품사에 따른 영화 리뷰 감성분석 연구

정세민*, 이세영*, 안유나*, 김보경*

A Study on the Effect of the Part of Speech on Movie Review Sentiment Classification Performance

Jung Semin*, Lee, Seyoung*, Ahn Yuna*, and Kim Bo Gyeong*

요 약

다양한 매체와 플랫폼의 등장으로 영화관에 가지 않더라도 언제, 어디서나 쉽게 영화를 접할 수 있다. 따라서 영화 리뷰와 평점은 영화 선택에 큰 영향을 미친다. 본 연구는 네이버 영화 포털에서 수집한 약 6만 건의 영화 리뷰 데이터로 감성 분석을 수행하였다. 먼저 평점을 기준으로 긍정과 부정 감성 라벨을 생성하였다. 리뷰 텍스트는 한국어 형태소 분석기를 통해 문장 형태소 분석을 한 후 문서단어행렬을 생성하였다. 데이터 셋은 명사, 형용사, 동사, 부사 등의 다양한 조합을 기반으로 생성하였으며, 각 데이터 셋에 머신 러닝 알고리즘을 적용하여 분류 모델을 생성하여 성과를 비교하였다. 단일 품사 데이터 셋으로 감성 예측 모델을 구축한 경우, 명사, 형용사, 동사, 부사 등의 순서로 높은 정확도를 보이는 것으로 나타났고, 두 개의 품사로 구성된 데이터 셋의 경우에는 명사와 형용사로 구성된 모델이 가장 좋은 성과를 나타냈다. 마지막으로, 세 개의 품사로 구성된 경우 명사, 형용사, 동사로 구성된 모델이 가장 좋은 성과를 나타냈다. 본 연구는 영화 리뷰의 감성 예측에 있어 서로 품사를 기반으로 한 데이터 셋이 모델의 성과에 영향을 미칠 수 있다는 사실을 검증하였다는 측면에서 의미가 있다.

Abstract

With the advent of various media and platforms, people can easily access movies anytime, anywhere without going to the movie theater. Therefore, movie reviews and ratings have a great influence on movie choices. This study conducted sentimental analysis with about 60,000 movie review data collected from the Naver movie portal. First, positive and negative sentimental labels were created based on the rating. For the review text, a document word matrix was created after sentence morpheme analysis was performed through a Korean morpheme analyzer. The dataset was generated based on various combinations of nouns, adjectives, verbs, and adverbs, and the performances were compared by creating classification models by applying machine learning algorithms to each dataset. When the sentimental prediction model was constructed with datasets of a single parts of speech, it was found that the accuracy was high in the order of nouns, adjectives, verbs, and adverbs, and in the case of datasets composed of two parts of speech, the model composed of nouns and adjectives showed the best performance. Finally, when composed of three basic parts of speech, the model composed of nouns, adjectives, and verbs showed the best performance. This study is meaningful in that it verified the fact that data sets based on different parts of speech can affect the performance of the model in predicting the sentiment of the movie reviews.

Key words

Text mining, Sentimental Analysis, Logistic Regression, Bernoulli Navie Bayes, XGBoost, Movie review

* 계명대학교, hiemily@naver.com, adltpdud@naver.com, sgvina@naver.com, jkn3323@naver.com

Ⅰ. 서 론

영화에 대한 리뷰는 관람객이 영화를 선택하는 데에 큰 영향을 미치는 요소 중 하나이다. 리뷰에 내포된 감성에 따라 영화 선택 여부가 달라질 수 있기 때문에 관람객이 작성한 리뷰에 대해 정확한 감성분석이 이루어져야 한다.[1] 리뷰 감성 분석 모델을 구축하는 과정에서 리뷰 데이터 토큰화의 품사 및 품사 조합에 따라 다른 성능을 보일 것이다. 따라서 본 논문에서는 영화 리뷰 데이터를 여러 가지 품사 조합별로 토큰화하여 데이터 셋을 생성한 감성 분류 모형을 구축하였다. 이를 통해 품사가 감성 분석 모델의 성능에 미치는 영향을 살펴보고자 한다.

Ⅱ. 본 론

2.1 연구 방법

본 연구는 <그림 1>과 같이 진행되었다. 우선 네이버 영화 포털에서 2021년 7월 19일부터 2021년 9월 7일까지 약 두 달 동안 데이터를 수집하였다. 수집된 데이터에서 영화 리뷰 결측치와 중복 리뷰를 삭제, ‘sentiment’ 변수를 만들어 영화 DB를 구축하였다. ‘sentiment’은 1~3점에 해당하는 영화 평점을 부정 리뷰, 8~10점에 해당하는 평점을 긍정 리뷰로 분류하여 생성하였다.

그 후 한국어 자연어 처리 패키지인 Konlpy를 사용하여 자연어 전처리를 하였다. 각 문장에서 형태소 분리, 토큰화, 불용어 제거, 품사 태깅 작업을 진행하여 형태소 DB를 구축하였다. 형태소 DB에서 태깅된 품사 중‘명사’,‘동사’,‘형용사’,‘ 부사’4가지 품사를 사용하여 15개의 조합과 영화 DB의 ‘sentiment'을 사용하여 분석 데이터를 생성하고 영화 리뷰 감성 분석을 진행하였다.

그 후 ' Permutation Importance'를 사용하여 분류에 가장 큰 영향을 미치는 요인을 추출하여 결과를 해석하였다.

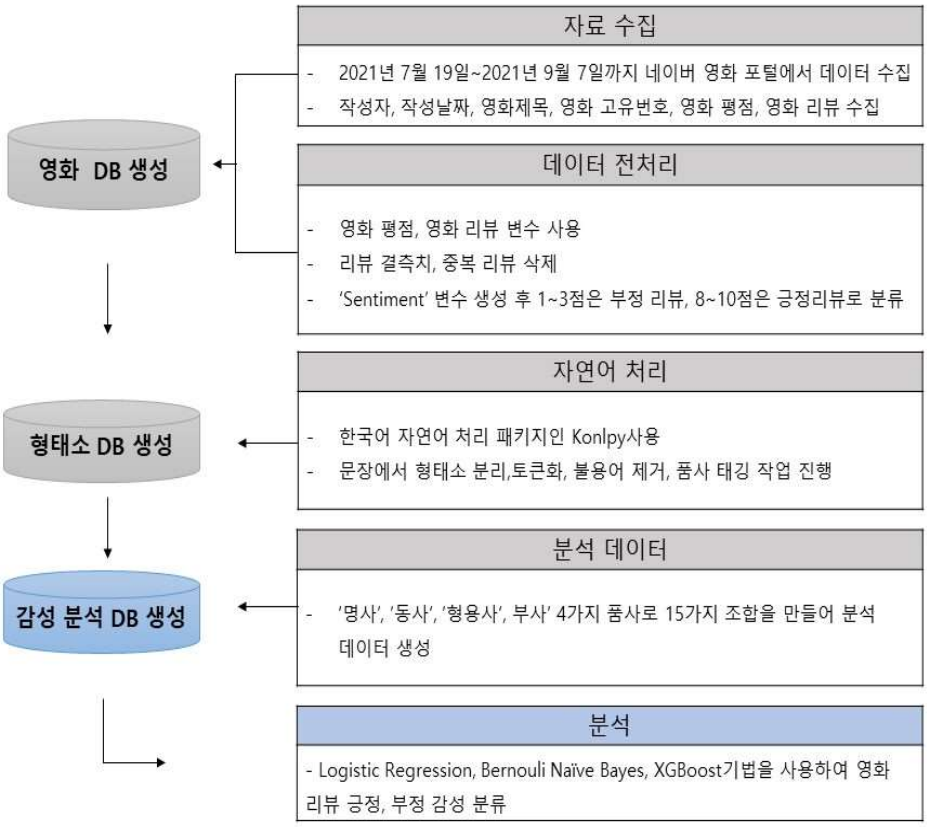

그림 1. 리뷰 감성 분류의 전체적인 과정
Fig. 1. Entire process of review sentimental classification

2.2 토큰화 형태소 분석

한국어 자연어 처리 패키지인 Konlpy를 사용하였다. 형태소 분석기인 Okt를 사용하여 문장을 정규화하고 각 단어의 어간을 추출하여 형태소를 분리, 토큰화를 진행하였다. 조사, 관사 등의 불용어를 제거하고 명사, 동사, 형용사, 부사 품사 태깅 작업을 진행하였다. 태깅 작업을 통해 정제된 토큰화로 Counter-Vectorize를 진행하였다.

Counter-Vectorize는 빈도를 기준으로 특징을 추출해 벡터화하는 방법으로 본 연구에서는 빈도수가 가장 높은 순으로 해당 개수만큼 특징을 추출하는 'max_features' 파라미터를 5,000으로 설정해 Counter-Vectorize를 진행하여 문서 단어 행렬을 생성하였다.

2.3 분석방법

평점에 따라 리뷰를 긍정, 부정으로 분류하여 감성 분류를 진행하기 위해서는 이진 클래스 분류 기법이 필요하다. 따라서 본 연구에서는 Logistic Regression(이하 LR), Bernoulli Naive Bayes(이하 BNB), XGBoost(이하 XGB) 기법으로 진행하였다.

LR는 종속변수가 범주형일 경우 하나의 종속변수와 한 개 이상의 독립 변수 사이의 인과관계 설명에 궁극적 목표를 둔 기법으로 다른 회귀 분석 기법에 비해 독립변수와 이진형 종속변수 간의 관계를 매우 유연하게 사용할 수 있다. 특정 이벤트가 발생할지 여부를 직접 예측하는 것이 아니라 해당 이벤트가 발생할 확률을 추정하는 분석 기법의 일종이다. 종속 변수의 예측 값은 항상 0에서 1사이의 확률 값을 갖게 되며 값이 0.5보다 크면 이벤트가 발생할 확률이 높고 0.5보다 작으면 이벤트가 발생할 확률이 낮다고 예측할 수 있다.[2,4]

BNB는 목표변수가 이진형인 경우 각각의 범주들이 정해져 있을 때 범주의 출현 횟수를 조건부 독립으로 가정하여 확률적으로 분류하는 확률 기반 분류 기법이다. BNB는 간결하고 효율적이며, 노이즈와 누락 데이터를 잘 처리하여 텍스트 분류에서 널리 사용되고 있다.[3]

XGB는 결측치를 자동 처리 하여 병렬적으로 트리를 생성하는 기법이다. XGB는 회귀와 분류 등 다양한 문제를 모두 지원하고 큰 규모 데이터 예측의 안정성과 훈련 속도가 높다. 성능과 자원 효율이 좋아 널리 사용되는 기법 중 하나이다.[5]

Train, Test 데이터를 8대 2 비율로 나눠 위 3개 모델로 모델링을 수행하였다. LR는 최적화에 사용되는 알고리즘을 결정하는 solver가 수렴하게 만드는 최대 반복 횟수를 200으로 파라미터를 설정 한 후 4품사 15개 조합을 모델링 하였다. BNB와 XGB는 기본 파라미터로 4품사 15개 조합을 모델링을 하였다. 모델 성능은 [표1]과 같다.

단일 품사를 사용할 때, 알고리즘 중에서는 LR이 가장 성능이 좋았다. 품사 중에서는 명사로 구성된 데이터 셋을 사용했을 때가 다른 품사를 사용했을 때보다 더 나은 성과를 보였다. 즉, 영화 리뷰의 감성 예측에 있어 명사가 좀 더 중요한 역할을 한다는 의미다. 두 개의 품사로 구성된 명사와의 조합으로 된 경우가 다른 조합보다 더 나은 성과를 보여준다. 특히, 명사와 형용사로 구축된 데이터 셋을 사용하는 경우 가장 좋은 성과를 보인 것을 알 수 있다. 마지막으로 세 개 품사 조합의 경우 명사, 동사, 형용사로 구성된 데이터 셋을 사용하는 경우가 가장 좋은 성과를 보이는 것을 알 수 있다.

표 1. 모델 성능 결과
Table 1. Model performance result

	LR	BNB	XGB
명사	0.8674	0.8595	0.8245
동사	0.8070	0.8046	0.7993
형용사	0.8341	0.8328	0.8201
부사	0.7940	0.7940	0.7939
명사+동사	0.8712	0.8619	0.8379
명사+형용사	0.8915	0.8791	0.8517
명사+부사	0.8685	0.8609	0.8353
동사+형용사	0.8481	0.8439	0.8241
동사+부사	0.8101	0.8069	0.7995
형용사+부사	0.8363	0.8336	0.8200
명사+동사+형용사	0.9004	0.8868	0.8475
명사+동사+부사	0.8741	0.8675	0.8265
명사+형용사+부사	0.8948	0.8871	0.8772
동사+형용사+부사	0.8491	0.8842	0.8249
명사+동사+형용사+부사	0.8985	0.8885	0.8501

2.4 변수 중요도 측정

영화 리뷰 감성 분류 모델 결과 성능이 가장 높은 4품사 4개 조합에서 어떠한 단어와 품사가 모델 성능에 많은 영향을 끼쳤는지 알아보기 위해 'Permutation Importance'를 사용하였다. 세 개 모델 중 평균적으로 성능이 좋은 LR를 가지고 변수 중요도를 측정하였다. 상위 6개의 중요 변수를 측정하였고 결과는 아래 [표2], [표3]과 같다. 각 품사 조합에서 상위 6개의 변수 중요도 결과에서 같은 단어와 품사들이 추출되었고 명사보다는 동사와 형용사가 모델 결과에 결정적 영향을 미치는 것을 확인할 수 있다. 따라서 명사보다는 동사와 형용사가 문장에 주 영향을 미치는 것으로 판단된다.

표 2. 변수 중요도 측정결과(1)
Table 2. Feature importance

명사 + 동사 + 형용사		명사+ 동사+형용사+부사	
Weight	Feature	Weight	Feature
0.0112±0.0014	재밌다	0.0098±0.0026	재밌다
0.0087±0.0007	최악	0.0086±0.0010	최악
0.0072±0.0008	재미없다	0.0061±0.0005	노잼
0.0068±0.0012	노잼	0.0060±0.0010	재미없다
0.0066±0.0016	아깝다	0.0058±0.0013	아깝다
0.0054±0.0014	좋다	0.0055±0.0024	좋다

표 3. 변수 중요도 측정결과(2)
Table 3. Feature importance

명사 + 형용사 + 부사		명사 + 형용사	
Weight	Feature	Weight	Feature
0.0106±0.0029	재밌다	0.0107±0.0023	재밌다
0.0094±0.0018	최악	0.0091±0.0009	최악
0.0076±0.0006	아깝다	0.0079±0.0015	아깝다
0.0068±0.0010	노잼	0.0070±0.0015	노잼
0.0044±0.0012	좋다	0.0051±0.0008	좋다
0.0043±0.0012	재미없다	0.0049±0.0009	쓰레기

III. 결 론

본 연구에서는 품사가 영화 리뷰 감성분석에 유의미한 영향을 미치는 것을 확인하였다. 한국어 기본 문장인 '주어+서술어' 구조에 해당하는 '명사+형용사+동사'가 감성분석의 성능에 크게 기여하는 것을 확인하였다. 그러나 본 연구에서 사용한 형태소 분석기인 Okt가 '진짜', '정말' 등 명사와 부사 역할을 하는 단어일 경우, 한 품사로만 분류하는 한계가 있다. 또한 '차라리', '역시' 등을 부사가 아닌 명사로 인식하는 등 제대로 작동하지 않는 한계가 존재한다. 향후 유사 연구에서는 Okt 형태소 분석기가 아닌 다른 분석기를 사용함으로써 본 연구의 한계점을 보완하며 더욱 완성도 높은 영화 감성 분석 연구로 이어질 것이라 기대한다.

참 고 문 헌

[1] 김진현, 김정현, 정성욱 and 강신재. (2017). SNS 게시글과 감성분류에 기반한 다단계 노래 추천 시스템. 예술인문사회 융합 멀티미디어 논문지, 7(3), 283-290.

[2] 이장택, 조현식 "로지스틱 회귀모형을 이용한 프로야구 홈경기의 이점에 관한 연구" Journal of The Korean Data Analysis Society 11.1 pp.533-543 (2009) : 533.

[3] 조희련, 임현열, 차준우, 이유미. (2021). KoBERT, 나이브 베이즈, 로지스틱 회귀의 한국어 쓰기 답안지 점수 구간 예측 성능 비교. 한국정보처리학회 학술대회논문집, 28(1), 501-504.

[4] 진수봉, 이종우 (2017). 로지스틱회귀분석 모델을 활용한 도시철도 사상사고 사고예측모형 개발에 대한 연구. 한국철도학회 논문집, 20(4), 482-49

[5] 하지은, 신현철, 이준기. (2017). RandomForest와 XGBoost를 활용한 한국어 텍스트 분류: 서울특별시 응답소 민원 데이터를 중심으로. 한국빅데이터학회지, 2(2), 95-104.