페이지부터 마지막 페이지까지 순회하며 리뷰를 크롤링한다. 결과적으로 10개의 영화에서 전체 리뷰를 크롤링한 데이터는 총 73,612개 였다. 현재 상영작 1위 영화는 '겨울왕국'으로, 리뷰 수는 14,269개 였고, 10번째 영화는 '결혼이야기'로 71개의 리뷰가 있었다. 또한 현재 사용된 데이터는 2019.12.01. 11시경을 기준으로 상위 10개의 영화로 구성되었다.

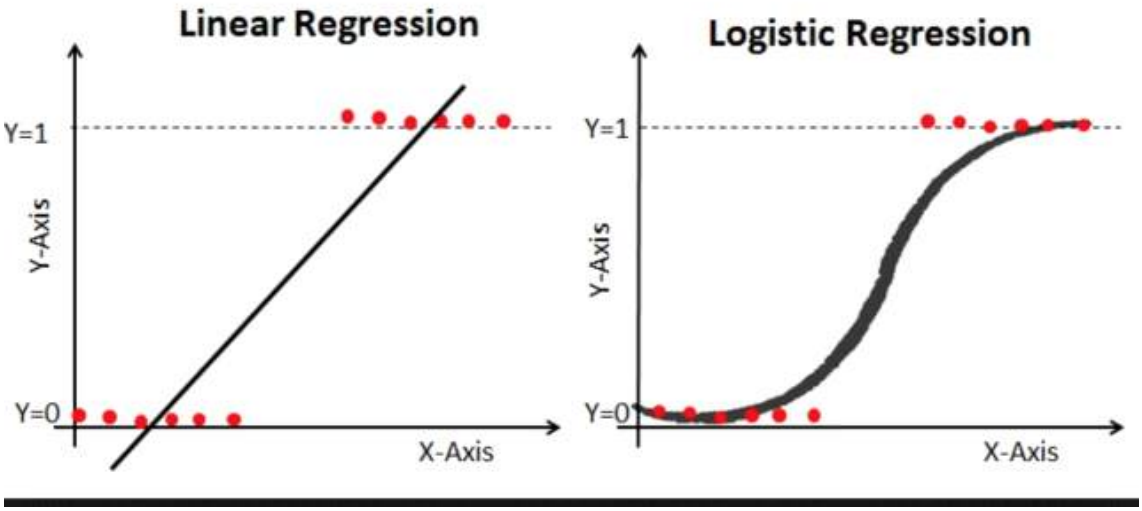
	text	score
1	올라프의 1편요약이 기가맥힙니다	10
2	크리스토퍼 뮤비에서 좀 흡칫함	10
3	애기들 울고 떠들고 하는거 보고 엘사 마법으로 얼릴뻔 했네요	10
4	엘사옷 보고 어머니들 긴장하는 영화	10
5	엘사님 Show yourself 장면 진짜 개오집니다 단연컨데 제2의 렛잇고는 into the unknown 아니고 Show yourself 입니다	10
6	관람객 미래가 보이지 않을 때는 지금 해야할 일을 해야 해	10
7	겨울왕국 역시는 역시였다. OST도 1편만큼이나 중독성있다고 생각함ㅋㅋㅋ1편만큼 존잼임	10
8	엘사 물 속에서 말타고 나올 때 대박ㅋㅋㅋㅋ	10
9	나는 개인적으로 2편이 더 좋았음. 더 깊어진 스토리에 아름다워진 영상미. 한번 더 볼 의향 있음.	10
10	을 생각하는지 알수있는 영화...1편 아동용 영화에서 1편을 보고 자란 성인들까지 보는데 재미를 느낄수 있게 만드는 2편	9

<표3-3> 영화 리뷰와 평점

3.1.2 전처리, 형태소 분석

먼저, 본 연구에서는 Scikit Learn 패키지의 로지스틱회귀분석을 이용하여 모델을 구축했다. Scikitlearn은 무료 소프트웨어 기계 학습 라이브러리 에 대한 파이썬 프로그래밍 언어이다. 다양한 분류 , 회귀 분석 및 클러스터링 포함 알고리즘 서포트 벡터 머신, 랜덤 포레스트, 그래디언트 부스팅, k-means 및 DBSCAN을 하고 파이썬의 NumPy, SciPy라이브러리와 상호 작용하도록 설계되었다.

로지스틱 회귀는 영국의 통계학자인 D. R. Cox가 1958년에 제안한 확률 모델로서 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는데 사용되는 통계 기법이다. 로지스틱 회귀의 목적은 일반적인 회귀 분석의 목표와 동일하게 종속 변수와 독립 변수 간의 관계를 구체적인 함수로 나타내어 향후 예측 모델에 사용하는 것이다. 이는 독립 변수의 선형 결합으로 종속 변수를 설명한다는 관점에서는 선형 회귀 분석과 유사하다. 하지만 로지스틱 회귀는 선형 회귀 분석과는 다르게 종속 변수가 범주형 데이터를 대상으로 하며 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류 (classification) 기법으로도 볼 수 있다.

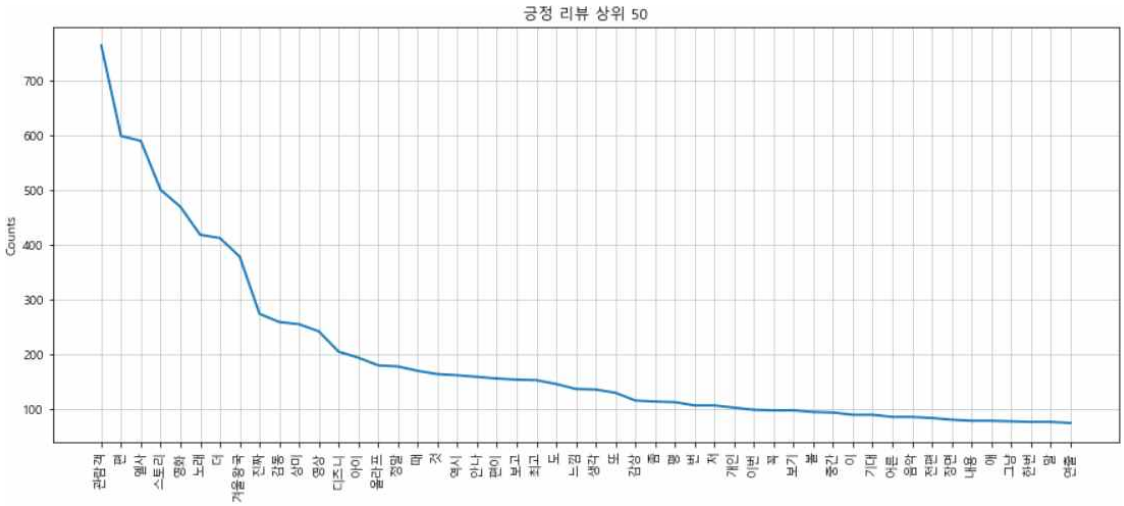


<그림3-1> 선형 회귀 분석과 로지스틱 회귀 분석

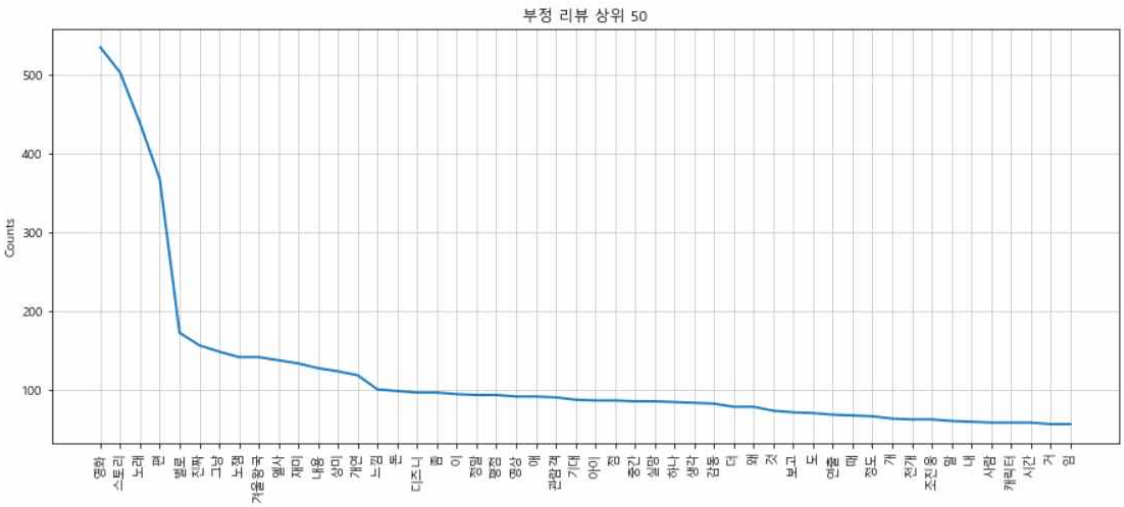
본 연구의 모델을 기계학습을 이용하므로 테스트 데이터와 트레인 데이터가 필요하다. 그래서 73,612개의 리뷰 데이터 중 70%인 51,528개는 학습을 위한 데이터로, 30%인 22,084개는 검증을 위한 데이터로 사용하였다.

형태소 분석은 okt.morphs()를 통해 텍스트를 형태소 단위로 나눴다. 종속 변수를 범주형으로 바꿔야 하기 때문에 별점을 5점을 기준으로 5점 이하는 부정으로 0, 6점 이상은 긍정으로 1을 부여했다. 이때 긍정리뷰는 53,549개였으며, 부정리뷰는 20,063개였다. 여기에서 볼 수 있듯이 상대적으로 긍정리뷰의 개수가 부정리뷰보다 훨씬 많았으며, 이 수치가 결과에 어떤 영향을 줄지에 대해서는 결론 부분에서 확인하도록 한다. 아래 그림은 긍정, 부정 리뷰에 많이 등장하는 단어 50개를 나타낸 것으로 관람객이 영화의 어떤 부분을

긍정, 부정적으로 평가하는지를 알 수 있다. 이렇게 토큰화 된 긍정, 부정 단어를 이용하여 감정 단어를 판별하는 모델을 구축하는데 사용한다. 이때 두 감정에 모두 흔하게 나타나는 관람객, 영화, 스토리, 노래 등은 흔하게 나오는 것으로 간주하여 사후에 가중치를 줄일 필요성이 있다.



<그림3-2> 긍정 리뷰 상위 50개 단어

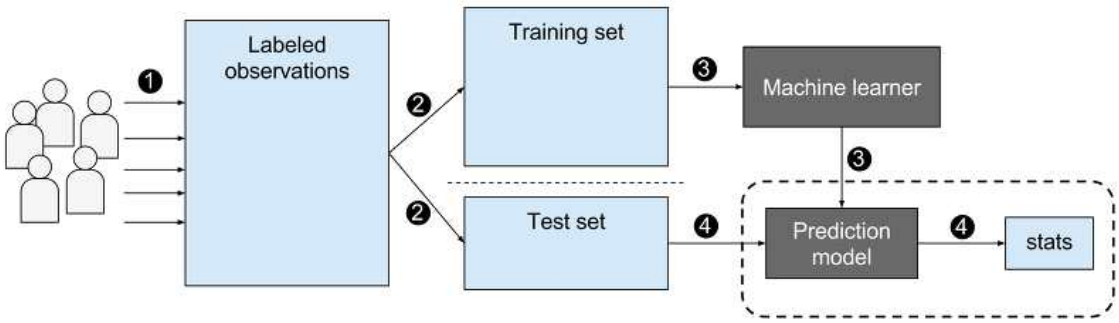


<그림3-3> 부정 리뷰 상위 50개 단어

3.2 긍정, 부정 판별 모델 구축

tf-idf를 이용해 주어진 데이터를 단어 사전으로 만들고 각 단어의 빈도수를 계산한 후 벡터화 하는 객체 생성했다. LogisticRegression을 이용해 문장별 나오는 단어 수를 세서 수치화, 벡터화해서 학습을 시켰다. 70%의 학습 데이터를 통해 긍정 리뷰 안에서 자주

등장하는 단어의 토큰, 즉 벡터를 인식하며 모델을 구축한다.



<그림3-4> 기계 학습

이는 가장 기본이 되고, 가장 구현하기 쉬운 알고리즘인 지도학습으로 일단 컴퓨터에게 문제(Feature)와 정답(Label)이 있는 데이터(Training Set)를 학습 시킨 후, 운영 데이터(Test Set)를 분류하거나 맞추는 것이다.

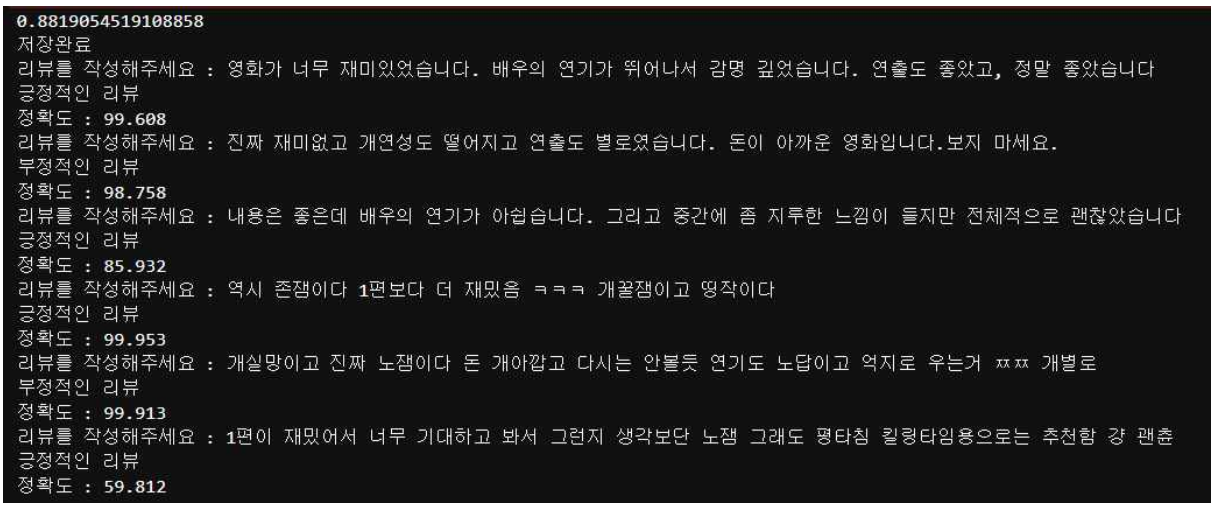
4. 연구 결과

4.1 분석 결과와 정확도

완성된 모델에 대해서 임의로 리뷰를 작성해보았다. 완성형 문장이나 문법상으로 하자가 없는 텍스트에 대한 분석은 기본적으로 진행하되, 본 연구에서 중점적으로 실험해보자 한 것은 오탈자 및 은어, 신조어에 대해서도 분석이 유의미한가를 알아보고자 했다. 이유는 텍스트 분석에 있어 지도 학습을 사용하는 것과 미리 구축된 단어 사전을 사용하는 것 사이에서 어느 것이 우세하다고 단정 지을 수는 없지만 적어도 오타나 신조어에 대해서는 본 연구에서 사용한 방법이 효과적이라는 것을 증명하고 싶었다. 그래서 테스트해 볼 텍스트로는 문법적으로 완벽한 긍정, 부정, 중립 문장과 신조어와 오타를 포함한 긍정, 부정, 중립 텍스트이다. 마지막으로 영화와는 전혀 관련이 없는 문장을 분석하여 향후 발전 방향에 대한 실효성을 검토해보았다.

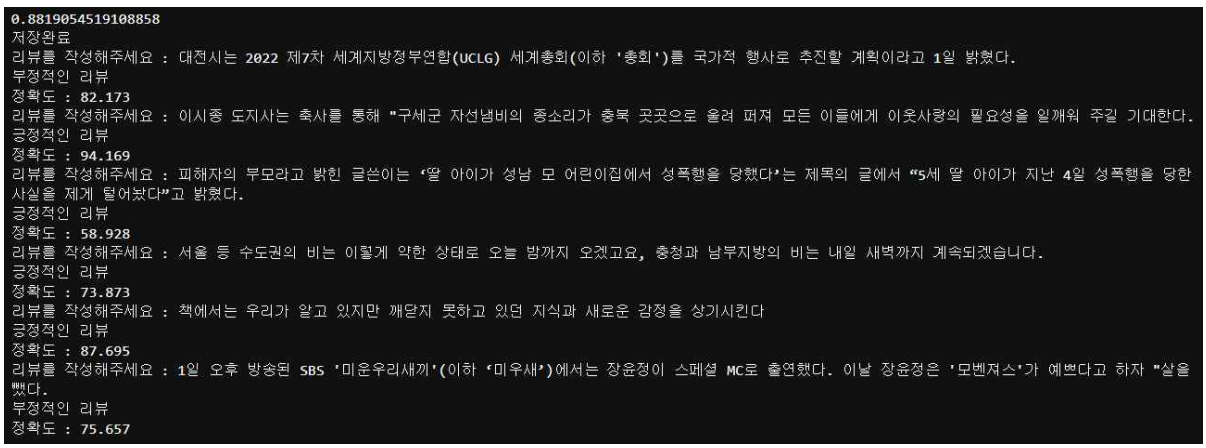
결과적으로 문법적으로 완벽하고 애매한 평가가 아닌 텍스트들은 높은 정확도로 분류가 바르게 되었다. 오타 역시 잘 분류가 되는 모습을 볼 수 있었다. 하지만 부정 리뷰의 경우 긍정으로 판별 되거나, 애매할 경우 긍정으로 판별 되는 경향이 있었다. 이는 앞서 데

이터 셋을 확인했을 때 5점을 기준으로 분류했을 때 부정 리뷰보다는 긍정 리뷰가 2배 이상 많았기 때문으로 볼 수 있다. 데이터의 양이 한쪽 범주로 치우쳐져 있었기 때문에 부정확한 결과가 나온 것으로 보인다. 또한 5점이라는 기준이 1부터 10까지 숫자 중에서 평균이기는 하지만 최빈값 혹은 중앙값을 확인하고 이를 적용한다면 결과의 정확도가 달라질 수 도 있을 것이다.



<그림4-1> 분류 모델 결과1

영화와 전혀 관련이 없는 시사, 사회, 정치, 그리고 연예 기사를 분석기에 입력해봤다. 일부는 어느 정도 맞게 분류된 것이 있었지만 그렇지 않은 것이 많았다. 이는 해당 입력 문장에서 쓰이는 어휘가 학습 데이터에는 전혀 없었기에 판단할 근거가 부족했기 때문이다.



<그림4-2> 분류 모델 결과2

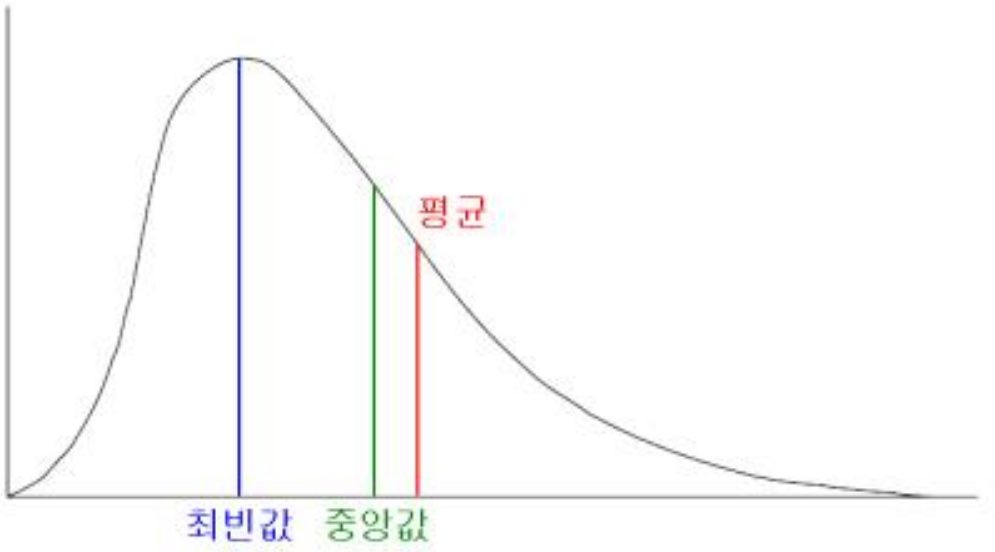
4.2 보완점, 모델의 공정성

애초에 AI나 머신러닝 역시 사람의 손으로 만들어지는 것으로 개인의 주관이 개입될 수 밖에 없다. 따라서 머신러닝 모델이라고 해서 본질적으로 객관적인 것은 아니다. 엔지니어는 학습 사례로 이루어진 데이터 세트를 입력하여 모델을 학습시키며 데이터의 사전준비와 선정에 사람이 관여하기 때문에 모델의 예측이 편향되기 쉽다.

보고 편향은 데이터 세트에 수집된 이벤트, 속성 및 결과의 빈도가 실제 빈도를 정확하게 반영하지 않을 때 나타난다. 이 편향은 사람들이 '말할 필요도 없다고 느끼는' 일반적인 상황은 언급하지 않고 특별히 기억할 만하거나 특이한 상황만을 기록하려는 경향이 있기 때문에 발생한다. 대표적인 예로 바로 본 연구와 비슷한 것이 있다. 일반적으로 영화에 관해 별다른 의견이 없는 사람들은 리뷰를 제출할 가능성이 적기 때문에 학습 데이터 세트의 리뷰 대다수는 극단적인 의견이 된다. 따라서 이 모델은 좀 더 미묘한 어휘를 사용한 영화 리뷰의 감정을 정확히 예측할 가능성이 적다.

표본 선택 편향은 데이터 세트의 사례가 실제 분포를 반영하지 않는 방식으로 선정된 경우 발생한다. 표본 선택 편향은 다음과 같은 여러 형태를 취할 수 있다. 본 연구의 편향은 무응답 편향이 작용했다. 무응답 편향(또는 응답 참여 편향)은 데이터 수집 시 참여도의 격차로 인해 데이터가 대표성을 갖지 못하는 것이다. 영화를 본 이후 크게 감명 받은 관람객은 리뷰를 비교적 많이 작성하고, 영화를 좋게 평가하는 이유를 자세하게 작성한다. 반면, 인터넷 상에서 부정적 평가는 대체적으로 거친 어휘로 작성되었거나, 극단적으로 영화를 부정적으로 평가하지 않는 이상 리뷰를 잘 작성하지 않는다.

따라서 이를 보완하기 위해서는 긍정, 부정을 분류하는 기준을 평균인 5점이 아니라, 최빈값과 중앙값 중에서 분류 정확도가 높은 것으로 설정할 필요가 있다. 또한 데이터 추출에 있어서 긍정, 부정 리뷰의 데이터 개수를 비슷한 수준으로 조정할 필요가 있으며 최대한 편향을 줄일 수 있는 공정한 데이터를 선택하는 것이 정확도를 높이는데 도움을 줄 수 있을 것이라고 생각된다.



<그림4-3> 최빈값, 중앙값, 평균

4.2 발전 방향과 의의

본 연구에서 사용한 영화 리뷰는 별점 스코어라는 자료를 범주화 할 수 있는 기준이 존재했기 때문에 감정을 분류할 수 있었다. 하지만 SNS 및 대부분의 인터넷 사이트에서 존재하는 텍스트는 이를 감정적으로 분류할 수치가 부여되어 있지 않다. 그렇기 때문에 지도학습 자체가 불가능하며, 일반적으로 사전에 구축된 단어사전을 이용한다. 예를 들어 '기쁨'이라는 단어는 감정 단어 사전에 의하면 긍정으로 분류되어 있을 것이다. 그러므로 텍스트에 '기쁨'이라는 단어가 포함된다면 텍스트는 긍정이라고 분류하는 것이다. 이것은 물론 직관적인 분류 방법이기는 하다. 하지만 텍스트 원문이 "생명 잉태의 기쁨이 없는 세상"이었다면 '없는'을 제외하고는 대부분 긍정 단어로 보기 쉽기 때문에 위 텍스트는 긍정으로 분류될 것이다. 하지만 그렇지 않다. 뿐만 아니라 기본적으로 한국어 단어 감정 사전이 잘 정의되어있지도 않다. 이와 같은 오류와 앞서 말했던 인터넷의 특성상 오타가 많은 텍스트를 분류하기 위해서는 단어 사전 이용이 그리 좋지 못하다고 생각된다. 그렇기 때문에 영화 리뷰와 같이 텍스트를 분류할 수 있는 기준이 부여된 자료를 이용해 모델을 학습시키면, 향후 영화 리뷰뿐만 아니라 다른 텍스트의 감정을 분석할 때에도 이용할 수 있을 것으로 기대된다. 예를 들어 상품의 리뷰, 도서에 대한 평가 등 좀 더 다양한 분야와 범위의 데이터를 수집하면 대부분의 텍스트를 감정 분석할 수 있을 것이다. 따라서 본 연구에서는 영화 데이터만을 사용했지만 향후 연구에서 다른 분야의 데이터를 추

가로 더하여 모델을 학습시킨 후, 사회적인 이슈 혹은 감정 분석을 원하는 주제에 대한 리뷰만을 추출하여 이를 분석한다면, 설문조사나 통계 전수 조사 없이 국민의 생각을 보다 솔직하고 빠르게 얻을 수 있을 것이라고 생각한다. 이것이 해당 연구의 발전 방향이며 사회적, 경제적으로의 범용적인 확장성이라는 의의이다.

참 고 문 헌

[1] 김건영, 이창기.(2016).Convolutional Neural Network를 이용한 한국어 영화평 감성 분석. 한국컴퓨터종합학술대회 논문집,747-749.

[2] 김유영, 송민.(2016).영화 리뷰 감성분석을 위한 텍스트 마이닝 기반 감성 분류기 구축. J Intell Inform Syst 2016 September,71~89.

[3] 이오준, 박승보, 정다울, 유은순.(2014).소셜 빅데이터를 이용한 영화 흥행 요인 분석. 한국콘텐츠학회논문지

[4] 박은정, 조성준.(2014).KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지. 제26회 한글 및 한국어 정보처리 학술대회 논문집

[5] google. (연도미상). 머신러닝 단기집중과정 중 공정성: 편향의 유형. https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias?hl=ko

[6]PSYCHY.(2014). 기본 감정(big six). https://m.blog.naver.com/PostView.nhn?blogId=yars&logNo=130184782526&proxyReferer=https%3A%2F%2Fwww.google.com%2F

[7]wikipedia.(2019).txet mining. https://en.wikipedia.org/wiki/Text_mining

[7]wikipedia.(2019).자연어 처리. https://ko.wikipedia.org/wiki/%EC%9E%90%EC%97%B0%EC%96%B4_%EC%B2%98%EB%A6%AC

[7]wikipedia.(2019).TF IDF. https://ko.wikipedia.org/wiki/Tf-idf

[7]wikipedia.(2019).txet mining. https://en.wikipedia.org/wiki/Text_mining

[8]KoNLPy.(2015).파이썬 한국어. NLPhttps://konlpy-ko.readthedocs.io/ko/v0.4.3/

[9]wikipedia.(2019).BeautifulSoup.https://en.wikipedia.org/wiki/Beautiful_Soup_(HTML_parser)

[10]wikipedia.(2019).Scikit learn.https://en.wikipedia.org/wiki/Scikit-learn

[11]wikipedia.(2019).로지스틱 회귀.https://ko.wikipedia.org/wiki/%EB%A1%9C%EC%A7%80%EC%8A%A4%ED%8B%B1_%ED%9A%8C%EA%B7%80

ABSTRACT

Text Emotion Analysis Using Natural Language Processing

PARK, JI-SU

Department of Big Data

Graduate School of Chungbuk University

Currently, emotional analysis mainly uses a method of preparing an emotional dictionary to match what emotion the word represents. However, it is difficult to include a new language, new words, or slang and requires a cumbersome task of classifying all Hangul words as emotional words. Therefore, this study aims to make more accurate judgment model by collecting data that categorizes emotions of texts such as movie review.

The word was vectorized with tf-idf and weighted. We used LogisticRegression to train our models with positive and negative reviews.

As a result, they showed high accuracy for typos and new words. In the case of ambiguous sentences, however, they tend to be biased into positive reviews.

In this study, only the film data is used, but in future studies, if the model is trained by adding data from other fields, the public's feelings on political and social issues can be analyzed. This can help to reduce national costs for social and public opinion polls.