Information Systems Review
Vol. 16, No. 3 December 2014 http://dx.doi.org/10.14329/isr.2014.16.3.111

영화 흥행성과 예측을 위한 온라인 리뷰 마이닝 연구: 개봉 첫 주 온라인 리뷰를 활용하여

Predicting Movie Revenue by Online Review Mining: Using the Opening Week Online Review

조 승 연 (Seung Yeon Cho) 연세대학교 정보대학원
김 현 구 (Hyun-Koo Kim) 연세대학교 정보대학원
김 범 수 (Beomsoo Kim) 연세대학교 정보대학원
김 희 웅 (Hee-Woong Kim) 연세대학교 정보대학원, 교신저자

요 약

온라인 리뷰는 네트워크 기술의 발전을 통해 그 영향력이 확대되고 있다. 특히, 사전 정보로 통해 소비가 결정되는 영화는 온라인 리뷰가 소비자들의 영화 결정에도 중요한 영향을 미치고 있다. 이에 본 연구는 영화관련 온라인 리뷰를 영화 소비 후 소비자들의 평가 정보라 가정하고, 이를 활용한 영화 흥행성과 예측모형을 제시하고자 한다. 선행 연구를 통하여 영화관련 온라인 리뷰에 감독, 배우, 스토리, 효과 등의 독립적인 속성 및 종합적인 평가가 있음을 확인하였으며, 본 연구에서는 각 속성을 2개 이상 평가하고 있는 복합형 리뷰 10가지를 추가하여 총 15가지로 온라인 리뷰 분류하였다. 2010년부터 2013년까지 개봉한 한국영화 중 상업영화 209개의 개봉 첫 주 온라인 리뷰를 온라인 리뷰 마이닝을 진행하고, 최종적으로 리뷰 마이닝 결과를 판별분석을 통한 영화 흥행성적 예측모형을 제시한다. 판별분석을 실시한 결과, 온라인 리뷰로부터 도출된 감독, 배우, 효과 및 스토리 관련 평가와 개봉 첫 주 전체 온라인 리뷰 수가 유의미하게 변별하였다.

키워드 : 온라인 리뷰 마이닝, 판별분석, 예측모형

Ⅰ. 서 론

최근 온라인 리뷰를 통해 제품이나 서비스에 대한 소비자들의 만족 정도(Wallace *et al*., 1993)와 불만족 요소를 찾아낼 수 있어(Litman and Kohl, 1989), 고객 확장 및 관리를 위한 필수적인 요소로 사용되고 있다(Eugene and Mary, 1993). 이와 같이, 온라인 리뷰는 Word-of-Mouth를 생성할 수 있는 중요채널로 인식되고 있다(Min and Lee, 2005). 특히, 영화와 같은 경험 재에서는 사전 정보 습득 및 평가정보가 영화 관람 결정에 중요한 영향을 미친다. 다시 말해, 영화를 관람

† 본 연구는 미래창조과학부 및 한국인터넷진흥원의 "고용계약형 정보보호 석사과정 지원사업"의 연구 결과로 수행되었음(과제번호 H2101-14-1001).

하기 전 관객들은 영화의 속성 및 주위 정보에 의존해서 영화를 선택하려는 경향이 있으며(De Vany and Walls, 1996), 영화 관람 후의 온라인 리뷰는 관람 전 관객의 영화 선택에 영향을 미칠 수 있다(Vapnik, 1995).

우리나라의 경우, 한국영화진흥원에서 1,530명을 대상으로 조사한 결과, 전체 51.1%가 인터넷을 통해 영화 정보를 얻고 있으며, 이들 중 70.1% 포털 사이트를 활용하고 있다. 즉, 한국 영화 소비자들의 대부분은 인터넷을 통해 관람 영화를 찾고 있으며, 동시에 온라인 리뷰에도 노출되고 있다.

이러한 추세에 따라 많은 연구들은 온라인 영화 리뷰가 영화 흥행성과에 미치는 영향에 대하여 연구가 진행되고 있다. 과거에 영화 흥행 성적 관련연구에서는 감독 및 출연배우의 영향, 제작비, 상영 스크린 수 등의 제품 속성을 중심으로 이뤄졌으나, 현재는 이러한 제품 속성 이외에도 온라인 리뷰의 영향력을 반영하는 시도가 이뤄지고 있다. 하지만, 이러한 연구들은 온라인 리뷰의 수와 리뷰의 평점만을 활용하는 한계를 가지고 있다. 이와 더불어, 텍스트 마이닝과 같은 분야에서도 영화 관련 온라인 리뷰를 분석하고자 하는 시도가 지속적으로 이뤄지고 있지만, 이러한 연구들에서는 각 리뷰의 감정분류를 통해, 긍정/중립/부정을 구분하는 특징을 찾아내거나, 텍스트 마이닝 알고리즘의 성능을 비교하는 연구가 주로 이루어지고 있는 실정이다. 이러한 상황에서 온라인 리뷰는 영화 관람 후 소비자가 직접 작성한 구매 후기이며, 해당 영화에 대한 만족도를 표현한다. 따라서, 영화 관련 온라인 리뷰를 통해 현재 상영 중인 영화들에 대한 소비자들의 만족도를 파악할 수 있는 동시에, 어떠한 요소가 만족도에 영향을 미쳤는지를 파악할 수 있다. 또한, 수명주기가 짧은 영화의 흥행은 불확실성이 높아 모형화 하기에는 쉽지 않지만(Addi and Wlliams, 2010), 영화의 흥행 정도는 영화 제작사, 배급사 등 이해당사자들의 수익과 직결되는 문제이므로 개봉 초기 정보를 활용한 예측 또한 매우 의미 있는 일이 될 것이다.

이에 본 연구는 연구의 범위를 국내에서 개봉한 한국영화로 한정하였으며, 개봉 첫 주 동안 작성된 온라인 리뷰에서 소비자들이 영화의 어떠한 요소에 의해 만족을 했는지를 온라인 리뷰 마이닝을 통해 분류하였다. 더 나아가 온라인 리뷰 마이닝 결과를 활용하여 최종 영화 흥행성과에 대한 예측 모형을 개발하고자 한다.

Ⅱ. 이론적 배경

2.1 영화 관련 정보와 흥행

영화의 흥행요인에 대한 연구는 크게 영화 제품의 속성과 비평가나 광고, 구전 등 정보를 제공하는 정보 제공자에 대한 연구로 구분된다(Eliashberg and Shugan, 1997). 제품 속성에 관련해서는 제작비와 장르, 감독 및 배우 파워가 영화의 흥행 수익에 중요 요인이며, 등장인물, 스토리, 연기속성이 중요 흥행요인임이 밝혀졌다(Li *et al*., 2006). Wallace *et al*.(1993)은 1950년대 영화 1,687개를 대상으로 출연 배우를 중심으로 연구를 수행하였으며, 영화 수익과 스타파워간에 상관관계가 있음을 확인하였다.

비평(전문가들에 의한 비평)의 경우, 과거 인터넷이 발전되기 전에는 흥행에 중요요인 이었으나(Reddy *et al*., 1998; 강문수 외, 2012), 온라인 발전에 따라 그 영향력이 감소함을 확인하였다(De Vany and Walls, 1996; Hatzivassiloglou and McKeown, 1997). 이와 더불어, 전문가들에 의한 비평이 영화 흥행과 부정적인 관계가 있다고 주장하는 연구도 있다. 영화관련 전문가들의 비평은 상업성보다는 예술성을 더 높게 평가하는 경향이 있기 때문에, 전문가들의 비평이 좋을수록 영화 수익이 나쁘다고 주장하고 있다(Liu, 2006; 성영신 외, 2002).

영화 광고와 관련해서, Lehmann and Weinberg

(2000)는 광고를 많이 한 영화일수록 인지도와 기대감이 높아지고, 이로 인해 흥행 성적이 높다고 주장하고 있다. 하지만, 광고 등을 통한 사전 정보를 획득한 소비자에게 영화가 주관적 기대 수치를 만족시키지 못할 경우에는 그 실망 정도가 그렇지 않은 관람객에 비해 높게 나타난다(Bowman *et al*., 2001). 실망 정도가 큰 소비자들은 영화 관람 후 부정적 의견을 주는 정보 제공자가 될 가능성이 높다.

온라인 리뷰는 온라인 환경을 통해 소비자들이 제품이나 서비스에 대해 직접 경험한 정보를 교환하는 커뮤니케이션 활동으로(박승현, 장정헌, 2012), 영화와 같이, 관람하기 전에는 그 속성을 평가하기 어려운 제품이나 서비스에서는 온라인 리뷰의 영향력이 더 크게 나타난다(De Vany and Walls, 1996). 또한, 온라인 리뷰에서 영화에 대한 평가는 주관적으로 이뤄지지만, 실제 소비자들은 온라인 리뷰의 평가 정보를 더욱 신뢰하는 경향이 있다(Eliashberg and Shugan, 1997).

전문가 비평과 일반 영화관람객이 작성한 온라인 리뷰를 비교하는 연구도 있다. Holbrook (1999)은 전문가 비평과 온라인 리뷰 간에는 유의미한 차이가 존재하며, 이 둘의 영향력은 서로 간 반비례의 관계가 있다고 주장하였다. Litman and Kohl(1989)은 개봉 전에는 전문가 비평의 영향력이 강하나, 영화가 개봉한 이후에는 일반인에 의한 구전효과가 영화 선택에 강하게 영향을 미친다고 했다. 즉, 온라인 리뷰를 통해 영화의 품질이나 특성에 대한 평가를 제시하는 경우가 증가함에 따라, 소비자에 의한 온라인 리뷰가 영화 흥행에 직접적인 영향을 미치고 있다(이성직, 김한준, 2009). 실제로 2000년대부터 전문가들에 의한 비평보다는 온라인 리뷰가 흥행성과에 주요한 영향을 주고 있는 것으로 분석되고 있다(Bo pang *et al*., 2002; Dellarocas, 2003). 특히, 온라인 리뷰의 수(Volume)는 영화 매출과 유효한 상관관계가 있음이 확인되고 있다. Duan *et al*. (2005)는 관객들의 평가 점수(Valence)는 영화의 매출에 유효한 설명력을 가지지 못하지만, 온라인 리뷰의 수(Volume)은 매출액과 상관관계가 있음을 확인하였다. Liu(2006)은 개봉 전과 개봉 후로 비교하여, 개봉 전 높은 기대로 인해 평가 점수는 개봉 후에 오히려 낮아지는 반면, 온라인 리뷰의 수는 영화 매출 규모에 유효한 설명력이 있다고 하였다.

하지만 기존 연구들에서의 영화 리뷰는 관객들의 평가 점수와 리뷰의 양에 집중되었으며, 각 리뷰가 어떠한 특징을 가지고 쓰여졌는지에 대한 고려가 부족하다. 이에 본 연구는 온라인 리뷰 마이닝을 통해 영화의 어떠한 요소가 관객들의 만족도에 영향을 주었는지를 분류하고, 이를 활용하여 영화 흥행성과 예측모형을 개발하였다. 더 나아가 어떠한 온라인 리뷰가 영화 흥행과 관련성이 있는지 알아보고자 한다.

2.2 영화 온라인 리뷰 분류

기존의 온라인 리뷰에서의 감정분류 관련 연구는 관련 지식을 바탕으로 이루어졌으며(Linton and Petrovich, 1988), 언어적 추론을 사용하여 의미 있는 단어나 구의 분류에 초점을 맞추고 있다(Bo pang *et al*., 2002; Dellarocas, 2003). Li *et al*. (2006)은 11개 분류 기준을 제시하고 분석하여, 각 분류 기준 별 주요 감정 키워드를 요약했다. Bo pang *et al*.(2002)은 긍정적인 리뷰 1300개, 부정적인 리뷰 750개를 대상으로 기계학습방법(Machine learning method)들의 성능을 비교하고 있다.

Neelamegham and Jain(1999)은 리뷰에 작성된 속성을 크게 3가지로 구분하고 있다. 첫 번째, 핵심적 속성은 영화의 내용과 직접적으로 연관된 속성으로 스토리, 연기, 출연배우를 포함한다. 두 번째 주변적 속성은 배경, 의상, 배경음악, 특수효과 등을 포함한다. 마지막 정서적 자극은 영화를 본 후 소비자가 느낀 감정을 의미한다. 예를 들어, 감동과 슬픔, 재미 등이 이에 해당된다.

마지막으로, Li *et al.*(2006)은 영화 속성 요소와 인물요소로 온라인 리뷰의 속성을 분류하고 있다. 영화 속성 요소로는 각본, 캐릭터, 시각효과, 음향효과, 특수효과 등 영화의 내용과 직접적으로 연관된 속성들을 포함하며, 인물요소는 영화 제작에 관련된 사람들에 관련된 속성으로, 제작자, 감독, 작가, 배우 등을 포함하고 있다. 최종적으로 이 연구에서는 온라인 리뷰 내에서 속성과 감성 간의 조합을 분류할 수 있는 알고리즘을 제시하고 있다. 이와 같이, 대부분의 온라인 리뷰 마이닝 연구들은 긍정/중립/부정 리뷰를 분류하는 것에 집중하고 있는 반면, 일반 관객들이 영화의 어떤 요소에 주목하여 영화 리뷰를 작성하는 지에 대한 체계적인 지식이 부족한 상황이다(Dellarocas 2003).

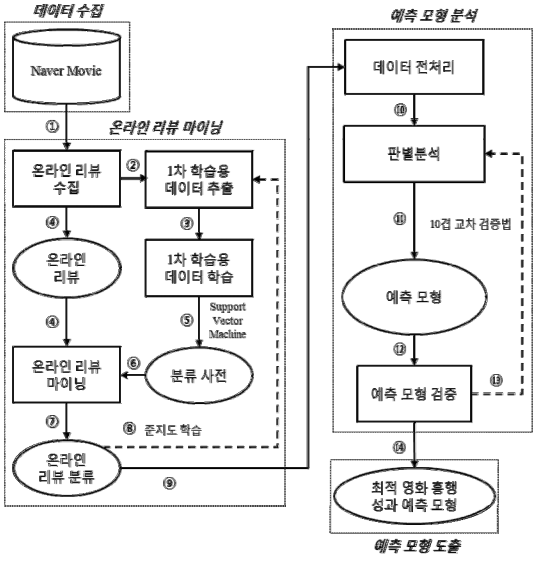
이와 달리, 특정 영화의 리뷰를 내용분석하고, 온라인 리뷰 간의 차이를 알아보고자 하는 연구들도 존재한다. 박승현, 송현주(2010)는 '박쥐(2009)'의 리뷰 300개를 내용 분석하여, 감독, 배우, 스토리, 대중성, 성정성, 시각적 스타일 등 총 11개의 분류 기준을 제시하고 비교하였다. 박승현, 장정헌(2012)은 애니메이션 영화인 "마당을 나온 암탉(2011)"의 전문가의 비평과 소비자가 작성한 온라인 리뷰를 비교했다. 이들은 영화의 품질, 연출력, 연기력, 스토리, 시각적 스타일로 분류 기준을 만들고, 연구를 수행하였다. 오은희, 전범수(2008)는 전문가 비평과 온라인 리뷰에서 장르, 제작 국가, 배우, 감독, 수상 이력 등의 평가가 이뤄지고 있음을 확인했으며, 특히, 소비자들은 유명 배우의 출연과 수상이력이 있을수록 평가를 좋게 하는 경향이 있다고 주장했다.

이렇듯 온라인 리뷰에 작성된 속성에 대하여 분류한 기준들은 각 연구 별로 차이가 있지만, 공통적으로 감독과 배우, 스토리 속성을 제시하고 있다. 또한, 기존 연구에서는 각 리뷰 별 속성을 하나씩 가지고 있다고 가정하고 진행하였다. 하지만 실제 온라인 리뷰는 140글자 내외로 작성될 수 있으며, 기존 연구에서의 가정과 달리, 한 리뷰에 여러 속성을 가지고 작성 될 수 있다. 이에 본 연구에서는 이러한 문제를 해결하고자, '복합형 리뷰'라는 분류 기준을 제시하고 연구를 진행하였다.

Ⅲ. 연구 방법

3.1 연구 절차

본 연구는 '네이버 영화(Naver Movie)'에 작성된 일반관객들의 온라인 리뷰를 온라인 리뷰 마이닝을 통해 분석하고, 이를 통한 영화 흥행성과 예측모형을 제시한다. 전체 연구에 대한 절차는 다음 <그림 1>과 같이 총 네 단계로 진행되었다. 먼저, 본 연구에서는 '네이버 영화'에서 일반 관객들이 작성한 온라인 리뷰를 수집하기 위한 Crawler를 개발하고, 각 영화에 대한 온라인 리뷰를 수집하였다(①). 다음 온라인 리뷰 마이닝 단계에서는 학습 데이터를 통한 지도 학습(Supervised-Learning)이 적용된 기존 연구들의 과정을 적용하여(Bo pang *et al.*, 2002; Li *et al.*, 2006), 수집된 리뷰 데이터를 두 분류로 구분하여 진행하였



〈그림 1〉 온라인 리뷰 마이닝을 통한 연구 절차

다(②, ④). 첫 번째는 단어 중심의 분류 사전을 만들기 위한 학습용 데이터로써 수립된 온라인 리뷰 분류 기준을 근거하여 분류되고, 서포트 벡터 머신(Support Vector Machine)을 통한 학습이 이뤄졌다(⑤). 학습용 데이터를 통해 구축된 분류 사전에 의해 다른 온라인 리뷰들의 분류를 진행했으며(⑥), 이를 통해 영화 별 리뷰 분류 결과가 도출되었다(⑦). 하지만, 1차 학습용 데이터에 의한 분류 사전의 다양성이 적어 분류의 정확도가 떨어질 수 있다(이동주 외, 2011). 이에 본 연구에서는 준지도 학습(Semi-supervised learning) 방식을 적용하여 분류사전의 다양성을 확충하고, 이를 통해 다시 온라인 리뷰 마이닝을 진행하였다(⑨).

하지만, 기존 지도학습이 이뤄진 연구들에서는 학습을 통해 얻어진 결과를 해석하는 것에서 연구가 마무리 되고 있다. 이와 달리, 본 연구에서는 각 영화 별 작성된 온라인 리뷰를 분류한 결과를 통해 영화 별 최종 흥행성과 예측모형을 개발하기 위해서, 앞서 진행한 온라인 리뷰 마이닝 결과들을 데이터 전처리(Data preprocessing)을 통해 변수화(⑩) 하여 그 다음 과정까지 연구를 진행하였다. 데이터 전처리 후 다음 단계에서는 K겹 교차 검증(K-fold Cross Validation) 방법을 적용한 판별함수 예측모형들을 도출하고, 검증하였다(⑪, ⑫). 최종적으로 최적 영화 흥행성과 예측모형을 제시한다(⑬).

3.2 온라인 리뷰 분류 기준

각 온라인 리뷰는 $R = r_1, r_2, \cdots, r_n$과 같은 단어들의 조합이며, 이러한 온라인 리뷰들은 기존 연구들에서 감독, 배우, 스토리 등으로 공통적으로 구분하고 있다(Dellarocas, 2003; Addi and Wlliams, 2010; 김광수, 2000; 박형현, 박찬수, 2001). 따라서 본 연구에서도 감독, 배우, 스토리 관련 분류를 동일하게 사용한다.

‘감독’과 관련된 온라인 리뷰는 감독의 연출력, 명성, 수상 이력 등 감독에 관련하여 직접적으로 평가하고 있는 리뷰로 정의한다. 예를 들어, “이 감독의 영화는 나의 기대를 저버리지 않는다.” “역시 디테일이 살아있는 연출력!”과 같은 리뷰들이 이에 해당한다.

‘배우’로 분류되는 온라인 리뷰들은 출연 배우의 명성 및 연기력 등을 언급하거나, 출연 배우의 기존 출연 작품들과의 비교가 이뤄진 리뷰로 정의한다. ‘배우’로 분류되는 온라인 리뷰들로는 “A의 연기력이 감동을 더한다.” “B의 연기가 이 영화를 가득 채운다.” 등이 있다.

‘스토리’로 분류된 온라인 리뷰는 해당 영화의 이야기 전개 및 시나리오 등 영화의 소재를 언급하는 것으로 정의하며, “이 영화의 스토리가 너무 좋아!” “대사 한마디 한마디가 내 심장을 파고 드는구나!” 등이 이에 해당된다.

Neelamegham and Jain(1999)은 영화의 주변적 속성으로서 배경음악 및 특수효과 등의 시각효과 등이 있으며, 온라인 리뷰의 3가지 구성 요소 중 하나라고 했다. 따라서 본 연구에서는 이러한 주변적 속성을 언급하는 온라인 리뷰를 ‘효과’로 분류한다. ‘효과’로 분류되는 온라인 리뷰에는 “마지막 장면에서의 음악 제목이 뭐야? 너무 좋아!” “컴퓨터 그래픽이 너무 실감난다.” 등이 포함된다.

영화 관련 온라인 리뷰에는 영화의 속성에 대한 언급 없이, 단순히 추천만이 이뤄지고 있는 리뷰도 존재한다. 예를 들어, “강추!” “후회 안함” 등이 있으며, 본 연구에서는 이러한 온라인 리뷰를 “전반적 평가”로 분류한다(김광수, 2000).

기존 연구들에서는 각 온라인 리뷰에는 한 가지 속성을 포함한다고 가정하고 연구가 이뤄졌다(Dellarocas, 2003; 김광수, 2000). 하지만 본 연구에서 수집한 온라인 리뷰 사이트에서는 한 명의 사용자가 140자 내로 리뷰를 작성할 수 있으며, 각 리뷰에 2가지 이상의 속성을 언급할 수 있다. 예를 들어, “탄탄한 시나리오 구성과 몸을 사리지 않는 OO의 열정이 XX감독의 디테일을 만나 영화를 더욱 재밌게 만들었다.”라는 온라인

리뷰는 앞서 제시한 “감독”, “배우”, “스토리”를 모두 언급하고 있다. 해당 리뷰를 독립적 속성에 따라 분류하기 어려우며, 거의 모든 영화에서 등장하였다. 따라서 본 연구에서는 “복합형 리뷰”라는 범주를 추가하여, ‘감독’, ‘배우’, ‘스토리’, ‘효과’를 동시에 평가하고 있는 리뷰를 다음 <표 1>과 같이 10가지 경우로 나누어 진행하였다.

〈표 1〉 복합형 리뷰 분류 기준 정의

복합형 리뷰	영화 속성
복합형_1	감독 + 배우 + 스토리
복합형_2	감독 + 배우 + 효과
복합형_3	감독 + 스토리 + 효과
복합형_4	배우 + 스토리 + 효과
복합형_5	감독 + 배우
복합형_6	감독 + 스토리
복합형_7	감독 + 효과
복합형_8	배우 + 스토리
복합형_9	배우 + 효과
복합형_10	스토리 + 효과

3.3 온라인 리뷰 데이터 수집

앞서 정의한 분류기준에 의해 영화 별 온라인 리뷰에 대한 온라인 리뷰 마이닝을 수행했다. 따라서 영화 별 소비자가 작성한 온라인 리뷰를 수집했다. 한국 영화진흥위원회에서는 2010년부터 2013년 3분기까지 한국에서 개봉한 영화는 총 2,150개이며, 그 중 한국에서 제작된 상업영화는 총 209개가 개봉되었다. 본 연구에서는 이들 상업 영화의 온라인 리뷰를 활용하여 연구를 수행하였다.[1)]

해당 영화들의 개봉 첫 주 동안 작성된 리뷰는 총 466,651개가 작성되었다. 기존 영화 리뷰를 활용한 연구들에서는 감성분석(Sentiment analysis)등을 활용하여 리뷰를 작성한 소비자의 감성 정도를 측정하였지만(Linton and Petrovich, 1988; 김광수 2000), 본 연구에서는 각 리뷰의 감성 정도를 소비자가 직접 입력한 별점으로 대체하고 진행하였다. ‘네이버 영화’의 경우, 소비자가 1~10점까지의 별점을 줄 수 있으며, 1~3점은 부정적 의견, 4~7점은 중립적 의견, 8~10점은 긍정적 의견이라고 가정할 수 있다(Kennedy and Inkpen, 2006). 그러나 부정적 의견과 중립적 의견의 리뷰들은 영화 속성 이외에 정치, 사회상황 등 영화 외적 요소를 가지고 작성되는 경우가 많았으며, 영화 관람 전 선입견을 가지고 작성된 리뷰가 많았다. 그러므로 본 연구에서는 8점 이상 긍정적 의견의 리뷰들을 중심으로 연구를 진행했다.

다음 <표 1>은 본 연구가 수행하는 데이터의 인구통계학적 특징을 요약한 것이다. 각 영화의 스크린 수와 관람객 수는 한국영화진흥위원회의 영화관입장권 통합전산망(www.kobis.or.kr)에서 수집하였다.

〈표 2〉 개봉 첫 주 데이터의 인구통계적 특성

(단위: 개, 명)

	총합	평균	표준편차
긍정 리뷰	307,385	1470.74	2139.28
상영관 수	90,236	429.11	218.20
관람객 수*	356,229,743	1,697,022	2,261,933

주) *관람객 수는 최종 관람객 총 수를 의미한다.

Ⅳ. 온라인 리뷰 마이닝

본 장에서는 영화 온라인 리뷰 마이닝의 결과를 설명한다. 영화 리뷰 마이닝을 진행하기에 앞서, 온라인 리뷰 수집기로 수집된 중 2,500개의 리뷰에 대한 내용 분석하였고, 이를 학습 데이터로 사용하였다. 온라인 리뷰 마이닝은 서포트 벡터 머신 분류기를 활용하였으며, TF-IDF 값을 단어 별 가중치로 사용하였다.

1) 209개 영화 중 2010년에 개봉한 영화는 53편이며, 2011년은 57편, 2012년은 62편, 2013년은 37편을 대상으로 연구하였다.

4.1 1차 학습 데이터

온라인 리뷰 마이닝을 수행하기 위한 학습 데이터를 만들기 위하여, 2010년 이전까지 관람객 수가 가장 많은 한국영화 5개(왕의 남자, 괴물, 과속스캔들, 국가대표, 해운대)의 8점 이상인 영화 리뷰를 내용 분석하였다. 이들 영화는 다음 <표 3>과 같이 다양한 장르로 구분되고 있으며, 실제 온라인 리뷰의 형태 또한 다양하게 작성되어있다. 따라서 1차 학습 데이터로써 분류 사전의 다양성을 확보하기 위해 적절한 데이터라고 판단하였다.

영화 리뷰에 대한 내용분석 결과의 객관성을 더하기 위해 연구자들은 개별적으로 온라인 리뷰를 분류한 후, 모두 동일하게 분류한 리뷰만을 사용하였다. 1차 학습 데이터는 영화당 500개씩, 총 2500개의 리뷰로 이루어졌으며, 결과는 <표 4>와 같이 분류되었다.

〈표 3〉 1차 학습 데이터 선정 영화의 특성

(단위: 원, 명)

영화제목	장르	누적 매출액	누적 관객수
왕의 남자	사극, 드라마	66,015,436,400	10,513,715
괴물	SF, 가족, 드라마	66,716,104,300	10,917,221
과속스캔들	코미디, 가족, 드라마	53,800,771,400	8,223,266
국가대표	드라마, 액션, 코미디	57,570,773,000	8,035,181
해운대	액션, 드라마, 어드벤처	81,025,004,000	11,324,433

자료: 영화입장권 통합전산망(www.kobis.or.kr).

〈표 4〉 1차 학습용 데이터 분류 결과

(단위: 개)

		왕의 남자 (2005)	괴물 (2006)	과속스캔들 (2008)	국가대표 (2009)	해운대 (2009)	합계
감독		28	72	12	46	8	166
배우		100	59	204	53	75	491
스토리		102	64	44	76	68	354
효과		31	94	11	57	78	271
전반적 평가		105	72	172	78	78	505
복합형	감독 + 배우 + 스토리	14	4	1	7	0	26
	감독 + 배우 + 효과	3	6	0	3	2	14
	감독 + 스토리 + 효과	0	1	0	2	1	4
	배우 + 스토리 + 효과	25	13	5	7	15	65
	감독 + 배우	21	14	7	17	1	60
	감독 + 스토리	4	9	0	6	0	19
	감독 + 효과	1	6	0	2	6	15
	배우 + 스토리	33	23	19	98	47	220
	배우 + 효과	15	29	17	12	29	102
	스토리 + 효과	18	34	8	36	92	188
Total		500	500	500	500	500	2,500

4.2 온라인 리뷰 분류 기법

리뷰의 분류를 수행하기 위하여 Java로 구현되어 있는 MALLET Package를 사용하였으며, 분류기로는 서포트 벡터 머신 분류기를 사용하였다. 서포트 벡터 머신은 Vapnik(1995)가 제안한 통계학적 학습 이론에 기반을 둔 기법으로 집단으로 구분된 학습 데이터를 분류할 때 기준이 되는 경계면을 찾아낸다. 이를 이용하여 분류기에 입력되는 다른 데이터가 어느 분류 기준에 속하는지를 구분한다. 이 연구에서는 먼저 서포트 벡터 머신 분류기에 학습 데이터를 입력하면 분류기가 분류기준과 그곳에 해당하는 단어가 무엇인지 학습하게 되고, 그 후 분석 대상이 되는 리뷰를 입력하면 어떤 기준에 적합한 리뷰인지 분류기에 의해 결정된다. 그 외의 분류 기법으로는 최대 엔트로피(Maximum Entropy)나 의사결정 나무(Decision tree), 신경망, 나이브 베이지안(Naïve Bayesian) 등 여러 가지 기법이 있으나, 불균형 데이터 집합을 분류할 경우 상대적인 차이점을 고려하지 않아 다수의 집합은 정확한 분류가 가능하지만, 소수의 집합은 무시되어 정확한 분류가 어렵다(강현철 외, 2010). 이와 달리, 서포트 벡터 머신은 불균형 데이터 집합에 대해 상대적으로 높은 성능을 보이기 때문에 본 연구에 적합하다고 판단하였다(박승현, 정완규 2009).

텍스트 마이닝을 위한 서포트 벡터 머신은 영어를 기본으로 만들어진 분류 기법이기 때문에 한글을 분석할 경우 형태소 분석기를 사용하여 리뷰의 형태를 바꾸어 주어야 한다. 본 연구에서는 온라인 리뷰를 분류하기 전 선행작업으로 형태소 분석을 진행하였다. 온라인 리뷰에서 흔히 존재하는 띄어쓰기 오류에 상대적으로 덜 민감한 한글 형태소 분석기인 꼬꼬마 형태소 분석기(KKMA, Kind Korean Morpheme Analyzer)를 사용하여 리뷰를 분석, 1차 학습 데이터와 분류 대상이 되는 모든 리뷰를 단어 단위로 나누었다(Prag and Casavant, 1994). 이 과정에서 문법적 오류가 심하거나 오타, 관용적 표현의 사용 등의 이유로 형태소 분석기가 분석하지 못하는 리뷰 19,197개가 제외되었다.

4.3 특징 단어 목록

1차적으로 본 연구에서는 내용분석을 통해 분류된 2,500개의 리뷰에 의한 1차 학습 데이터와 서포트 벡터 머신 분류기를 이용하여 1차 분류 사전을 구축하고, 이를 활용하여 모든 리뷰를 자동분류하였다. 하지만, 1차 학습 데이터를 통해 구축된 분류 사전에 포함된 단어는 제한적이다. 이러한 문제를 해결하기 위하여, 준지도 학습(Semi- supervised Learning)방법을 적용하여 2차 분류 사전을 구축하였다. 준지도 학습은 1차 분석결과로 나온 영화 중21개를 임의로 추출하였다.[2] 그 뒤 1차 학습 데이터와 임의로 선정된 영화의 리뷰를 포함한 총 58,737개의 리뷰를 2차 학습 데이터로 사용하여, 모든 리뷰를 다시 한 번 분류하였다. 2차 학습 데이터에 대한 리뷰는 다음 <표 5>와 같다.

〈표 5〉 2차 학습용 데이터

(단위: 개)

분류	리뷰 수	분류	리뷰 수	분류	리뷰 수	분류	리뷰 수
감독	3,541	복합형_4	182	전반적평가	35,643	복합형_8	2,607
배우	4,753	복합형_5	488	복합형_1	65	복합형_9	310
스토리	5,283	복합형_6	24	복합형_2	15	복합형_10	2,295
효과	3,479	복합형_7	47	복합형_3	4	총 합	58,736

2) 맨발의 꿈, 설국열차, 반가운 살인자, 두여자, 더 테러 라이브, 두 개의 달, 풍산개, 특수본, R2B_리턴투베이스, 7번방의 선물, 화이트 저주의 멜로디, 더 웹툰-예고살인, 시라노 연애조작단, 무서운 이야기 2, 꿈은 이루어진다, 적과의 동침, 댄싱퀸, 라스트 갓파더, 조선명탐정-각시투구꽃의 비밀, 네버엔딩 스토리, 하모니

분석 결과, ‘감독’ 분류에서는 감독, 작품, 연출, 표현력 등 감독 및 감독의 연출력에 관련된 단어가 주를 이루고 있다. ‘배우’ 분류의 경우 배우, 연기력, 아역, 캐스팅 등 배우의 특징을 나타내는 단어뿐 아니라 ‘대박’이라는 단어가 특징 단어로 도출되었다. 이는 “하정우 대박 멋지십니다!!”, “송강호 연기 대박임 웃기고 재미있슴~” 등 영화 전체적인 배우들의 평가 외에도 특정 배우에 대한 평가가 이루어지고 있기 때문이다. ‘스토리’ 분류의 경우 감동, 장면, 마지막, 풍자 등 특정 장면에 대한 평가와 스토리, 시나리오, 전개 등 전반적인 스토리에 대한 평가에 관련된 단어들이 포함된 것을 확인 수 있다. ‘효과’ 분류는 그래픽이나 CG, 음악, 실감 등 시각적, 청각적 요소와 관련된 단어가 포함 되었다. 특히, “우리나라의 CG가 이 정도라니~!”와 같이, 국내 그래픽 기술에 대한 평가가 상당수 등장했다. 전체 온라인 리뷰 중 가장 많은 수를 차지하고 있는 ‘전반적 평가’는 최고, 대박, 후회, 완전 등 단순한 내용을 포함하고 있다.

본 연구에서 추가한 ‘복합형 리뷰’는 감독, 배우, 스토리, 효과 중 리뷰에 두 개 이상의 단일 분류 기준이 동일 리뷰에 작성된 것으로, “배우들의 연기, 영화의 완성도, 스토리, 모두 만점을 줘도 아깝지 않다”, “감독의 열정과 배우들의 노련미와 엄청난 그래픽” 등의 리뷰가 이에 해당된다. 각 ‘복합형 리뷰’를 구성하는 단어도 마찬가지로 감독, 배우, 스토리, 효과를 구성하는 단어가 조합된 형태로 구성되어 있다. <표 6>은 각 분류 기준 별 TF-IDF 가중치[3] 값이 높은 상위 10개의 단어를 정리한 것이다(Jo *et al.*, 2004).

〈표 6〉 분류 기준 별 분류 사전 예시

감독	감독, 작품, 연출, 짜임새, 표현력, 디테일, 연출력, 구성, 전개, 신인
배우	연기, 감동, 배우, 대박, 연기력, 아역, 캐스팅, 주연, 조연, 멋
스토리	가슴, 장면, 때, 마지막, 최고, 시나리오, 스토리, 전개, 풍자, 압권
효과	우리나라, 그래픽, 괴물, 한국, CG, 음악, 실감, 효과, 장면, 리얼
전반적평가	영화, 최고, 감독, 10점, 대박, 후회, 한국, 완전, 강추, 한국영화
복합형_1	연기, 배우, 연출, 내용, 스토리, 최고, 감독, 구성, 삼박자, 시나리오
복합형_2	연기, 연출, 영상, 감독, 최고, 배우, 효과, 연기력, 연출력, 그래픽
복합형_3	영화, 감동, 음악, 연출
복합형_4	연기, 스토리, 배우, 음악, 최고, 연기력, 감동, 내용, 역할, 각본
복합형_5	감독, 배우, 연기, 아역, 최고, 작품, 연기력, 연기자, 디테일, 연출력,
복합형_6	감독, 스토리, 연출, 최고, 전개, 실제, 감동, 속, 짜임새, 시나리오
복합형_7	노래, 그래픽, 감독, 작품, 구성, 한국, 음악, 표현력, 영상, 재난
복합형_8	감동, 최고, 재미, 내용, 연기력, 짜임새, 스토리, 완벽, 캐스팅, 기대
복합형_9	노래, 연기력, 감동, 그래픽, CG, 영상, 음악, 배경, 영상미, 노력
복합형_10	감동, 내용, 음악, 조화, 배경, 기대, 영상미, 그래픽, 재난, 기대이상

3) 문서 내부의 단어 간 상대적 중요도를 평가하기 위해 고안된 값이다. TF(Term Frequency) 값은 한 문서에서 특정 단어가 출현한 빈도이며, IDF(Inverse Document Frequency)값은 전체 문서 집합의 수를 특정 단어가 출현한 문서의 수로 나눈 값이다. 즉, TF-IDF 값은 TF와 IDF를 곱한 값이다.

〈표 7〉 온라인 리뷰 마이닝 결과

(단위: 개)

분류	리뷰	평균	Std.	분류	리뷰	평균	Std.
감독	32,212	153.39	268.18	복합형_4	699	3.33	8.05
배우	25,387	120.89	196.35	복합형_5	1,643	7.82	21.80
스토리	9,515	45.31	180.92	복합형_6	14	0.07	0.36
효과	15,472	73.68	120.26	복합형_7	127	0.60	2.64
전반적 평가	179,475	854.64	1,348.67	복합형_8	12,028	57.28	139.92
복합형_1	240	1.14	3.06	복합형_9	1,779	8.47	16.78
복합형_2	4	0.02	0.14	복합형_10	9,593	45.68	81.49
복합형_3	0	0.00	0.00				

4.4 온라인 리뷰 마이닝 결과

1차, 2차 학습 데이터를 통해 특징 단어 목록을 도출하여, 분류 사전을 구축하였다. 이를 활용하여 본 연구의 분석 대상인 307,385개의 리뷰에 대한 마이닝을 진행하였다. 그 중 형태소 분석기가 분석하지 못하는 리뷰를 제외한 288,188개의 리뷰를 다음 <표 7>과 같이 분류하였다.

'복합형 리뷰_3(감독+스토리+효과)'의 경우, 학습 데이터 중 4개의 리뷰만이 이에 해당되었다. 즉, '복합형 리뷰_3'의 분류 사전에 포함된 단어의 수가 적으며, 최종적으로 온라인 리뷰 마이닝을 통해 분류되는 리뷰가 없었다.

Ⅴ. 분석결과 및 예측모형

본 장에서는 위에서 진행한 온라인 리뷰 마이닝 결과를 활용한 예측모형에 대하여 설명한다. SAS 9.2를 활용하여 판별분석을 수행하였으며, 10겹 교차 검증법으로 도출된 예측모형의 타당성을 검증하였다.

5.1 데이터 전처리

5.1.1 종속변수의 범주화

2010년부터 2013년까지 개봉한 한국영화는 1,460편이며, 관람객 수는 총 393,612,036명이다 즉, 본 연구에서 분석하는 영화의 수는 209편이지만, 관람객 수는 356,229,743명으로 같은 기간 영화 관람객 점유율의 90% 이상을 차지한다.

본 연구는 앞서 진행한 영화 별 온라인 리뷰 마이닝 결과를 활용하여 최종 영화 흥행 성적을 예측하고자 한다. 따라서, 최종 영화 관람객 수를 다음 <표 8>과 같이 구분하였다.

〈표 8〉 영화 등급 구분

등급	설명	영화
1	개봉연도 관람객 수 상위 10%	22
2	개봉연도 관람객 수 상위 10%~50%	46
3	개봉연도 관람객 수 50%~하위 10%	121
4	개봉연도 관람객 수 하위 10%	20

1등급에 해당하는 영화의 총 관람객은 143,839,470명(평균 = 6,538,158, Std. = 3,058,393)으로 전체 209편 영화 관람객의 약 40.38%를 점유했다. 2등급 영화의 경우, 총 137,196,808명(평균 = 2,855,251, Std. = 1,230,538)이 관람했다. 하위등급인 3등급과 4등급에 해당하는 영화는 총 141편으로 가장 많은 영화들이 이에 해당한다. 하지만 하위등급 영화의 총 관람객 수는 78,193,465명으로, 209편 영화 관람객의 22.04%를 점유했다.

5.1.2 변수 변환

온라인 리뷰 마이닝의 결과에서 확인할 수 있듯이, 209개 영화 사이에서도 작성된 리뷰 수의 편차가 크다. 따라서 본 연구에서는 이러한 작성된 리뷰 수에 대한 문제를 보완하기 위해서 온라일 리뷰 마이닝 결과를 비율척도로 변환하였다. 다음 <표 9>는 예측모형을 만들기 위해 본 연구에서 사용한 변수들에 대한 설명이다.

복합형_2(감독+배우+효과)와 복합형_6(감독+스토리)는 온라인 리뷰 마이닝에 의해 분류된 리뷰의 수가 각각 4개, 14개 이므로, 본 예측모형 분석에서는 제외하였다. 이와 더불어, 본 연구에서는 온라인 리뷰 마이닝 결과 이외에 3개의 변수를 추가하였다. 기존 연구들에서는 온라인 리뷰의 양과 평가 정도를 활용하여 영화 매출액과의 관계를 파악하였다 (Duan *et al.*, 2005; Liu, 2006). 이에 본 연구에서는 총 온라인 리뷰 수(ln_review)와 긍정 리뷰 비율(Positive)를 추가하였다. 또한, Elberse and Eliashberg(2003)는 배급사와 극장 간의 협상을 통해 개봉 첫 주 상영관 수가 결정된다고 하였다. 즉, 본 연구에서는 통제변수로써 배급사 파워 및 극장과의 교섭력을 '개봉 첫 주 누적 상영관 수'(ln_screen)로 측정하여 연구를 진행하였다.

5.1.3 요인 분석 및 해석

온라인 리뷰 분류 기준에서 복합형 리뷰는 감독, 배우, 스토리, 효과 등 단일 기준이 2개 이상 복합적으로 등장하는 리뷰이다. 따라서, 분류 결과를 그대로 사용할 경우, 단일 분류 기준과의 중복이 발생할 수 있다. 이러한 문제를 제거하기 위해서, 본 연구에서는 데이터의 전처리 기법 중 요인 분석을 수행하였다(Zhu, 2005). 요인분석 방법으로는 주성분 분석과 직교 회전 방법 중 베리맥스(Varimax) 방법을 사용하였다.

요인분석을 통해 고유값이 1 이상인 주성분 3개가 추출되었으며, 이 과정을 통해 Comb_7항목을 제거하였다. 이를 통해 구성된 요인구조를 중심으로 크론바하 알파(Cronbach's α) 계수를 사용하여 신뢰도 검정을 실시하였다. 요인분석 결과는 다음 <표 10>과 같다.

〈표 10〉 요인 분석 및 신뢰도 분석 결과

구 분		요인 적재치			신뢰도*
		1	2	3	
배우 중심 평가	Comb_8	**0.846**	0.121	0.032	0.753
	Actor	**0.799**	-0.027	-0.441	
	Comb_4	**0.755**	0.444	0.100	
	Comb_9	**0.727**	-0.021	0.202	
감독 중심 평가	Comb_5	0.182	**0.815**	-0.204	0.705
	Comb_1	0.374	**0.758**	-0.017	
	Director	-0.218	**0.718**	0.037	
스토리 및 효과 평가	Comb_10	0.231	-0.066	**0.808**	0.738
	Effect	-0.243	0.093	**0.737**	
	Story	0.134	-0.175	**0.735**	

주) *Cronbach's α.

〈표 9〉 변수 정의

변수명	설명	변수명	설명
ln_review	ln(개봉 첫 주 총 온라인 리뷰 수)	Comb_1	전체 리뷰 중 '복합형_1' 리뷰 비율
ln_screen	ln(개봉 첫 주 누적 상영관 수)	Comb_4	전체 리뷰 중 '복합형_4' 리뷰 비율
Positive	전체 리뷰 중 8점 이상 리뷰 비율	Comb_5	전체 리뷰 중 '복합형_5' 리뷰 비율
Director	전체 리뷰 중 '감독' 리뷰 비율	Comb_7	전체 리뷰 중 '복합형_7' 리뷰 비율
Actor	전체 리뷰 중 '배우' 리뷰 비율	Comb_8	전체 리뷰 중 '복합형_8' 리뷰 비율
Story	전체 리뷰 중 '스토리' 리뷰 비율	Comb_9	전체 리뷰 중 '복합형_9' 리뷰 비율
Effect	전체 리뷰 중 '효과' 리뷰 비율	Comb_10	전체 리뷰 중 '복합형_10' 리뷰 비율

요인분석 결과, '배우 중심 평가 요인'은 온라인 리뷰를 작성 시 배우의 연기력 및 명성 등을 중심으로 작성된 온라인 리뷰를 의미한다. 복합형_4, 복합형_8, 복합형_9는 모두 '배우' 관련 평가요소와 기타 요소들을 복합적으로 평가하는 리뷰들이다. '감독 중심 평가 요인'은 'Director', 'Comb_1', 'Comb_5'로 묶였다. 특히, 복합형_1과 복합형_5는 감독뿐만 아니라 배우에 대한 평가가 복합적으로 이뤄지고 있는 리뷰이기 때문에, 다소 해석 상 논란의 여지가 있을 수 있다. 하지만, 이들 복합형 리뷰는 감독의 연출력과 배우의 연기력의 시너지 효과를 주로 언급하므로, 감독 연출력을 중심으로 한 평가로 생각할 수 있다. 마지막 '스토리 및 효과 평가 요인'은 'Effect', 'Story', 'Comb_10'로 구성된 요인으로, 영화의 시나리오 및 구성 요소에 관련된 요인으로 판단된다. 본 연구에서는 요인분석을 통해 묶인 변수들의 합으로 변수들을 통합하였다.

다음 <표 11>은 요인분석을 통한 요인들과 추가된 세 변수(ln_review, ln_screen, Positive) 간의 상관관계 분석한 결과이다. 상관관계 분석 결과, '개봉 첫 주 총 온라인 리뷰 수'와 '개봉 첫 주 누적 상영관 수'간의 상관관계가 높음을 알 수 있다. 이에 본 연구에서는 예측모형 분석 시, '개봉 첫 주 총 온라인 리뷰 수'와 '개봉 첫 주 누적 상영관 수'의 상호작용항을 추가하여 분석을 진행하였다.

5.2 판별분석을 통한 예측모형

5.2.1 판별변수 기초통계량 분석

다음 <표 12>는 각 등급별 판별변수들의 평균과 일변량 분산분석 결과를 제시하고 있다. 이를 통해, 6개 판별변수 모두 등급 간 유의한 차이가 있음을 확인할 수 있다. 각 등급 별 판별변수 간 차이를 보면, 하위 등급은 3등급과 4등급의 영화에서 '감독중심 평가'와 '스토리 및 효과 평가'가 상위 등급(1등급, 2등급)에 비해 높음을 알 수 있다. 이와 달리, 상위 등급에 해당하는 영화들은 '배우 중심 평가'가 높으며, 전반적으로 긍정적 리뷰의 비율과 전체 리뷰의 수가 높음을 확인할 수 있다.

이는 현재 상위 흥행 성적의 영화들은 배우에 대한 긍정적 평가가 높음을 알 수 있으며, 더 나아가 개봉 첫 주 동안에는 감독과 스토리에 대한 평가보다는 배우에 대한 평가 비율이 높을수록 상위 흥행 성적을 얻을 수 있음을 유추 할 수 있다.

〈표 11〉 상관관계 분석 결과

	배우 중심 평가	감독 중심 평가	스토리 및 효과 평가	Positive	ln_review	ln_screen	평균 (STD.)
배우중심 평가	1						0.149
							(0.093)
감독중심 평가	0.00596	1					0.034
	0.9316						(0.042)
스토리 및 효과 평가	-0.1683	-0.07712	1				0.177
	0.0146	0.2659					(0.064)
Positive	0.01746	-0.01424	0.09855	1			0.643
	0.8014	0.8375	0.1547				(0.143)
ln_review	0.20721	-0.02019	0.10516	0.23027	1		6.987
	0.0026	0.7717	0.1297	0.0008			(1.309)
ln_screen	0.26524	-0.10699	-0.02853	0.09628	0.76943	1	5.930
	< .0001	0.1222	0.6811	0.1645	< .0001		(0.532)

〈표 12〉 영화 등급 별 평균 비교

변수	1등급	2등급	3등급	4등급	Pr > F
배우중심 평가	0.189	0.159	0.145	0.114	0.048
감독중심 평가	0.038	0.030	0.032	0.053	0.046
스토리 및 효과 평가	0.157	0.161	0.187	0.181	0.032
Positive	0.744	0.655	0.619	0.646	0.002
ln_review	8.430	7.958	6.664	5.124	<.001
ln_screen	6.577	6.332	5.814	4.990	<.001
ln_review×ln_screen	55.56	50.52	38.95	25.57	<.001

주) *상호작용항의 평균은 41.97, 표준편차는 10.80.

5.2.2 다변량 판별분석 예측결과

영화의 흥행성과 예측모형을 개발하기 위해 다변량 판별분석(Multivariate Discriminant Analysis)을 수행하였다. 다변량 판별분석은 종속변수가 범주형 변수이고 독립변수가 연속형 변수일 때, 선험적으로 정의된 두 개 이상의 집단들을 판별할 수 있는 둘 이상의 독립변수의 선형조합을 찾아내는 분석기법이다(강현철 외, 2011).

본 연구에서는 판별분석의 타당성을 검증하기 위해, 10겹 교차 검증법을 사용하였다. 학습 데이터를 통해 도출된 판별함수의 카이제곱(Chi-square)값은 148.86, 유의확률이 < 0.0001이므로 판별함수의 유의성을 확인할 수 있다. 다음 식 (1)부터 식 (3)까지는 판별분석을 통해 도출된 3개의 선형 판별함수이다. 또한, 각 판별함수의 유의확률도 0.05 미만이므로, 3개의 판별함수 모두 유의한 수준의 판별력을 가지고 있음을 확인하였다.

$$\text{판별함수 } 1(v_1) = -0.142z_1 - 0.066z_2 - 0.407z_3 + 0.043z_4 + 2.823z_5 + 2.320z_6 - 3.181z_7 \quad (1)$$

$$\text{판별함수 } 2(v_2) = 0.019z_1 - 0.019z_2 - 0.164z_3 + 0.454z_4 - 11.527z_5 - 5.868z_6 + 16.568z_7 \quad (2)$$

$$\text{판별함수 } 3(v_3) = 0.50z_1 + 0.255z_2 + 0.577z_3 + 0.647z_4 - 0.239z_5 + 1.136z_6 - 1.065z_7 \quad (3)$$

위 결과에서 표준화 된 계수를 나타내며, 계수의 절대값 크기를 이용하여 각 변수가 판별함수에 기여하는 정도를 평가할 수 있다(강현철 외, 2011)[4]. 즉, 판별함수 1에서는 '개봉 첫 주 누적 상영관 수(ln_screen)'와 '개봉 첫 주 총 온라인 리뷰 수(ln_review)', '스토리 및 효과 평가'가 다른 변수에 비해 판별 기여도가 크다고 할 수 있으며, 판별함수 2에서는 '개봉 첫 주 긍정적 리뷰 비율(Positive)'와 '개봉 첫 주 총 온라인 리뷰 수(ln_review)', '개봉 첫 주 누적 상영관 수(ln_screen)'의 판별 기여도가 높음을 알 수 있다. 마지막으로, 판별함수 3에서는 2개의 판별함수에 비해 온라인 리뷰 마이닝 결과를 통해 계산된 세 변수(배우 중심 평가, 감독 중심 평가, 스토리 및 효과 평가)의 판별기여도가 높아졌음을 확인 할 수 있다.

3개의 판별함수 모두 '개봉 첫 주 누적 상영관 수(ln_screen)'의 판별기여도가 높으며, 이는 영화 관람객 수와 영화 상영 횟수는 상관관계가 있다는 기존 연구 결과와 일치한다(Neelamegham and Jain, 1999).

'개봉 첫 주 총 온라인 리뷰 수'와 '개봉 첫 주 누적 상영관 수'의 상호작용 효과를 의미하는

4) x_i는 순서대로 배우 중심 평가, 감독 중심 평가, 스토리 및 효과 평가, Positive, ln_review, ln_screen, ln_review×ln_screen이며, $z_j = (x_j - \bar{x}_j)/s_j$이다.

'ln_review*ln_screen' 또한 판별함수에서 중요 변수로 작용하였다. Chen *et al.*(2004)는 제품의 판매량과 온라인 리뷰 간의 인과관계는 양방향으로 작용한다고 하였으며, 위에서 언급한 것과 같이 개봉 첫 주 상영 횟수는 관람객 수에 유의적인 영향을 작용한다. 다시 말해, 온라인 리뷰와 상영 횟수 간의 상호작용은 영화 관람객 수에 영향을 미친다고 할 수 있다.

다음 <표 13>는 각 등급 별 3개의 판별함수 결과를 정리한 것이다. 관객 수 상위 10%에 해당되는 영화들은 판별함수에서 모두 양수의 값이 계산되었으며, 상위 10%부터 50%에 해당하는 영화들은 판별함수 3에서만 음의 값이 나왔다.

위 결과에서 알 수 있듯이, 1등급과 2등급의 영화를 구분할 수 있는 기준은 판별함수 3에 의해서 제시되고 있으며, 3등급과 4등급 간의 차이는 판별함수 2와 3에 의해서 분류될 수 있다. 즉, 온라인 리뷰에 의한 변수들의 판별기여도가 높은 판별함수 3에 의해, 상위등급과 하위등급을 좀 더 세부적으로 분류할 수 있음을 확인할 수 있다.

본 연구에서는 3개의 판별함수로 구성된 예측

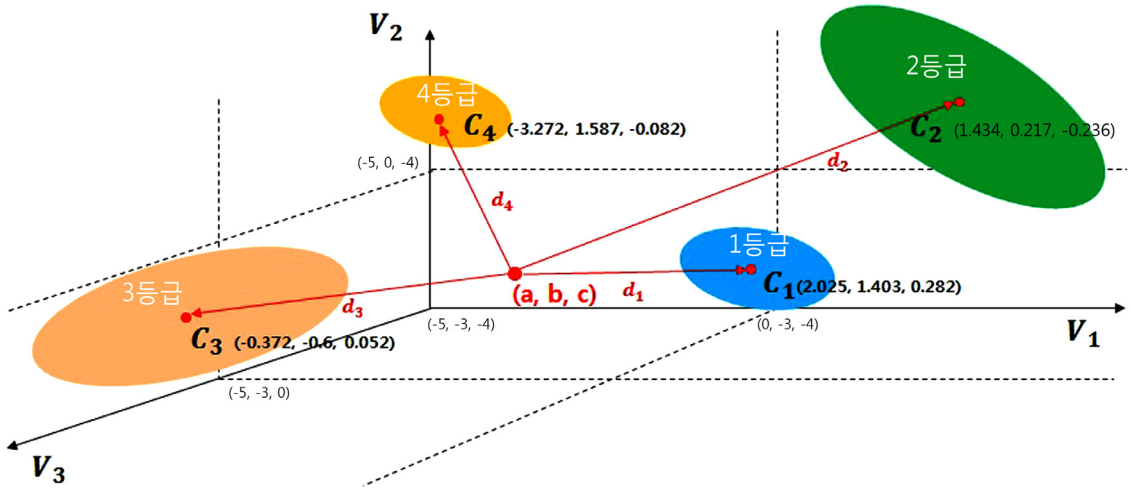
〈표 13〉 영화 등급 별 판별 결과

	판별함수 1	판별함수 2	판별함수 3
1등급	+	+	+
2등급	+	+	-
3등급	-	-	+
4등급	-	+	-

〈표 14〉 예측모형 검증 결과

	1	2	3	4	
오분류율	0.182	0.239	0.198	0.100	0.196

모형을 개발하였으며, 학습 데이터를 통해 각 그룹의 중심값을 계산하였다. 그 후 본 모형을 검증하기 위해 다음 <그림 2>와 같이, 각 영화 별 예측 모형에 의한 결과 값과 각 그룹 별 중심값과의 거리를 측정하여, 가장 가까운 거리를 갖는 그룹으로 해당 영화를 예측하였다. 예측한 결과는 다음 <표 14>와 같이, 10겹 교차 검증법을 통해 검증하였으며, 10회 평균 각 등급 별 및 전체 오분류율을 정리한 것이다. 예측 모델을 검증한 결과, 도출된 예측모형의 정확도는 약 82.02%라 할 수 있다.



〈그림 2〉 예측모형 적용 방법

Ⅵ. 연구 결과 및 의의

6.1 연구 결과 토의

본 연구는 개봉 첫 주 동안 작성된 온라인 리뷰를 통해 최종 영화 관객 수를 예측하였다. 이를 위해, 서포트 벡터 머신 알고리즘이 적용된 온라인 리뷰 마이닝을 수행하였으며, 흥행 등급을 구분하기 위한 3개의 판별함수를 제시하였다.

온라인 리뷰 마이닝을 진행하기 위한 온라인 리뷰 분류 기준을 기존 연구들을 바탕으로 '감독', '배우', '스토리', '효과' 등으로 정리하였으며, 현재 온라인 리뷰 작성 환경을 고려한 '복합형 리뷰'를 10가지 경우로 구분하였다. 온라인 리뷰 마이닝 결과, '복합형_3'(감독+스토리+효과)과 '복합형_2'(감독+배우+효과), '복합형_6'(감독+스토리)에 해당하는 리뷰가 각각 0개, 4개, 14개로 분류되어, 본 연구에서는 해당 분류 결과를 제거하고 연구를 진행하였다.

예측모형을 개발하기 위한 전 단계로 데이터 전처리를 통한 온라인 리뷰 마이닝 결과의 통합과 영화 관람객 수의 범주화를 진행하였다. 요인분석으로 온라인 리뷰 마이닝 결과를 '배우 중심 평가'와 '감독 중심 평가', '스토리 및 효과 평가'로 변수들을 통합하였으며, 영화 관람객 수를 상위 10%부터 하위 10%까지 4개의 등급으로 범주화하였다.

판별분석을 통해 도출된 3개의 판별함수에서는 공통적으로 '개봉 첫 주 누적 상영관 수'에 해당하는 변수의 기여도가 컸다. 이는 Dellarocas *et al.*(2007)의 연구 결과와 같이, 영화 상영 횟수와 영화 관람객 수 간의 상관관계가 있음을 확인할 수 있었다. 또한, '개봉 첫 주 누적 상영관 수'와 '개봉 첫 주 총 온라인 리뷰 수' 간에 상호작용이 있음을 확인하고, 판별분석에서도 중요 변수로 작용함을 알 수 있었다.

온라인 리뷰에 의한 변수인 '배우 중심 평가', '감독 중심 평가', '스토리 및 효과 평가'는 판별함수 3에서의 판별 기여도가 다른 2개의 판별함수에서 보다 높게 나왔으며, '배우 중심 평가'와 '스토리 및 효과 평가'가 '감독 중심 평가'보다 높은 판별 기여도가 있었다. 이러한 판별함수 3은 영화 흥행성과를 세부적으로 분류하는데 기여하였다. 다시 말해, 1등급에 해당하는 영화들은 2등급에 해당하는 영화에 비해, '감독', '배우' 및 '스토리 및 효과'에 대해서 긍정적으로 평가하는 리뷰가 많은 영화라고 할 수 있다. 더 나아가, '감독 중심 평가'보다 '배우 중심 평가'와 '스토리 및 효과 평가'가 높을수록, 1등급에 속할 확률이 높아진다. 예를 들어, '연가시(2012)'와 '감기(2013)'의 경우, 두 영화 모두 재난영화이며, 개봉 첫 주 총 온라인 리뷰 수는 약 5,000건, 긍정적 리뷰의 비율은 약 60%이다. 또한 개봉 첫 주 누적 상영관 수는 '감기'가 약 50개 정도 더 많았다. 하지만, 본 연구의 예측모형을 통해 분석한 결과, '연가시'와 '감기'가 1등급에 속할 확률이 각각 52.5%, 47.2%로 나왔다. 실제 온라인 리뷰 분석결과를 비교하면 '연가시'에서의 '배우 중심 평가' 비율이 '감기'보다 약 10% 정도 더 높음을 확인할 수 있었다.[5)]

이와 같이, 개봉 첫 주 동안 작성된 온라인 리뷰를 활용하여 최종 영화 관람객 규모를 예측할 수 있었으며, 관객 수가 많은 영화들은 '감독'과 관련된 평가보다 '배우' 및 '스토리', '효과'관련 평가가 많다는 것을 확인할 수 있었다.

6.2 연구의 학술적/실무적 의의

본 연구는 영화 개봉 후 작성된 온라인 리뷰를 활용한 최종 영화 흥행성과 예측모형을 제시하

5) '연가시'의 최종 관객 수는 4,515,833명이며, 2012년 국내 박스오피스 10위(해외영화 포함)이다. '감기'는 총 3,119,023명이 관람했으며, 2013년 박스오피스 17위를 기록했다. 실제 두 영화 모두 1등급에 포함되는 영화이며, 예측모형에서도 1등급으로 분류되었다.

였으며, 이러한 본 연구의 학술적 의의를 살펴보면 다음과 같다. 첫째, 온라인 리뷰 마이닝을 통한 온라인 리뷰 중심의 예측모형을 개발하였다. 기존 영화 수익 관련 연구에서의 온라인 리뷰는 양(Volume)과 의견 정도(Valence)에 집중하여 연구를 진하였다(Min and Lee 2005). 이와 달리, 본 연구에서는 15개의 분류기준을 정리하고, 온라인 리뷰 마이닝을 통해 각 분류 기준 별 분류 사전을 구축하여 온라인 리뷰를 분류하였다. 온라인 리뷰 마이닝 결과, '감독'에 관련된 온라인 리뷰는 감독의 연출력과 관련된 언급이 많으며, 배우와 관련해서는 배우의 연기력에 대한 평가가 많음을 확인 할 수 있었다.

둘째, '복합형 리뷰' 속성을 제시하고, 연구를 진행하였다. 기존 연구에서는 각 영화 리뷰의 속성을 한 개의 속성을 가진다고 가정하거나(김광수, 2000; 박형현, 박찬수, 2001), 감성분석을 통해 영화 속성과 감성 간의 관계를 파악하기 위한 연구를 진행하였다(Linton and Petrovich, 1988; Dellarocas, 2003). 하지만 실제 온라인 리뷰는 140글자까지 작성이 가능하며, 한 리뷰 내에서 '감독', '배우', '스토리', '효과' 등의 독립적 속성을 동시에 평가할 수 있다. 이에 본 연구에선 '복합형 리뷰'라는 분류 기준을 추가하고, 이를 10가지의 경우를 분류하여 분석하였다. 분석 결과, '복합형 리뷰'에서도 '배우와 스토리', '스토리와 효과'가 복합적으로 평가하는 리뷰가 주를 이루고 있음을 확인 할 수 있었다.

마지막으로 본 연구에서 제시한 15가지의 분류기준을 요인분석을 통해, 3가지 요인으로 정리하였다. 기존 연구들은 각 연구에 따라 온라인 리뷰 내 영화속성을 다르게 제시하고 있다(Dellarocas, 2003; Addi and Wlliams, 2010; 김광수, 2000; 박형현, 박찬수 2001). 이에 본 연구는 기존연구와 실제 온라인 리뷰 상황을 고려하여, 15가지의 분류기준을 제시하였으며, 요인분석을 통해 온라인 리뷰 분류 기준을 3가지로 축소하였다. 요인분석 결과, '배우', '배우+스토리+효과'(복합형_4), '배우+스토리'(복합형_8), '배우+효과'(복합형_9)가 하나의 요인으로 묶여 '배우 중심 평가'로 정리하였으며, '감독', '감독+배우+스토리'(복합형_1), '감독+배우'(복합형_5)를 '감독 중심 평가' 요인으로 하였다. 또한, '스토리', '효과', '스토리+효과'(복합형_10)가 한 요인으로 묶임으로써, '스토리 및 효과 평가'로 하였다. 이와 같이, 본 연구는 영화 관련 온라인 리뷰의 분류 방법 및 분류 기준을 제시하였다는 것에 학술적 의의를 두고 있다.

본 연구는 학술적 의의 외에도 다음과 같은 실무적 시사점을 갖는다. 첫째, 영화 제작단계에서의 감독 및 배우의 평가기준으로 활용될 수 있다. 현재 영화 제작단계에서 감독과 배우의 스타파워는 중요한 변수이다(Li *et al*., 2006). 이는 전작의 관객 수 및 현재까지의 연출 또는 주연급으로 캐스팅된 영화의 수 등으로 간접적으로 측정되고 있다. 하지만 본 연구에서 제시한 분류기준 및 온라인 리뷰 마이닝 방법을 적용한다면, 전작에서의 감독과 배우에 대한 일반관객들의 평가 비중을 측정할 수 있을 것이다.

둘째, 온라인 리뷰와 상영관 수 등 공개된 정보만을 활용하여 예측모형을 제시하고 있다. 영화가 개봉한 후 경쟁 영화와의 비교는 영화 수익 및 마케팅 전략에도 중요한 요인으로 활용되고 있다. 하지만, 기존 연구에서의 영화 수익 예측모형에서는 제작비와 같이 개봉 당시에는 공개되지 않는 변수가 활용되고 있다. 이와 달리, 본 연구에서는 온라인 리뷰를 중점적으로 사용하고 있으며, '누적 상영관 수'라는 공개된 정보를 추가적으로 사용하였다. 즉, 실제 개봉 영화간의 예상 수익 정도의 비교에 용이할 것이다.

셋째, 본 연구에서는 개봉 초기의 데이터를 활용하여 최종 영화 흥행성과를 예측하고 있다. 이는 개봉 첫 주 관객의 반응을 점검하고, 흥행성과를 향상시키기 위한 전략을 개발하는데 있어서 활용될 수 있을 것이라 판단된다. 판별함수 3을 통해 확인하였듯이, '감독' 관련 평가에 비해 '배우' 평가의 비중이 높을수록 상위등급으로

갈 확률이 높아진다. 즉, 개봉 초기에 흥행성과를 향상시키기 위한 조건을 탐색함에 있어서 본 연구 결과를 활용할 수 있을 것이다.

6.3 연구의 한계 및 향후 연구 방향

본 연구는 온라인 리뷰 마이닝을 통해 온라인 리뷰를 분류하고, 최종적으로 영화 흥행성과 예측모형을 개발하였다. 하지만, 다음과 같은 한계점을 가지고 있으며, 이는 향후 연구에서 보완되어야 할 것이다. 첫 번째로, 1차 학습 데이터를 만들기 위해 내용분석법을 사용하였다. 내용분석법은 다양한 장점에도 불구하고 연구자의 주관성을 완전히 배제되지 못 할 가능성을 가지고 있다(Eugene and Mary, 1993). 본 연구에서는 연구자가 개별적으로 약 3,000개 리뷰를 분류하고, 동일하게 분류된 결과만을 사용함으로써 가급적 주관성을 배제하기 위해 노력하였다. 하지만, 내용분석법의 근본적 한계로서 어느 정도 연구자의 주관성이 개입되었을 수도 있다. 향후 연구에서는 온라인 리뷰에 대한 객관성이 입증된 분류사전에 대한 연구가 필요하다.

두 번째로, 본 연구는 온라인 리뷰 마이닝을 하기 위해 '꼬꼬마 형태소 분석기'를 사용하였다. 해당 형태소 분석기는 띄어쓰기 등 오류에 덜 민감하다는 평가를 받고 있다. 하지만, 실제 온라인 리뷰에서는 여러 오・탈자와 신조어들이 존재한다. 이에 최초 수집한 긍정적 리뷰 307,385개 중 형태소 분석이 되지 않는 19,197개의 리뷰를 제거하고 온라인 리뷰 마이닝이 수행하였다. 따라서, 향후 연구에서는 온라인 환경에 적합한 형태소 분석기의 개발과 다양한 은어, 신조어 관련 단어 사전의 구축이 필요할 것이다.

세 번째로, 본 연구에서는 각 리뷰에 대한 감성을 각 리뷰 별 평점을 통해 분류하였으며, 그 중 8점 이상의 리뷰만을 대상으로 진행하였다. 하지만 8점 미만의 리뷰 또한 영화 관람객의 평가정보이며, 이들 또한 영화 흥행에 영향을 미칠 수 있다. 향후 연구에서는 중립적 리뷰와 부정적 리뷰에 대한 분류 기준을 정리하여 보다 포괄적인 온라인 리뷰 분석을 진행되어야 할 것이다.

네 번째는 데이터 수집 기한의 제한이다. Liu (2006)은 온라인 리뷰와 영화 매출액 간의 관계를 알아보기 위해 2002년 5월부터 9월까지 'Yahoo! Movies'에 작성된 12,136개의 온라인 리뷰와 40개의 영화를 활용하였다. 이는 기술적 한계로써, 데이터 양의 증가에 따라 계산 양이 기하급수적으로 증가하기 때문이었다. 본 연구에서는 온라인 리뷰 마이닝을 통해 307,385개의 온라인 리뷰를 분석하고, 209개의 영화를 대상으로 예측모형을 개발하였지만, 본 연구를 수행함에 있어서도 동일한 한계가 존재하였다. 그러나 이러한 문제는 하둡(Hadoop)으로 대표되는 빅데이터 플랫폼 및 컴퓨터 자원의 발전에 따라 해결 가능할 것이다.

마지막으로, 수집된 온라인 리뷰 중 관람 전 작성된 리뷰와 거짓 온라인 리뷰의 존재 가능성이다. 현재 여러 매체를 통해, '댓글 알바'에 대한 문제가 지적되고 있으며,[6] 리뷰 작성자가 실제 영화를 관람하고 평가를 했는지에 대한 문제가 있다. 이에 본 연구는 '감독', '배우', '효과', '스토리' 등의 속성에 대한 평가가 없이 작성된 '전반적 평가'를 분류하고, 이렇게 단순 추천이 이뤄진 리뷰는 예측모형에 개발단계에서 배제하였다. 하지만 이러한 방식은 임시적인 방법이며, 객관성이 부족하다고 판단된다. 따라서 실제 관객이 작성한 리뷰만을 분류할 수 있는 기준 및 방법에 대한 연구가 필요하다.

영화 산업에서의 수익 예측은 영화 상영과정에서 필수적인 전략이다. 본 연구의 예측모형은 다양한 마케팅 전략 및 영화 제작 단계에서 활용될 수 있을 것이며, 더 나아가 한국 영화산업 발전에도 기여할 수 있을 것으로 기대한다. 또한,

6) 이승재, "영화평론은 죽었다", 동아일보, 개제일: 2014. 03.27, http://news.donga.com/3/all/20140327/62034242/1.

현재 온라인 리뷰는 영화 산업 이외에 제품/서비스 이용 후기의 형태로 오픈마켓, 블로그, SNS 등에서 등장하고 있다. 즉, 본 연구에서는 영화 산업에 한정하여 온라인 리뷰 마이닝을 진행하였으나, 영화 산업뿐만이 아니라 다른 산업의(예: 전자제품 도소매, 도서 구매, 관광 여행 상품 등) 유형에 따라 판매 초기 정보를 활용한 최종 수익 규모 파악이 이뤄질 수 있을 것이다. 이를 위해서는 본 연구에서 마련한 분류기준 이외에 제품 속성에 맞는 분류기준이 마련되어야 할 것이며, 소비자 중심적인 마케팅 전략에 활용될 수 있을 것으로 기대한다.

참 고 문 헌

강문수, 백승희, 최영식, “맵리듀스를 이용한 통계적 접근의 감성 분류”, 감성과학, 제15권, 제4호, 2012, pp. 425-440.

강현철, 한상태, 김기영, 전명식, “예제로 배우는 SAS 다변량 자료 분석 입문”, 자유아카데미, 2011, pp. 203-234.

고정민, “미국영화와 한국영화의 흥행요인에 관한 비교연구-애국심 유발 요인을 중심으로”, 문화산업연구, 제10권, 제2호, 2010, pp. 71-96.

김광수, “영화 선택 및 평가에 관한 연구”, Korean Association for Advertising and Public Relations, 제48권, 2000, pp. 139-164.

박동권, 분산분석과 반복측정자료, 민영사, 2007, pp. 38-71.

박승현, 송현주, “영화 관객의 온라인 평가 분석: <박쥐>를 중심으로”, 언론과학연구, 제10권, 제4호, 2010, pp. 157-191.

박승현, 장정헌, “온라인 영화 리뷰의 내용과 품질에 관한 탐색적 연구: <마당을 나온 암탉>을 중심으로”, 언론과학연구, 제12권, 제4호, 2012, pp. 221-256.

박승현, 정완규, “한국 영화시장의 흥행결정 요인에 관한 연구: 2006~2008년 개봉작품을 중심으로”, 언론과학연구, 제9권, 제4호, 2009, pp. 243-276.

박형현, 박찬수, “영화 평론과 흥행성과 간의 관계-인터넷 시대에도 유효한가?”, 마케팅연구, 제16권, 제4호, 2001, pp. 71-85.

성영신, 박진영, 박은아, “온라인 구전정보가 영화 관람 의도에 미치는 영향”, 광고연구, 제57권, 2002, pp. 31-52.

양소영, 김형수, 김영걸, “온라인 고객 리뷰의 분류 항목별 차이 분석: 채널, 제품속성, 가격을 중심으로”, Asia Marketing Journal, 제10권, 제2호, 2008, pp. 125-151.

오은희, 전범수, “네티즌과 영화 평론가의 영화 평가 결정 요인 비교”, 한국방송학보, 제22권, 제6호, 2008, pp. 267-289.

이경재, 장우진, “베이지안 선택 모형을 이용한 영화흥행 예측”, 2006년 대한산업공학회 춘계학술대회 논문집, 2006, pp. 1428-1433.

이동주, 연종흠, 이상구, “한국어 문장의 띄어쓰기 오류 과정과 최적 형태소 분석을 위한 통합 확률 모델”, 한국정보과학회 2011한국컴퓨터종합학술대회 논문집, 제38권, 제1호(A), 2011, pp. 237-240.

이성직, 김한준, “TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법”, 한국전자거래학회지, 제14권, 제4호, 2009, pp. 59-73.

Addi, H. and L. J. Williams, “Principal component analysis”, *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol.2, No.4, 2010, pp. 433-459.

Alistair, K. and I. Diana, “Sentiment classification of movie reviews using contextual valence shifters”, *Computational Intelligence*, Vol.22, 2006, pp. 110-125.

Bausuroy, S., S. Chatterjee, and S. A. Ravid, “How critical are review? The box office effects of film critics, star, power, and budget”, *Journal of Marketing*, Vol.67, No.4, 2003, pp. 103-117.

Bo, P., L. Lee, and Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", *Proceedings of EMNLP 2002*, 2002, pp. 79-86.

Bowman, D. and Das Narayandas, "Managing Customer-Initiated Contacts with Manufacturers: The Impact on Share of Category Requirements and Word-of-Mouth Behavior", *Journal of Marketing Research*, Vol.38, No.3, 2001, pp. 281-297.

Brian, E., "Sentiment classification of movie reviews using linguistic parsing", 2006, (Available online at http://pages.cs.wisc.edu/~apirak/cs/cs838/eriksson_final.pdf).

d'Astous, A. and N. Touil, "Consumer Evaluations of Movies on the Basis of Critics' Judgments", *Psychology and Marketing*, Vol.16, No.8, 1999, pp. 677-694.

De Vany, A. and D. Walls, "Bose-Einstein dynamics and adaptive contraction in the motion picture industry", *The Economic Journal*, Vol.106, 1996, pp. 1493-1524.

Dellarocas, C., "The digitization of word of mouth: promise and challenges of online feedback mechanisms", *Management Science*, Vol.49, No.10, 2003, pp. 1407-1424.

Dellarocas, C., xiaoquan(michael) zhang, and neveen f. Awad, "Exploring the value of online product reviews in forecasting sales: the case of motion pictures", *Journal of interactive marketing*, Vol. 21, No.4, 2007, pp. 23-45.

Duan, W., B. Gu, and A. Whinston, "Do Online Reviews Matter? An Empirical Investigation of Panel Data", working paper, Department of Management Science and Information Systems, University of Texas at Austin, 2005.

Elberse, A. and J. Eliashberg, "Demand and Supply Dynamics for Sequentially Released Products in International Markets: The Case of Motion Pictures", *Marketing Science*, Vol.22, No.3, 2003, pp. 329-354.

Eliashberg, J. and S. Shugan, "Film Critics: Influencers or Predictors?", *Journal of Marketing*, Vol. 61, 1997, pp. 68-78.

Eugene, W. A. and W. S. Mary, "The Antecedents and Consequences of Customer Satisfaction for Firms", *Marketing Science*, Vol.12, No.2, 1993, pp. 125-143.

Hirschman, E. and A. Pieros, "Relationships Among Indicators of Success in Broadway Plays and Motion Pictures", *Journal of Cultural Economics*, Vol.9, 1985, pp. 35-63.

Jo, T. and N. Japkowicz, "Class Imbalances versus Small Disjuncts", *ACM SIGKDD Exploration*, Vol.6, 2004, pp. 40-49.

Lehmann, D. R. and C. B. Weinberg, "Sales through Sequential Distribution Channels: An Application to Movies and Videos", *Journal of Marketing*, Vol.64, No.3, 2000, pp. 18-33.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu, "Movie Review Mining and Summarization", *CIKM '06 Proceedings of the 15th ACM international conference on Information and knowledge management*, 2006, pp. 43-50.

Linton, J. and J. Petrovich, "The application of the consumer information acquisition approach to movie selection: An exploratory study", *In B.A. Austin(Ed.), Current research in film: Audiences, economics, and law*, Vol.4, 1988, pp. 24-44, Norwood, NJ: Ablex.

Litman, B. and L. Kohl, "Predicting Financial Success of Motion Pictures: The Early '80s experience", *Journal of Media Economics*, Vol.2, 1989, pp. 35-50.

Liu, Y., "Word-of-mouth for movies: Its dynamics and impact on box office revenue", *Journal of Marketing*, Vol.70, 2006, pp. 74-89.

Marsha L. Richins, "Negative Word-of-Mouth by Dissatisfied Consumers: A Pilot Study", *Journal of Marketing*, Vol.47, No.1, 1983, p. 68.

Min, J. H. and Y. C. Lee, "Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernel Function Parameters", *Expert Systems with Applications*, Vol.28, 2005, pp. 603-614.

Neelamegham, R. and D. Jain, "Consumer choice process for experience goods: An econometric model and analysis", *Journal of Marketing Research*, Vol.36, 1999, pp. 373-386.

Prag, J. and J. Casavant, "An Empirical Study of Determinants of Revenues and Marketing Expenditures in the Motion Picture Industry", *Journal of Cultural Economics*, Vol.18, 1994, pp. 217-235.

Reddy K. R., Y. Wang, W. F. DeBusk, M. M. Fisher, and S. Newman, "Forms of soil phosphorus in selected hydrologic units of the Florida Everglades", *Soil Science Society of America Journal*, Vol.62, 1998, pp. 1134-1147.

Vapnik, V., "The Nature of Statistical Learning Theory", Chapter 5, Springer-Verlag, New York, 1995.

Vasileios, H. and Kathleen R. McKeown, "Predicting the Semantic Orientation of Adjectives", ACL '98 Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 1997, pp. 174-181.

Wallace, W. T., A. Seigerman, and M. B. Holbrook, "The Role of actors and actresses in the success of films: How much is a movie star worth?" *Journal of Cultural Economics*, Vol.17, No.1, 1993, pp. 1-27.

Xiaojin, Z., "Semi-Supervised Learning with Graphs", *Language Technologies Institute*, CMU-LTI-05-192, School of Computer Science, Carnegie Mellon University, 2005.

Information Systems Review
Volume 16 Number 3
December 2014

Predicting Movie Revenue by Online Review Mining: Using the Opening Week Online Review

Seung Yeon Cho* · Hyun-Koo Kim* · Beomsoo Kim* · Hee-Woong Kim**

Abstract

Since a movie is an experience goods, purchase can be decided upon preliminary information and evaluation. There are ongoing researches on what impact online reviews might have on movie revenues. Whereas research in the past was focused on the effect of online reviews. The influence of online reviews appears to be significant in products like a movie because it is difficult to evaluate the feature prior to "consuming" the product. Since an online review is regarded to be objective, consumers find it more trustworthy. Contrary to prior research focused on movie review ratings and volume, we focus moves on movie features related specific reviews. This research proposes a predictive model for movie revenue generation. We decided 15 criteria to classify movie features collected from online reviews through the online review mining and made up feature keyword list each criterion. In addition, we performed data preprocessing and dimensional reduction for data mining through factor analysis. We suggest the movie revenue predictive model is tested using discriminant analysis. Following the discriminant analysis, we found that online review factors can be used to predict movie popularity and revenue stream. We also expect using this predictive model, marketers and strategic decision makers can allocate their resources in more parsimonious fashion.

Keywords: Online Review Mining, Discriminant Analysis, Prediction Model.

* Graduate School of Information, Yonsei University
** Corresponding Author, Graduate School of Information, Yonsei University

◐ 저 자 소 개 ◑

조 승 연 (infostat_24@yonsei.ac.kr)

현재 연세대학교 정보대학원에서 석사과정 재학 중이며, 연세대학교 정보통계학과를 졸업했다. 주요 연구분야는 Big data analytics, Security intelligence 등이다. Asia-Pacific Decision Sciences Institute Conference (APDSI) 2014에서 논문을 발표하였다.

김 현 구 (newredmoon@yonsei.ac.kr)

현재 연세대학교 정보대학원에서 석사과정 재학 중이며, 충북대학교 컴퓨터 공학과를 졸업했다. 주요 연구분야는 Computer security, IT management 등이다.

김 범 수 (beomsoo@yonsei.ac.kr)

현재 연세대학교 정보대학원에서 부원장으로 재직 중이다. 주요 연구분야는 정보보호정책 및 제도, 프라이버시 권리, 전자상거래, 정보경제학 등이다. International Journal of Information Management, Electronic Commerce Research and Applications, GLOBAL ECONOMIC REVIEW, Information Systems Review (ISR), Journal of Information Technology, Decision Support Systems 등에 논문이 게재되었다.

김 희 웅 (kimhw@yonsei.ac.kr)

National University of Singapore에서 근무한 후, 현재 연세대학교 정보대학원에 재직 중이다. 연세대학교 언더우드 특훈 교수(Underwood Distinguished Professor)로 선정되었다. 주요 연구분야는 소셜미디어, 디지털 비즈니스, 정보시스템관리 및 활용이다. IEEE Transactions on Engineering Management (IEEE TEM), Information Systems Research (ISR), MIS Quarterly 등에 논문이 게재되었다.

논문접수일 : 2014년 07월 28일 **게재확정일 : 2014년 11월 25일**
1차 수정일 : 2014년 09월 30일 **2차 수정일 : 2014년 11월 10일**