SNS를 활용한 영화 흥행 분석 및 예측
Analysis and prediction of movie box-office using SNS

강희석* 장은지** 전유원*** 정예지** 현민지**
*아시아나 IDT
**서울시립대학교 컴퓨터과학부
***서울시립대학교 통계학과

요약

최근 다양한 SNS가 활성화되면서 관객의 관람후기가 영화 흥행의 결정적 요소가 되고 있다. 본 연구는 SNS에서 수집할 수 있는 소비자 반응에 집중해서 영화의 흥행을 예측한다.
다양한 SNS에서 관객 반응을 수집하고, 그 중 비정형 데이터인 관람후기는 긍부정 반응을 분석해 수치화한다. 전처리한 데이터를 사용해 다중선형회귀모델과 인공 신경망 모델을 통해 분석하여 실제 영화 관객 수와의 관계를 분석한다. 또한 이를 기반으로 통계분석을 통해 흥행 예측모델을 구성한다.

연구목적

영화산업에서 영화의 흥행 결정요인에 대한 연구와 함께 상업적 측면에서 흥행예측에 대한 관심이 증대되고 있다. 또한, 최근 다양한 SNS가 활성화되면서 관객의 관람후기는 영화의 흥행의 중요한 요소로 주목받고 있다. 따라서, SNS의 텍스트, 좋아요 수 , 공유 수 등을 수집하여 영화의 관객수를 예측하는 데에 본 연구의 목적이 있다.

연구방법

- **데이터 수집 및 전처리**

Python과 R을 이용해서 Facebook, Twitter, 네이버, 다음에서 영화 개봉 전 후 일주일 동안의 영화 관람 후기 데이터를 수집했다. 데이터는 좋아요 수, 평점, 댓글 수 등의 정형데이터와 비정형 데이터인 텍스트 데이터로 이루어져있다.
텍스트데이터는 koNLPy를 사용하여 형태소 분석을 하고, KOSAC을 사용해 긍부정 반응을 분석해 변수를 생성했다. 전처리 과정을 모두 마친 후에는 총 30개의 특성변수가 생성되었다.

	moviename	review	date	rate	reply	likes	retweet	snsflag
0	택시운전사	EDIT)택시운전사- 바비 ⓒHDK#BOBBY#바비#김지원#iKON#택시운전사pic...	2017-07-26 0:00	0	113	188	twitter
1	택시운전사	택시운전사주연은 물론이고 조연에도 여캐 한명 없는거 실화냐? 5.18은 남자끼리 하...	2017-07-26 0:00	45	407	3316	twitter
2	택시운전사	좌파의 특징이 역사나 과거를 현재로 끌고와 사고를 (좌파체제)미화하는데 적극적이...	2017-07-26 0:00	2	48	80	twitter
3	택시운전사	택시운전사랑 정년경찰 보고 하다보면... 글든슬럼버가 개봉하겠지 아pic.twitt...	2017-07-26 0:00	0	15	5	twitter
4	택시운전사	<영화 '택시운전사' 영화 번개> 다음주는 방학을 맞아 강서목민관학교는 한 주 쉽니...	2017-07-26 0:00	1	75	128	twitter

그림1

- **예측 모델링**

30개의 변수 중 가장 영향력 있는 10개의 변수를 선택했다. 관객수를 예측하기 위해 다중선형회귀모형(그림2)과 인공신경망 모형을 사용했다. 인공신경망은 그림3와 같이 구성하였다.

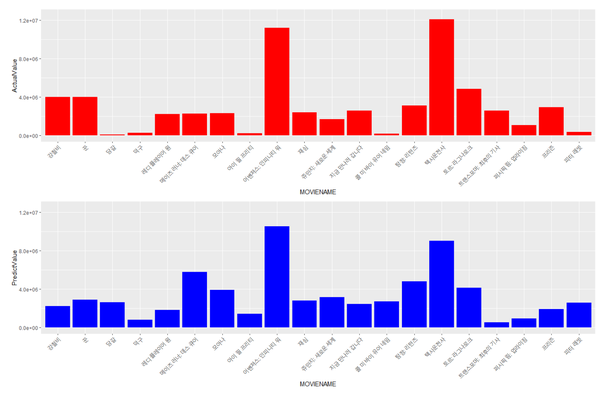

그림2

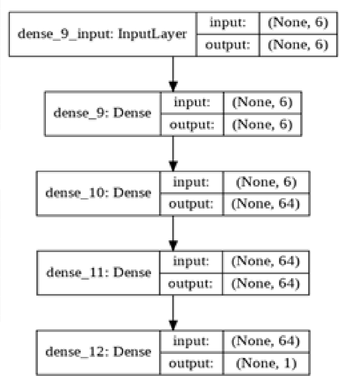

그림3

연구결과

- **다중선형회귀 예측 결과**

구분	변수 30개	변수 10개	변수 5개
p-value	4.493e-09	5.968e-12	3.196e-13
Adjusted R-squred	0.6174	0.5426	0.5201
정확도	62.75%	92.54 %	92.87 %

변수선택을 했을 때, 정확도가 증가했다.

- **인공신경망 예측 결과**

예측 관객 수(명)	실제 관객 수(명)	상대오차
264,977	240,148	10.3 %
5,415,578	11,211,627	51.7 %
3,248,625	1,335,193	143.3 %
2,541,335	2,614,601	2.8 %
736,734	903,195	18.4 %
평균 상대오차	41.6 %	

결론

- 다중선형회귀 모델을 사용하여 관객 수를 예측한 결과, 92.87%의 높은 정확도를 보였다. 이는 영화 흥행 예측에 실제로 사용할 수 있을만큼 높은 신뢰도를 갖고 있는 것으로 볼 수 있다. 또한 이 때 사용된 변수는 리뷰 텍스트를 이루는 형태소의 긍부정 비율, 좋아요 수, 공유 수 등을 포함하고 있었다.
- 인공 신경망 모델로 예측한 결과 평균 상대오차가 크고 시행마다 값의 편차가 커서 신뢰성이 높지 않았다. 이는 부족한 학습 데이터의 양이 원인인 것으로 판단된다. 이러한 문제는 더 많은 데이터를 수집하고, Gaussian Noise를 이용한 Data Augmentation 기법을 적용하여 보완할 수 있을 것이다.
- 시간의 흐름에 따라 SNS의 반응을 고려할 수 있는 시계열 분석을 더하여 SNS 반응의 변화 흐름에 따른 관객 수의 추이를 보다 정확하게 예측할 수 있을 것으로 판단된다.