2018 – 가을학기 온라인데이터 분석

**Final Project Paper**

**영화 댓글 감정분석에서 영화 장르가 미치는 영향**

4조 : 2018321084 안주영

2017321060 박가람

Code&Data : <https://www.dropbox.com/sh/x47zr8eq3vc4oxd/AABWhuast4QkZ5DI-n-tHbqsa?dl=0>

**1. 서론**

감성분석에는 크게 감성어사전을 이용한 분석과 머신러닝 알고리즘을 사용한 분석 두가지 분석이 사용된다. 감성어사전을 만드는 과정은 단순한 데이터분석 뿐만 아니라 언어학적인 접근이 필요하기 때문에 머신러닝 기반의 감성분석에 대해서 연구하고자 한다.

특히 머신러닝기반의 감성분석에 있어서 가장 많이 사용되는 영화평점 데이터를 활용한 분석에서 영화의 장르별로 모델을 만드는 것과 장르를 구분하지 않고 모든 데이터를 사용하여 모델을 만드는 것 두가지를 비교해보고자 한다.

머신러닝에 있어서 보통 데이터의 크기가 커질수록 높은 모델의 정확도를 갖는 것으로 알려져 있다. 영화 감상평의 감성분석을 하는데 있어서 장르별 특수성이 존재한다면 영화의 장르별로 각각 모델링하는 것이 더 좋은 예측 정확도를 보여줄 것이다.

연구에 사용한 데이터는 2018년 11월 28일 19시 30분을 기준으로 ‘네이버 영화’사이트에 등록 되어있는 2018년도 개봉한 모든 영화의 한줄평 데이터를 사용하였다.

**2. 이론적 논의와 선행연구 조사**

각 장르별로 댓글이 장르의 특수성을 반영한다면 각 장르별 구분하여 모델링한 감성분석 모델이 좋은 예측 정확도를 보여줄 것이고, 그렇지 않다면 일반적으로 많은 데이터를 사용하여 모델링한 전체데이터 모델이 더 높은 예측 정확도를 보여줄 것이다.

2016년도에 이미 IMDb의 영화평 데이터를 사용하여, 장르별로 영화의 감성어사전을 생성하여 영화평점에 대한 분석이 이루어진 바 있다[1]. ‘영어’ 데이터를 사용하여 감성어 분석을 ‘감성어 사전’기반으로 이루어진 선행 연구와는 다르게, 이번 연구에서는 ‘한글’ 데이터를 활용한 ‘머신러닝 기반’의 분석을 통해서 장르별로 댓글이 장르의 특수성을 반영하는지 알아보도록 하겠다.

**3. 방법론**

1) 데이터

2018년 11월 28일 19시 30분을 기준으로 ‘네이버 영화’사이트에 등록 되어있는 2018년도 개봉한 모든 영화를 대상으로 데이터를 수집하였다. 댓글내용과 평점으로 이루어진 데이터를 각 영화별로 수집하였다. 또한 장르별 분석을 위하여 네이버에서 관리하는 영화의 고유 코드와 영화의 장르를 함께 수집하였다.

2018년 개봉한 영화 중 영화평이 존재하는 737개영화에대한 데이터를 수집하였고, 총773,123개의 평점과 영화평 데이터 관측치를 수집하였다.

2) 수집 방법

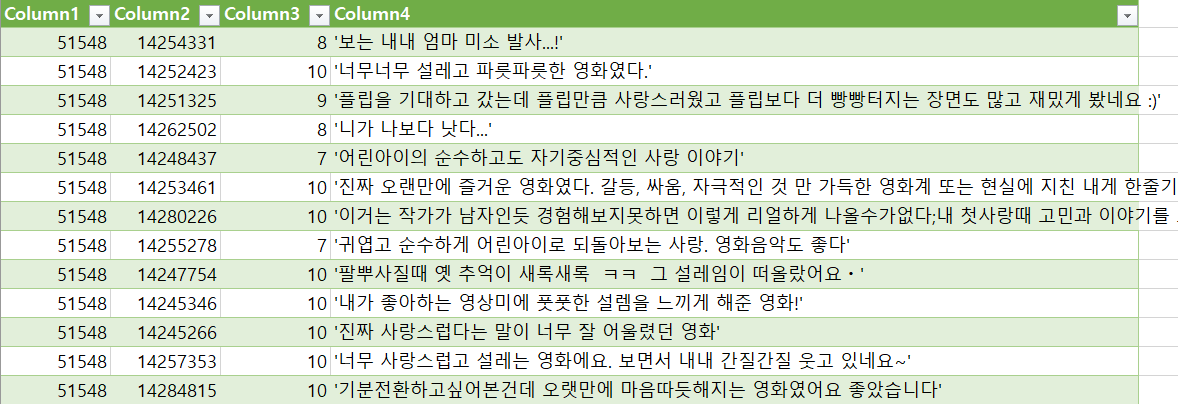
우선 네이버영화의 고유 영화코드를 기반으로 해당 영화의 페이지에 접근할 수 있기 때문에 각 영화의 코드를 input으로 받아서 영화평점을 수집하는 함수를 생성하였다.

두번째로는 처음 만들어진 크롤링 함수에 영화의 고유코드를 입력해 주기 위해서 2018년도에 개봉한 영화의 고유코드를 수집하는 함수를 생성하였다. 또한 장르별 분석을 위해서 해당 영화의 장르도 함께 수집하였다.



[그림 1] 영화고유번호와 장르가 저장된 *MvDf.csv*

Request 모듈의 경우 로컬환경에서 수집이 제대로 되지 않는 버그가 존재하여. Urllib 모듈의 request함수를 이용하여 네이버영화에 등록된 2018년 개봉작들의 고유번호와 영화의 장르를 함께 수집하였다. 앞에서 수집한 영화의 고유번호를 기반으로 각 영화별 영화평과 평점을 수집하여서 영화별로 저장하였다.



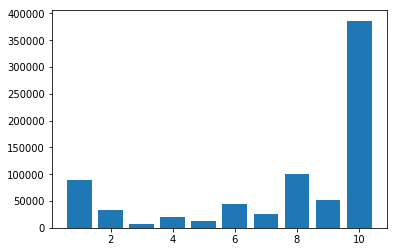
[그림 2] 각 영화별로 영화평과 평점이 저장된 파일 예시 *51548.tsv*

분석을 위하여 수집된 영화평에서 간단한 전처리를 통해 특수문자등을 제거하고 먼저 모든 영화의 평점과 댓글이 담긴 테이블을 생성하였고, 두번째로는 장르별 분석을 위해서 장르별로 영화평과 평점을 합쳐서 총 21개의 장르별로 분석 테이블을 생성하였다.

3) 분석 방법

모델을 만들고 평가하기위해서 데이터를 전체 데이터의 70%를 train데이터로 30%를 train데이터로 사용하였다.

모델의 간결성을 위해서 1~10까지 주어진 평점데이터를 긍정과 부정 두가지로 나누어서 평가하였다. 평점의 분포를 확인해 보았을 때 관측치의 개수가 2점에서 3점으로 넘어가는 구간에서 확연하게 감소하였고 7점에서 8점으로 넘어가는 구간에서 확연하게 증가하는 추세가 확인되었다. 이를 통해 2점 이하인 경우를 부정적인 감성으로 8점 이상인 경우를 긍정적인 감성으로 라벨링하여서 사용하였다.

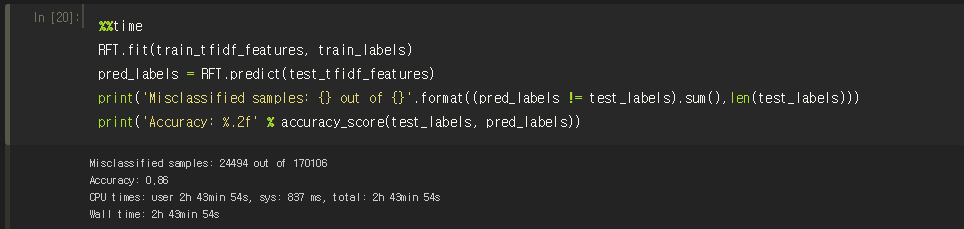


[그림 3] 영화 평점 별 영화평 빈도 수

분석에 사용한 word matrix는 단어의 빈도수를 계산해면서 모든 문서에 공통적으로 등장하는 단에어는 penalty를 부여하는 tf-idf matrix를 사용하였다.

머신러닝 알고리즘의 경우는 KNN, Random Forest등 다양한 알고리즘을 사용해 보았지만 데이터의 크기가 크기 때문에 로컬 환경에서 작업에 어려움이 있었다.

Random Forest의 경우 약 2시간 40분의 시간의 소요되었고 예측 정확도의 경우는 약 86%의 결과를 보여주었다. 이 때의 예측 정확도가 L2페널티를 이용한 Lasso logistic regression과 큰 차이를 보여주지 못하였다. Grid Search를 통한 하이퍼파라미터 튜닝을 통해 좀 더 높은 예측 정확도를 얻을 수 있겠지만, 컴퓨팅 파워의 제한으로 인해 Grid Search를 시행는데 어려움이 있었다. 따라서 큰 computing resource가 필요한 알고리즘을 제외하였다.



[그림 4] 전체장르 Random Forest 알고리즘 사용 결과

최종적으로 분석에 사용한 알고리즘은 L2 penalty를 사용한 Lasso logistic regression, Naïve Bayes, Stochastic Gradient Descent에서 hinge loss를 사용한 Support Vector Machine세가지를 사용하였다.

분석 결과에서 test 데이터의 크기가 50개 이하인 경우는 데이터의 크기가 너무 작아서 정확한 결과를 얻기 어려울 것이라고 판단하여서 영화평의 개수가 50개 이하인 장르와 영화장르 정보가 없는 영화들은 제거하였다.

최종적으로 각 전체와 각 장르별로 각각 3가지 알고리즘을 사용하였을 때의 예측 정확도를 얻을 수 있었다. 이를 바탕으로 최고의 예측정확도를 비교하는 모비율 검정을 통해 전체영화의 예측정확도와 유의미한 차이를 갖는지 R을 활용하여서 모비율 검정을 수행하였다.

**4. 결과**

[표 1] 예측 정확도 및 모비율검정 결과

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Genre** | **lr\_acc** | **nb\_acc** | **svm** | **max** | **p-value** | **test\_n\_row** | **total\_n\_row** |
| **전체** | 0.874 | 0.863 | 0.846 | 0.874 |  | 198508 | 661691 |
| **장르별 예측 정확도가 전체 예측 정확도보다 높은 장르** | | | | | |  |  |
| **뮤지컬** | 0.955 | 0.955 | 0.959 | 0.959 | 0 | 2074 | 6911 |
| **가족** | 0.904 | 0.904 | 0.947 | 0.947 | 0.049 | 94 | 313 |
| **애니메이션** | 0.934 | 0.935 | 0.940 | 0.940 | 0 | 14024 | 46745 |
| **다큐멘터리** | 0.897 | 0.899 | 0.921 | 0.921 | 0 | 4328 | 14424 |
| **멜로/애정/로맨스** | 0.879 | 0.882 | 0.914 | 0.914 | 0 | 9640 | 32133 |
| **서스펜스** | 0.864 | 0.864 | 0.896 | 0.896 | 0.029 | 1109 | 3696 |
| **판타지** | 0.839 | 0.848 | 0.892 | 0.892 | 0 | 10762 | 35871 |
| **액션** | 0.888 | 0.880 | 0.885 | 0.888 | 0 | 59988 | 199960 |
| **전체 예측 정확도가 장르별 예측 정확도보다 높은 장르** | | | | | |  |  |
| **모험** | 0.748 | 0.810 | 0.871 | 0.871 | 0.62 | 4387 | 14622 |
| **미스터리** | 0.809 | 0.825 | 0.870 | 0.870 | 0.293 | 9471 | 31568 |
| **드라마** | 0.859 | 0.863 | 0.866 | 0.866 | 0 | 40763 | 135875 |
| **범죄** | 0.786 | 0.824 | 0.865 | 0.865 | 0.013 | 9702 | 32340 |
| **스릴러** | 0.669 | 0.846 | 0.841 | 0.846 | 0 | 5922 | 19740 |
| **코미디** | 0.742 | 0.816 | 0.843 | 0.843 | 0 | 10494 | 34979 |
| **SF** | 0.640 | 0.837 | 0.829 | 0.837 | 0 | 5229 | 17430 |
| **공포** | 0.722 | 0.791 | 0.821 | 0.821 | 0 | 8211 | 27370 |

총 16개의 장르 중 모험과 미스터리 장르를 제외한 16개의 장르가 전체의 예측정확도와 유의미한 차이를 갖는 것으로 나타났다. 8개의 장르가 전체의 예측정확도보다 높은 장르별 예측 정확도를 보여줌으로써 장르의 특수성이 존재하는 것을 추론할 수 있었고, 6개의 장르가 전체의 예측정확도보다 낮은 정확도를 보이는 것을 통해 장르의 특수성이 존재하지 않는다고 추론해 볼 수 있었다.

또한 가장 높은 예측정확도를 갖는 알고리즘의 차이를 볼 수 있었다. 전체의 예측정확도는 로지스틱 회귀분석에서 가장 높은 값을 갖는 반면 장르별 예측 정확도의 결과는 대부분 SVM에서 가장 높은 값을 갖는 것을 확인 할 수 있었다.

**5. 논의**

우선 전체모델의 경우 로지스틱 회귀분석이 가장 좋은 예측정확도를 보이는데 반해 장르별 모델에서 hinge loss 기반의 Stochastic Gradient Descent가 가장 좋은 결과를 보여주는 것은, 데이터의 크기로 해석할 수 있다. 회귀모형의 경우 변수의 개수가 많을수록 높은 정확도를 보여주는데 데이터 크기가 크다는 것은 그만큼 tf-idf매트릭스의 크기 즉 x변수의 개수가 많다는 것을 의미한다. 이는 장르별 감성분석에서 가장 많은 관측치를 갖는 액션영화에서도 로지스틱 회귀분석이 가장 높은 예측정확도를 갖는 것을 통해서 확인할 수 있다.

두번째로 전체모형에 비해 높은 예측정확도를 갖는 8개의 장르의 경우 각 장르별 특수성을 나타내는 분야라고 생각 할 수 있다. 각 영화별로 댓글에서 나타나는 단어들의 의미가 보편적으로 사용되는 의미와 다르게 사용될 수 있다는 것을 추론해 볼 수 있다. 실제로 선행연구에서 제시한 ‘scary’의 경우 공포영화에서는 긍정적이지만 다른 장르에서는 부정적으로 사용되어지는데 한국어에서도 마찬가지로’공포’라는 단어가 동일하게 사용될 수 있다. 이러한 특성을 반영한 것이라고 해석할 수 있다.

**6. 결론**

영화평점을 통해 감성분석을 시행할 경우 데이터의 크기가 클 경우(약 20만건 이상)로지스틱 회귀분석을 사용한 감성분석이 효과적인 것을 확인 할 수 있다. 반면 데이터의 크기가 작은 경우는 hinge loss를 사용한 분석 즉 SVM이 더 효과적으로 분석할 수 있다.

또한 평점분석에 있어서 뮤지컬, 가족, 애니메이션, 다큐멘터리, 멜로, 서스펜스, 판타지, 액션 장르의 경우는 장르별 특수성이 존재하므로 전체 영화에 대한 평점을 통해 감성분석을 시행하기보다는 해당하는 장르내에서 감성분석을 시행할 때 더 좋은 결과를 얻을 수 있을 것이다.

**7. 추가연구과제**

이번 연구에 사용된 모델들의 경우 cross validation을 통해서 hyperparameter를 최적화한 모델이 아니기 때문에 추가적으로 hyperparameter 조정을 통해서 예측율을 끌어올린 후 비교를 진행할 수 있을 것이다. 또한 computing resource를 확보하여 머신러닝에 많이 사용하는 tree 기반의 random forest나 gradient boosting기반의 알고리즘들을 추가하여 비교해 볼 수 있을 것이다.

또한 실제로 전체데이터를 기반으로 예측한 결과와 개별 데이터를 기반으로 한 감성분석의 예측 결과가 상이하게 나오는 경우, 이러한 예측결과에 영향을 미치는 단어들을 찾아내고 이러한 단어들이 갖는 의미에 대해서 연구해 보도록 하겠다.

[참고문헌]

[1] 이상훈, 최정, 김종우, “영역별 맞춤형 감성사전 구축을 통한 영화리뷰 감성분석” 한국지능정보시스템학회, 지능정보연구 22(2), 2016.6, 97-113 (17 pages)