J. Korea Saf. Manag. Sci. Vol. 19 No. 4 December 2017 ISSN 1229-6783(Print) 169
http://dx.doi.org/10.12812/ksms.2017.19.4.169 ISSN 2288-1484(Online)

개봉 전후 트윗 개수의 증감률과 영화 매출간의 상관관계

박 지 윤* · 유 인 혁* · 강 성 우*
*인하대학교 산업공학과

A Study of Correlation Analysis between Increase / Decrease Rate of Tweets Before and After Opening and a Box Office Gross

Ji-Yun Park* · In-Hyeok Yoo* · Sung-Woo Kang *
*Department of Industrial Engineering, INHA University

Abstract

Predicting a box office gross in the film industry is an important goal. Many works have analyzed the elements of a film making. Previous studies have suggested several methods for predicting box office such as a model for distinguishing people's reactions by using a sentiment analysis, a study on the period of influence of word-of-mouth effect through SNS. These works discover that a word of mouth (WOM) effect through SNS influences customers' choice of movies. Therefore, this study analyzes correlations between a box office gross and a ratio of people reaction to a certain movie by extracting their feedback on the film from before and after of the film opening. In this work, people's reactions to the movie are categorized into positive, neutral, and negative opinions by employing sentiment analysis. In order to proceed the research analyses in this work, North American tweets are collected between March 2011 and August 2012. There is no correlation for each analysis that has been conducted in this work, hereby rate of tweets before and after opening of movies does not have relationship between a box office gross.

Keywords: Sentiment analysis, Box office prediction factors, Social Network Service, Twitter, Word of Mouth

1. 서 론

본 연구의 주된 목적은 영화 개봉 전 사람들의 반응에 비해 개봉 후 사람들의 반응 변화의 비율이 매출과의 상관관계가 있는지 파악하는 데 있다. 또한, 본 연구는 영화 흥행 성과의 대부분이 개봉 후 3주 이내에 결정된다는 연구 결과에 따라 (Kwon, 2014) 사람들의 반응 변화 비율이 이를 따르는지를 파악한다. 단순한 변화가 아닌 변화 비율을 파악하면, 그 변화 추이를 파악할 수 있다. 영화 반응의 변화 추이가 매출과 상관성이 있다면, 그 반응 변화의 경향을 파악함으로써 이후 마케팅 전략을 바꿔야 할지에 대한 결정에 도움이 될 것이다.

영화 산업에서 대부분 제작회사, 투자회사뿐만 아니라 배급사의 같은 목표는 영화 흥행이며, 그 영향 요소로 내적 요인으로는 감독, 배우, 관람 등급이, 외적 요인으로는 스크린 수, 배급사 파워, 소셜미디어 등이 영화 관객을 유인하는 요인이다 (Kim & Hong, 2011).

† This paper was supported by Inha University
† Corresponding Author : Sung-Woo Kang, Industrial Engineering, INHA UNIVERSITY, 100, inha-ro, Nam-gu, Incheon, M·P : 010-6343-9721, E-mail: kangsungwoo@inha.ac.kr
Received April 20, 2015; Revision Received May 11, 2015; Accepted June 11, 2015.

신상품과 같이, 과거 데이터가 존재하지 않는 영화는 매출을 예측하는 것이 중요한 과제이다. 그렇기에 많은 연구에서 영화 매출에 영향을 미치는 요인과 흥행을 예측하는 모델을 개발한다. 그 중, '구전 효과 (WOM: Word of Mouth)' 는 사람들이 영화를 선택하는데 영향을 미치므로 영화의 흥행을 결정한다 (Austin, 1989; Bayus, 1985, Faber & O`Guinn, 1984). 또한, 'Facebook', 'Instagram', 'Twitter' 와 같은 많은 소셜 네트워크 서비스(SNS: Social Networking Service)와 스마트폰의 발달로 2015년의 소셜 미디어 이용량이 2014년에 비해 평균 3.2% 정도 증가하였으며, 그에 따라 자기 생각을 간편하게 표현 가능해졌다 (Lee & Kim & Jung & Jamg & Kim, 2011).

본 연구에서는 SNS 데이터를 이용하여 단순한 반응이 아닌 개봉 전후의 반응 변화를 나타내는 증감률과 영화 매출과의 상관성을 확인한다. 그 중, 세계에서 가장 인기 있는 SNS 중 하나인 '트위터' 를 이용하여 특정 영화와 관련된 사람들의 의견이 그 영화 흥행에 얼마나 영향을 미치는가를 파악한다. 이때, 영화 흥행이 개봉 후 3주 안에 결정된다는 연구 결과에 따라 개봉 후 3주라는 기간과 개봉 전의 트윗 증감률이 흥행과 관련이 있는지를 먼저 파악하였다. 본 논문에서 파악한 문제는 다음과 같다.

연구 문제 1. 개봉 전과 개봉 후 3주까지 트윗 개수의 증감률과 영화의 매출이 상관관계가 있는가?
연구 문제 2. 개봉 전후 트윗 개수의 증감률과 영화의 매출이 상관관계가 있는가?
연구 문제 3. 긍정적 혹은 부정적인 트윗 개수의 증감률과 영화의 매출이 상관관계가 있는가?

본 논문에서는 미국 내 흥행 수입과 감성 분석(sentiment analysis)을 통해 특정 영화와 관련된 트윗 개수의 증감률과 매출과의 상관관계를 탐색한다. 언급한 연구 문제1과 2를 확인해보고, 이를 확장해 전체적인 반응만이 아닌 각 개인의 감정이 긍정적인지, 부정적인지를 판별하여 이런 반응들의 변화가 영화 매출과 상관관계가 있는지를 확인하고자 하는 연구 문제3을 진행한다. 이 연구는 2011년 3월부터 2012년 8월 사이에 수집된 북미 지역 트윗을 수집하여 그 중 무작위로 선정된 10편의 영화와 관련된 트윗을 추출해 개봉 전후의 증감률과 매출의 관계를 분석하고 그 관계의 의미를 파악한다.

다음 목차에서는 사전에 연구된 감성 분석(sentiment analysis)과 영화 예측과 관련된 방법론에 대하여 설명한다. 그리고 차례로, 이 논문의 방법론과 가설 검정 결과를 통한 결론을 다루며 이 논문은 마친다.

2. 사전 연구

이 분야에서는 기존의 박스 오피스를 예측하는 모델과 감성 분석 방법을 사용하여 박스 오피스 매출을 예측하는 최근 연구들, SNS 구전 효과가 영화에 영향을 미치는 기간에 관한 연구들을 다루고자 한다.

2.1 박스 오피스 예측 (Box office prediction)

2016년 전 세계 영화산업 총 매출은 전년 대비 1% 증가한 $386억을 기록하였으며, 북미지역은 전년 대비 2% 증가한 $114억을 기록하였다 (Motion Picture Association of America, 2016). 이 중 매출 상위 25위까지 차지한 영화의 매출이 전체 북미지역 매출의 53%를 차지하였다 (Motion Picture Association of America, 2016). 이렇듯 영화 산업에서 성공한 영화와 그렇지 않은 영화 매출 간의 차이는 아주 크다. 또한, 영화 사업은 위험하고 불확실한 사업으로 영화의 재정적인 성공 또는 실패를 예측하기 어렵다 (De Vany & Walls, 1999; Elberse, Eliashberg, & Leenders, 2006; Ghiassi, Lio, & Moon, 2015; Hennig-Thurau, Houston, & Walsh, 2007; Zhang & Skiena, 2009). 재정적인 손실을 막기 위해 많은 연구가 영화 예측 모델을 제안했으며 영화 성공에 영향을 미치는 요인을 분석하였다.

Litman 외 1인 연구에서는 제작 예산 규모, 등장인물의 스타성, 장르, 속편 여부, 관람 등급 등과 같은 제작 관련 요인이 영화 흥행에 긍정적 영향을 미친다고 분석하였다 (Litman & Kohl, 1989). 하지만 실제로 제작과 관련된 요인이 흥행에 미치는 영향력이 크지 않다는 연구 결과들도 많다. Sochay 연구에서, 속편 여부가 흥행에 영향을 미치지 않는다고 하였으며 (Sochay, 1994), Pokorny 외 1인은 등장인물의 영향력은 제작비 규모에 따라 차이가 있다고 하였다 (Pokorny & Sedgwick, 1999). 또한, Prag 외 1인은 프린트와 광고 비용, 평론가의 비평과 같은 마케팅에 투입된 비용이 영화 흥행에 가장 큰 영향을 미치는 요인으로 분석하였으며, 이런 요인들은 영화 장르, 스타 출연 여부, 제작비 규모 등과 높은 관련성을 보인다고 하였다 (Prag & Casavant, 1994). 최근 김연형 외 1인 연구에서는 2010년 상영된 영화를 대상으로 영화 흥행 결정 요인을 파악하고 흥행 성과를 예측하였다. 그 결과, 영화 내적 요인으로는 감독, 배우, 관람

등급이, 영화 외적 요인으로는 스크린 수, 배급사 파워, 소셜미디어 등이 영화 관객을 유인하는 요인으로 분석되었다 (Kim & Hong, 2011).

이렇듯 영화 산업에서 흥행에 영향을 미치는 요인에 관한 많은 연구들이 존재한다. 하지만 De Vany 외 1인의 연구에 따르면, 영화에 대한 고객의 의견은 영화 제작 요소(배우, 감독 등)를 반영한 전통적인 예측 모델을 사용하기 어렵게 한다. 또한, 많은 연구에서 '구전 효과 (WOM: Word of Mouth)'는 사람들이 영화를 선택하는데 영향을 미치므로 영화의 흥행을 결정한다고 한다 (Aus-tin, 1989; Bayus, 1985, Faber & O`Guinn, 1984). 영화에 대한 리뷰는 영화 출시부터 폐막까지 기간에 대한 흥행과 상관이 있지만, 초기 판매와는 강한 상관관계가 없다 (Eliashberg & Shugan, 1997). 이 연구에서 WOM은 심지어 비평가보다 더 많은 흥행에 영향을 미친다는 사실이 밝혀졌다. WOM은 제품이나 서비스에 관심이 있는 개인이나 조직이 정보를 전달하기 때문에 새로운 제품이나 서비스를 구매할 때 가장 믿을 만한 정보 중 하나이다 (Katz & Lazarsfeld, 1955; Y. Liu, 2006). 영화 사업에서 '구전 효과 (WOM)'는 사람들이 영화를 선택하는데 영향을 미치므로 이는 영화 흥행을 결정하는 요소이다 (Austin, 1989; Bayus, 1985, Faber & O`Guinn, 1984). 또한, 웹 2.0이 개발되면서 온라인 구전효과 (eWOM: electronic word of mouth)를 통해 일반 사용자가 자기 생각과 의견을 공유하는 것이 가능해졌다 (Ravi & Ravi, 2015).

본 논문은 소셜 미디어에서 eWOM을 효율적으로 활용하는 데 초점을 둔다. 특히, 트위터의 실제 사용자가 만든 엄청난 문자 데이터(eWOM 등)를 다루며, 약 10억 개의 트윗으로부터 42,590개의 트윗을 추출하여 사용한다. 거대한 비정형 데이터에서 원하고자 하는 트윗을 검색하기 위해 영화의 이름과 해시태그를 사용하며, 분리된 트윗에 대한 정확성을 검증한다. 그런 다음 분석하고자 하는 트윗에 대해 감성 분석이 수행되고 그 트윗의 개봉 전후 증감률과 박스 오피스의 매출 간의 상관관계를 분석한다.

2.2 감성 분석 (sentiment analysis)을 이용한 박스 오피스 예측

최근 많은 분야에서 SNS를 통한 의견을 분석하고 이를 활용하고 있다. '감성 분석(sentiment analysis)은 '오피니언 마이닝(Opinion Mining)' 혹은 '감성 분류(Sentiment classification)'로도 불리며, 텍스트에 나타난 사람들의 태도, 의견, 성향과 같은 주관적인 데이터를 분석하는 자연어 처리 기술이다 (Chen & Zimbra, 2010, B Liu, 2012; Wiebe, 1994). 감성 분석의 주된 목적은 주어진 문서에서 의견, 표현을 분석하여 감성을 파악하는 것이다 (T;un Thura Thet, Na, & Khoo, 2010). SO(Semantic Orientation)는 주어진 문서의 감성을 결정하는 감성의 극성 (sentiment polarity)과 감성을 점수화한 감성 강도 (sentiment intensity)를 나타낸다 (Taboada, Brooke, Tofilo-ski, Voll, & Stede, 2011). 문서 안의 감성 극성을 측정하고 분석함으로써 마케팅을 포함한 다양한 분야에 도움될 것이다. 그 중, 최근에 SNS가 영화 흥행의 중요한 변수가 되면서 SNS의 비정형 텍스트 데이터 분석을 통해 영화 흥행의 추이를 예측하고 분석하는 연구들이 활발하게 이루어지고 있다 (Lee et al, 2014).

Mishne 외 1인은 영화 개봉 전 블로그에 게시된 글을 감성 분석을 함으로써 긍정적인 감정이 영화 흥행 요소에 영향을 미치는 요인임을 증명하였다 (Mishne & Glance, 2006). 또한, Asur 외 1인은 2009년부터 3개월 동안 미국에서 개봉된 24편의 영화에 대한 트윗의 감성을 분류하였다. 그 결과, 영화 개봉 전 트위터에서 언급된 횟수와 개봉 첫 주의 영화 수입 간의 상관관계가 있음을 파악하였다 (Asur & Huberman, 2010). Lica 외 1인 은 트윗의 감성 분석을 통해 30편의 영화에 대한 흥행을 예측하였다 (Lica & Tuta, 2011). 국내에서는 허민회 외 2명은 네이버의 영화 평점을 이용하여 관객의 긍정, 부정 표현과 영화 흥행의 상관성이 있음을 밝혔다 (Heo & Kang & Jo, 2013).

2.3 SNS 구전 효과가 영화에 영향을 미치는 기간

최근 많은 분야에서 SNS를 통한 의견을 분석하고 이를 활용하여 마케팅 전략을 내세우기 위해 연구한다. Litman 외 1인은 개봉 전에는 전문가의 비평이 더 중요하나, 개봉 이후에는 일반인들의 구전 효과가 영화를 선택할 때에 있어 더 많은 영향을 미친다고 한다 (Litman & Kohl, 1989). 실제로, 2000년 이후부터 전문가의 비평보다 SNS가 주요한 흥행 성과 요인으로 분석되었다 (Dellarocas, 2003). 이후 SNS를 통한 구전 효과가 영화에 미치는 영향은 크다는 사실을 파악하였고, 사람들은 그 영향력이 유효한 기간에 관해 관심을 두기 시작했다.

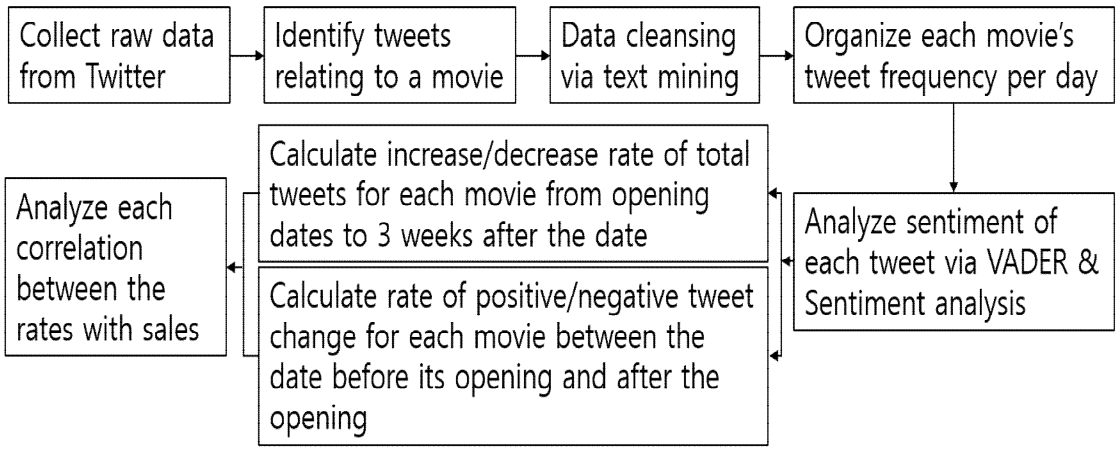
영화 전체 흥행성과의 대부분이 개봉 3주 안에 이루어진다는 연구에서 상업적 영화 마케팅이 개봉 전과 초기에 집중되어야 한다고 말했다 (Kwon, 2014). 또한, 다른 제품에 비해 흥행하지 못한 영화는 교체가 쉬우므로, 개봉 후 2~3주 안에 성공하지 못하면 즉시 다

른 영화로 교체가 된다고 한다 (Sawhney & Eliashberg, 1996, Swami, Eliash-berg, & Weinberg, 1999). 정재엽 외 1인의 연구에서 트위터에서 언급된 영화에 대한 호감도는 개봉 후 2~3주에 관람한 관람객들의 경우가 가장 높았다 (Jeong & Kim, 2016). 이 연구에서 개봉 후 2~3주 후에 관람한 사람들이 구전의 대상이 되는 영화에 가장 많은 호감을 느끼고 있으며 가장 많은 영향을 받고 있음을 보여준다. 또한, 초기에 관람하는 소비자들이 구전 확산 의도가 높은 만큼 영화의 초기 평가가 매우 중요하다는 점을 시사한다 (Jeong & Kim, 2016). 2015년 4월, CGV리서치 센터에서 발표한 빅데이터 분석을 통한 영화 흥행에 대한 보고서에서는 한국 영화 흥행의 결정적 요소가 구전효과 (WOM: Word Of Mouth)이며, 그중 SNS (Social Network Services)를 통한 구전 효과의 영향력이 중요하다고 밝혔다 (CGV research center, 2014). 이로써 SNS 구전 효과가 영화 흥행에 영향을 미치며, 개봉 후 약 3주 정도까지 크게 영화에 영향력이 있음을 파악할 수 있다.

3. 연구 방법 (Methodology)

본 연구는 트윗 개수의 증감률과 영화 매출 간의 상관관계를 파악한다. 또한, 사전 기반 감성 분석 (lexicon based sentiment analysis)을 사용하여 트윗이 긍정적인지 부정적인지를 판별한 후, 긍정적인 트윗 개수의 증감률과 부정적인 트윗 개수의 증감률 각각이 영화 매출과의 상관관계가 있는지를 확인한다. 이 연구의 전반적인 분석 과정은 다음 <Fig. 1>과 같다.

트윗을 수집한 후, 특정 영화에 따른 트윗을 구분하여 각 영화에 따라 데이터셋을 만든다. 각각의 데이터셋을 정제하고, 사전 기반 감성 분석(lexicon based sentiment analysis)을 이용하여 트윗별 강도를 부여한다. 이를 총 트윗, 긍정적인 트윗, 부정적인 트윗 각각의 경우에 대해서 날짜별 빈도수로 정리한 후, 개봉 전후의 증감률을 계산한다. 이때, 개봉일을 기준으로 개봉 전과 개봉 3주 후까지의 증감률, 개봉 전부터 폐막까지의 증감률을 따로 계산한다. 그 후, 각각의 증감률과 영화 매출 간의 상관관계가 있는지 파악한다.



[Figure 1] Overall process of the methodology in this work

3.1 데이터 전처리

3.1.1 원시 데이터로부터 목표 트윗 분리

본 연구는 사용자가 기분이나 감정을 표현한 트윗을 수집하는 것으로 시작한다. 대량의 트윗 데이터로부터 영화의 이름과 해시 태그(Hashtag)를 이용하여 영화와 관련된 트윗을 추출한다. 해시 태그(Hashtag)는 공식 트위터 계정에서 가져온다. 예를 들어, 영화 'The Girl with the Dragon Tattoo' 에 대해 아래와 같은 2개의 해시 태그를 얻을 수 있다.

#Thegirlwiththedragontattoo, #TGWTDT

이 두 해시 태그는 같은 영화를 가리킨다. 이 해시 태그를 사용하면 주어진 영화에 대해 분리된 트윗의 엔트로피를 줄일 수 있다. (Shannon, 2001).

그 후, 다음 작업에서 공식 영화 예고편이 YouTube에 업로드된 날짜, 즉 마케팅 시작일로부터 영화 종료 날짜까지의 데이터를 누적한다. 중복되는 트윗은 이 연구의 목표인 상관관계를 살펴보는 것에 문제가 될 수 있기에 비 중복 트윗만 수집한다. 트윗에는 각 트윗, 트윗이 작성된 날

짜, 텍스트 메시지와 해시 태그를 식별하는 사용자의 ID가 들어있다. 트윗의 예는 <Table 1>에 나와 있다.

3.1.2 분리된 트윗으로부터 샘플 데이터를 추출

영화의 이름이 일상적인 단어로 구성되거나 다른 개체를 참조하는 동일 어인 경우 대상이 아닌 트윗을 포함할 수 있다. 예를 들어, 본 논문의 사례 연구에서 'Real steel' 은 실제 강철을 나타내는 일반적인 단어이다. 따라서 분리된 데이터는 쓸모없는 의견으로 구성될 수 있다. 분리된 데이터에 쓸모없는 의견이 있는지 확인하기 위해 샘플 데이터를 추출하여 유효성 검사를 수행하고자 한다. 이 방법은 대용량 데이터를 처리하기 때문에 모든 트윗에 적용하지 않는다.

<Table 1> Example of collected tweets

Tweet_ID	Date	Text	Hashtags
146307841651769000	20111212	two best movies comin out #darkknightrises and #battleship	['darkknightrises', 'battleship']
149903000213786000	20111222	Which scared you more? 1973's #Exorcist or 1979's #Alien I am looking forward to #RidleyScott return and touch in #Prometheus	['Exorcist', 'Alien', 'RidleyScott', 'Prometheus']

3.1.3 텍스트 정제 과정

본 과정에서는 영화에 대해 감성 분석을 하기 앞서서 트윗에 대해 전처리 과정을 진행한다. 이 부분은 전체 텍스트 마이닝 (text mining) 방법에 대해 필수적이고 기초적인 과정이다.

첫 번째 단계에서는 트윗의 중요하지 않은 기능을 제거한다. 트윗에는 다른 문서와 도드라지는 특성이 있는데, 이는 주소, 해시 태그 및 외부 링크 (URL) 등이 있다. 이러한 특성은 본 논문에서 제시하는 감성 분석에 있어 의미있는 단어들이 아니므로, 텍스트에서 제거해야 한다. 속어, 약어 및 이모티콘은 감성 분석을 방해하는 또 다른 특성이다. 하지만 이들은 감성을 담고 있는 중요한 단어들이므로, 원래의 감성 강도를 반영할 수 있도록 점수로 대체한다.

그 후, 반복되는 문자 또한 처리한다. 어떤 트윗은 감정을 강조하기 위해 같은 단어를 반복하는데, 예를 들면 'good' 를 'goooooood' 로 표현하는 때도 있다. 정확한 감성을 평가하기 위해, 이렇게 적어도 세 번 이상 반복되는 단어는 올바른 형태로 고쳐야 한다.

3.2. 사전 데이터베이스 구축

때때로 일반적인 단어 사전은 제대로 작동하지 않을 수 있다. 일반적인 단어 사전은 단어 'horrible' 에 대해 부정적인 감정을 나타낼 수 있으나 본 논문의 사례연구에서 'Woman in black' 이라는 공포 영화에 적용할 때, 이 단어는 영화에 대해 긍정적인 감정을 내포한다고 볼 수 있다. 따라서 장르에 맞는 사전을 구축해야 영화에 대한 정확한 감성을 나타낼 수 있다.

3.3. 감성분석 (Sentiment analysis)

이 부분에서는 사전 기반의 감성 분석 (Lexicon-based sentiment analysis)을 통해 각 트윗의 감정을 결정한다. 사전 기반의 감성 분석은 단어를 사전에 결정된 어휘로 대응시킴으로써 문서 또는 문장 내의 각 단어에 대해 감성 정도를 계산한다. 그 후, 전체 감성 극성을 긍정, 부정 또는 중립으로 분류한다.

부정문은 감성 분석의 중요한 개념이다. 예를 들어, 'I don't recommend this movie.' 는 영화에 대한 부정적인 견해라고 판별되어야 한다. 본 문장이 부정문인 것을 고려하지 않으면, 문장은 긍정적인 의견으로 취급될 것이다. 따라서 이를 제대로 처리하는 것이 중요하다. 또한, 강조하는 단어도 중요한데, 예를 들어, 'I like this movie very much.' 은 'I like this movie.' 보다 긍정적이다. 따라서 이 또한 고려되어야 한다.

본 연구에서는 VADER 사전을 사용한다. VADER 사전은 소셜 미디어 서비스에 게시되는 이모티콘 또는 약어의 감성을 파악한다 (Hutto & Gilbert, 2014). 이 사전은 연구 방법론이 감성의 극성과 강도를 분석할 수 있게 한다. VADER는 -1 (매우 부정)에서 +1 (매우 긍정) 사이의 감성 점수를 반환한다. 본 논문은

VADER를 활용하여 트윗의 감성을 분석한다. 트윗에는 일반적으로 사용되는 이모티콘, 약어와 속어 같은 SNS의 특정 문자가 있다. VADER는 위의 특정 문자들에 대하여 상응하는 감성 점수를 부여한다.

3.4 상관 분석 (Correlation analysis)

상관 분석은 확률론과 통계학에서 두 변수 간에 어떤 선형적 관계를 맺고 있는지를 분석하는 방법이다. 상관관계의 정도를 파악하는 상관계수는 두 변수 간의 연관된 정도를 나타낼 뿐 인과관계를 설명하는 것은 아니다. 상관계수는 -1~1 사이의 값을 갖게 되는데, 상관계수의 절댓값이 1에 가까워질수록 선형적 상관관계가 있다고 이야기할 수 있다. 이 연구에서는 특정 영화를 언급한 개봉 전후의 트윗 개수 증감률이 영화의 총 매출과 어느 정도의 선형적 상관관계가 있는지를 확인하고자 상관 분석을 하였다.

이때, 상관분석에서는 기본적으로 '선형성', '동변량성', '두 변인의 정규분포성', '무선독립표본'을 가정하므로, 이 가정에 어긋나는지를 판별 후, 만약 정규분포를 만족한다면 피어슨 상관계수(Pearson correlation coefficient)를 사용하며, 정규분포를 만족하지 못하거나 두 변수 중 최소한 한 변수가 순위 변수인 경우, 혹은 표본 수가 적은 경우, 스피어만 상관계수(Spearman correlation coefficient)를 사용하여 두 변인이 선형적 상관관계가 있는지 없는지를 판별할 수 있다.

4. 실험

4.1 실험 절차

4.1.1 데이터 수집 및 정제

본 연구는 2011.03~2012. 09까지 북미 지역의 총 1,237,454,287개의 트윗을 수집한다. 트윗은 JSON (Ja-vaScript Object Notation) 형식으로 저장된다. JSON은 사용자가 쉽게 내용을 이해할 수 있도록 해주는 공개 표준 파일 형식이다. 트윗 데이터에는 사용자 위치, 프로필 색상 등과 같은 다양한 정보가 포함되어 있다. 효율적인 분석을 위해 관련성 없는 내용은 삭제하였다. 이 과정에서 생성 날짜, 해시 태그, ID와 게시글 내용은 유지된다. 위에서 밝힌 기간 내에 발표된 영화 중 무작위로 10개를 선정 후, 흥행 수입과 폐막일과 같은 영화 정보는 boxofficemojo.com에서 수집하였다. Boxofficemojo 는 가장 큰 온라인 박스 오피스 웹 사이트 중 하나이다. 영화 공식 예고편의 경우, youtube 공식 영화 계정에서 가져왔으며, 해시 태그는 영화사에서 공표한 공식 해시 태그를 기준으로 수집하였다. 실험을 위하여 무작위로 선정된 영화와 해당 공식 해시 태그는 <Table.2>에 표기되어 있다.

<Table 2> Selected movies and official hashtags

Movie	Hashtag
The Hunger Game	#TheHungerGames
Captain America	#CaptainAmerica, #FirstAvenger
The girl with the dragon tattoo	#Thegirlwiththedragontattoo, #TGWTDT
Battleship	#Battleship
real steel	#realsteel
Immortals	#Immortals
Prometheus	#prometheus
woman in black	#WomanInBlack
the iron lady	#theironlady
Tower Heist	#TowerHeist

해당 영화들의 트위터에 적혀 있는 내용은 VADER 사전을 이용하여 감성 분석을 하였다. 그 후, 감성 강도가 -1~-0.5는 부정, 0.5~1은 긍정으로 판별하였다. 이 데이터로부터 각 트윗의 날짜별 트윗 빈도수를 정리하였으며, <Table 3>은 영화 'Battleship'의 일별 트윗 언급 빈도수를 이 영화의 공식 영화 예고편부터 폐막일까지의 나타낸 데이터셋이다.

<Table 3> Frequency of tweets per day relating to 'Battleship'

Date	Freq
2011-12-12	18
2011-12-13	10
2011-12-14	5
2011-12-15	5
2011-12-16	6
...	...
2012-05-17	201
2012-05-18	595
2012-05-19	576
2012-05-20	485
...	...
2012-07-30	6
2012-07-31	12
2012-08-01	14
2012-08-02	12

<Table 4> Total dataset for 10 movies

Movie	Before	After	Change	Gross
Battleship	3031	4253	0.4031673	63135260
Captain America	391	1149	1.9386189	176654505
The girl with the dragon tattoo	79	316	3.0000000	102515793
The hunger games	5001	8949	0.7894421	408010692
Iron lady	3	35	10.6666667	30017992
Immortals	715	4191	4.8615385	83504017
Prometheus	2305	11073	3.8039046	126477084
Real steel	92	156	0.6956522	85468508
Tower heist	83	159	0.9156627	78046570
Woman in black	128	312	1.4375000	54333290

'Battleship' 은 2011년 12월 12일에 공식 영화 예고편을 공개하고, 2012년 5월 18일에 개봉하였다. <Table 3>을 살펴보면, 개봉일 즈음에 빈도수가 증가했음을 확인할 수 있으며, 폐막일인 2012년 8월 2일까지의 자료를 수집하였다.

또한, <Table 3>와 같은 방법으로 나머지 9편의 영화에 대하여 데이터를 수집하고 <Table 4>와 같은 총 데이터 셋을 완성하였다.

4.1.2 데이터 탐색

먼저, 총 데이터 10개에서 종속변수는 개봉 전후 트윗 개수의 증감률로, 목표변수는 총매출액으로 하여 그 상관관계를 알아보고자 한다. 총 데이터 수는 10개로, 같은 확률분포를 가진 독립확률변수 n개의 평균의 분포가 n이 적당히 크다면 정규분포에 가까워진다는 중심극한정리를 가정할 수 없으므로, 정규성 검정을 하여야 한다. 이론적으로 10개의 총 매출과 개봉 전후 총 트윗의 증감률이 각각 정규분포를 따르는지를 판별하기 위해 'Shapiro-Wilk nor-mality test' 를 사용하였다.

- H_0: 총 매출/ 개봉 전후 총 트윗의 증감률은 정규분포를 따른다
- H_1 : 총 매출/ 개봉 전후 총 트윗의 증감률은 정규분포를 따르지 않는다.

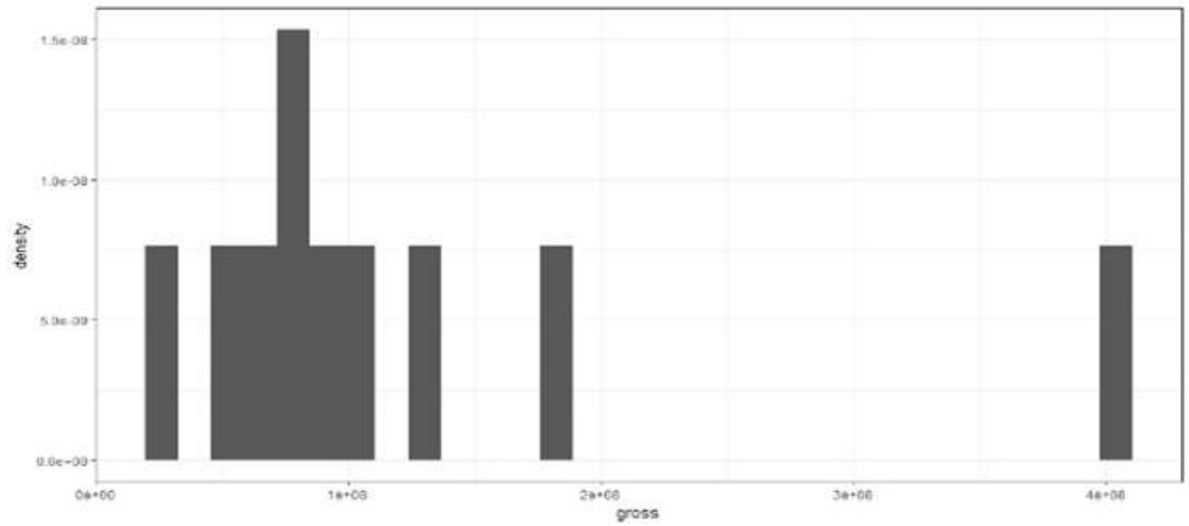
Shapiro-Wilk normality test결과, 매출과 개봉전후의 트윗개수 의 증감률은 각각 p-value가 0.000968, 0.005391으로 귀무가설을 기각하여 둘 다 정규분포를 따르지 않는다고 할 수있다. 이를 시각적으로 확인하기위해 [Figure 2]와 [Figure 3]에서 각각의 분포를 나타내었다.

4.2 가설 검정

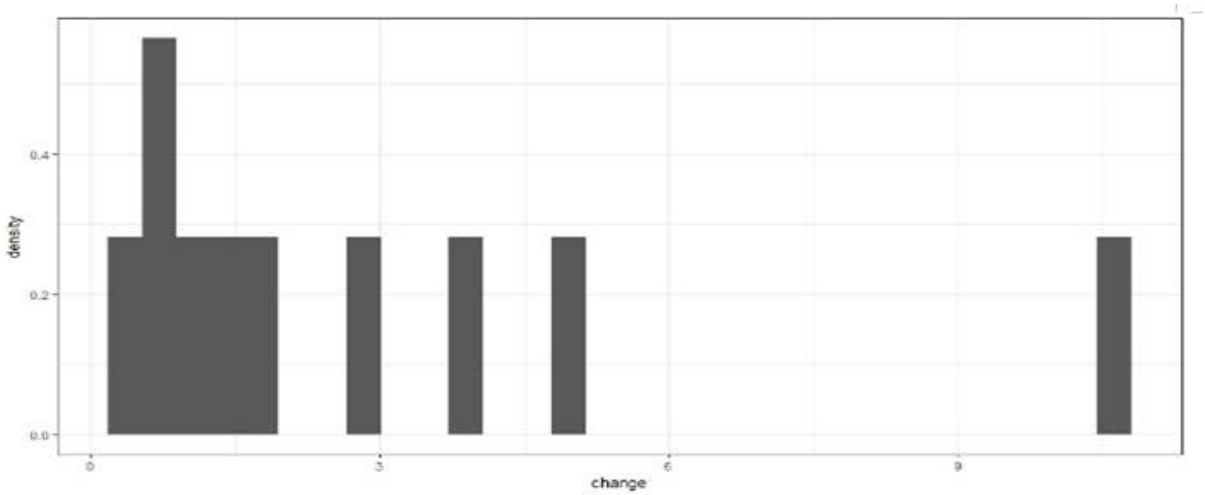
앞에서 확인한 결과에 의해, 개봉 전후 트윗 개수 증감률, 총 매출은 정규성을 만족하지 못하므로, 비모수통계 스피어만 순위상관계수(Spearman Rank Correlation Coefficient)를 이용하여 상관관계가 있는지 확인하였다. Spearman Rank Correlation Coefficient는 다음과 같은 식을 가지며, 주로 표본 집단이 정규성을 만족하지 못하는 경우, 혹은 목표 변수가 순위 형태의 값을 지닐 때 상관관계를 보기 위하여 사용된다.

$$r = 1 - \frac{6\sum(x_i - y_i)^2}{n(n^2-1)} \quad (1)$$

위에서, x는 종속변수, y는 목표 변수를 의미하며 n은 총 데이터 수를 의미한다.



[Figure 2] Distribution of total sales



[Figure 3] Increase / decrease rate of total tweets before and after opening

4.2.1 개봉 전과 개봉 후 3주까지 총 트윗 개수의 증감률

사전 연구에서 밝힌 영화 흥행에 개봉 후 3주까지가 가장 영향을 많이 끼친다는 연구를 고려하여 (Kwon, 2014) 개봉 전과 개봉 후 3주까지의 트윗 개수의 증감률이 총 매출 과의 상관성을 확인해보고자 한다. 이 때의 가설은 다음과 같다.

- H_0 : 개봉 전과 개봉 후 3주까지 트윗 개수의 증감률과 총 매출은 특정한 선형관계가 없다. (r=0)
- H_1 : 개봉 전과 개봉 후 3주까지 트윗 개수의 증감률과 총 매출은 특정한 선형관계가 있다. (r≠0)

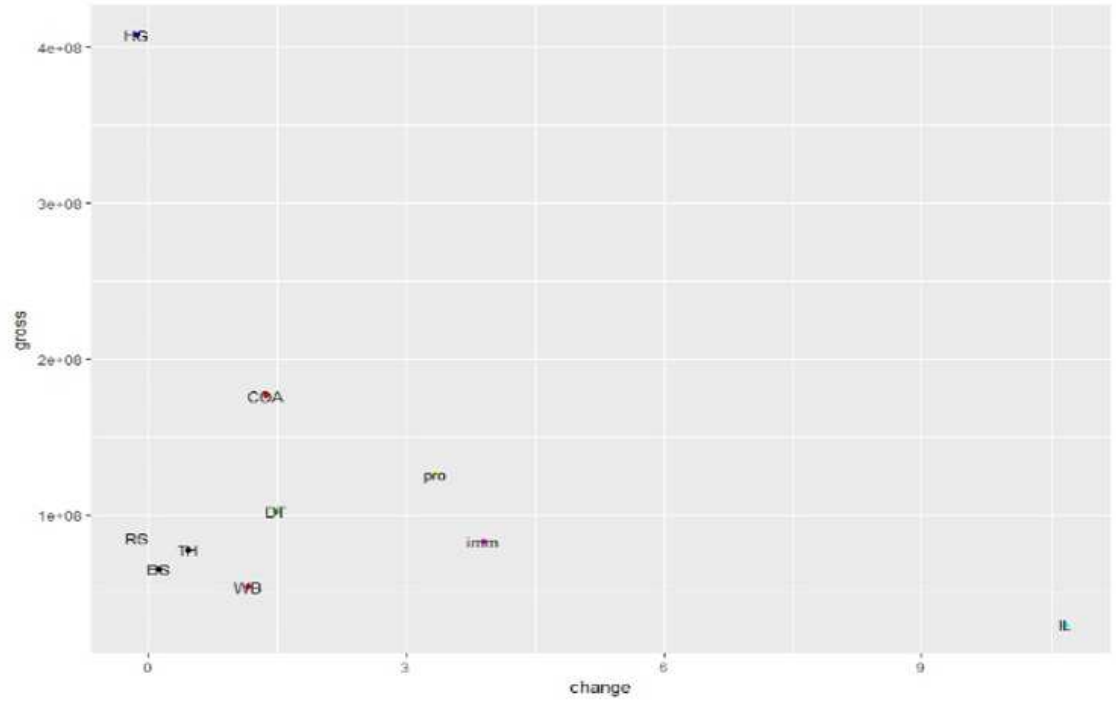
[Figure 4]를 참고하여 가설검정 결과를 살펴보면, p-value가 0.5139로 귀무가설을 기각하지 못하며, 스피어만 순위상관계수가 약 -0.2364로 선형적 상관관계가 없음을 확인할 수 있다.

4.2.2 개봉 전후 총 트윗 개수의 증감률

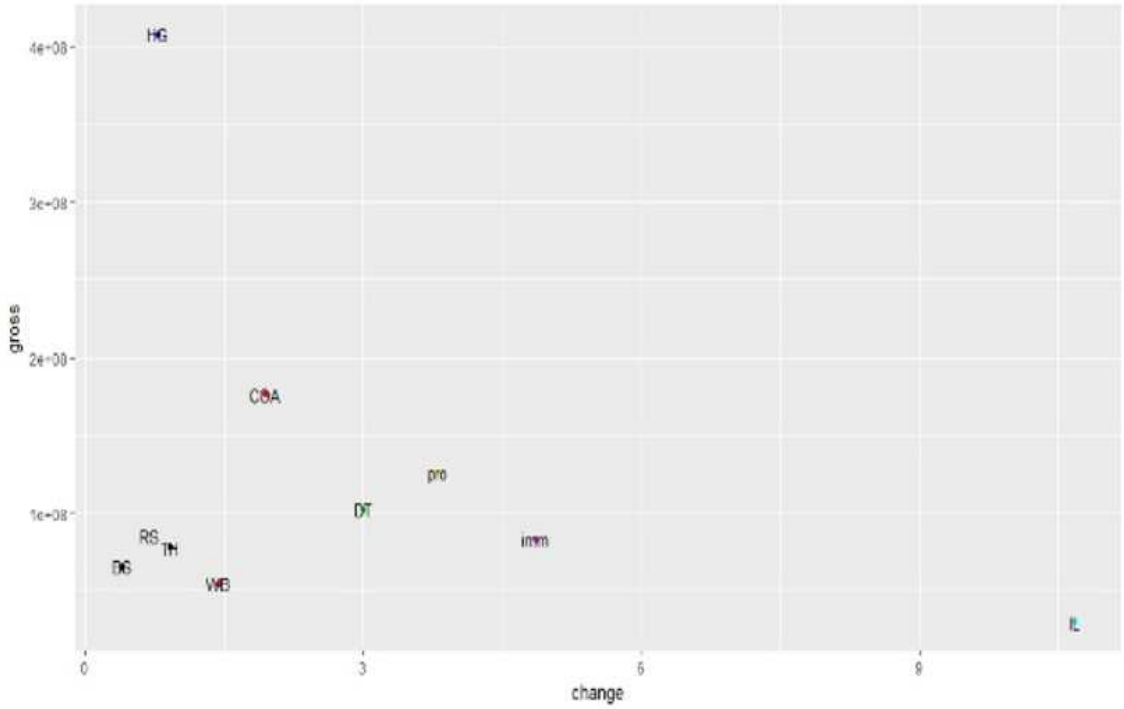
개봉 전후의 트윗 개수의 증감률과의 총 매출 간의 관계는 어떻게 나타날지를 확인하기 위해 그래프를 살펴보았고 이는 <Fig. 5>와 같다. 이 때의 가설은 다음과 같다.

- H_0 : 개봉 전후의 트윗 개수의 증감률과 총 매출은 특정한 선형관계가 없다. (r=0)
- H_1 : 개봉 전후의 트윗 개수의 증감률과 총 매출은 특정한 선형관계가 있다. (r≠0)

가설검정 결과를 살펴보면, p-value가 0.7588로 귀무가설을 기각하지 못하며, 스피어만 순위상관계수가 약 -0.1152로 선형적 상관관계가 없음을 확인할 수 있다.



[Figure 4] Distribution of total tweet growth rate and total sales before opening and after three weeks from the opening



[Figure 5] Distribution of total tweet rate before/ after opening and total sales

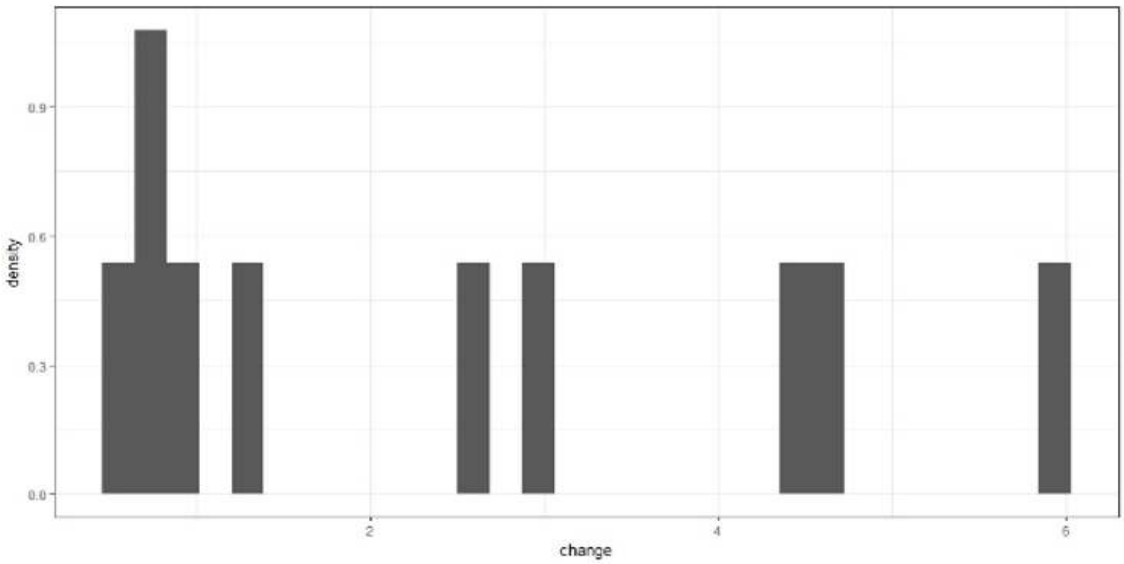


Fig. 6 Increase / decrease rate of positive tweets before and after opening

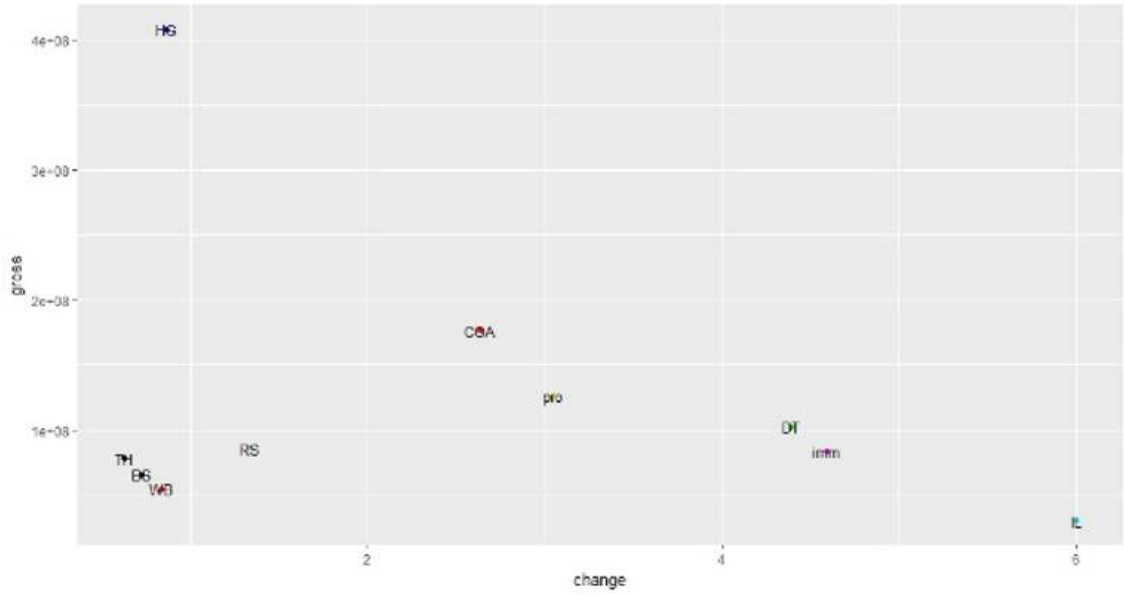
4.2.3 개봉 전후 긍정적인 트윗 개수의 증감률

위의 결과에 개봉 전후의 총 트윗 개수 증감률은 큰 관계가 없음을 확인했다. 그렇다면, 감성 분석(sentiment analysis)을 통해 구분한 긍정적인 트윗과 부정적인 트윗을 이용하여 각각의 트윗 개수 증감률이 총 매출과 관련이 있는가를 확인해보고자 먼저 긍정적인 트윗에 대하여 다음과 같은 가설을 세우고, 검정을 실시해보았다.

- H_0 : 개봉 전후의 긍정적인 트윗 개수의 증감률과 총 매출은 특정한 선형관계가 없다. (r=0)
- H_1 : 개봉 전후의 긍정적인 트윗 개수의 증감률과 총 매출은 특정한 선형관계가 있다. (r≠0)

총 10개의 영화에 대한 긍정적인 트윗 개수의 증감률은 다음 [Figure 6]과 같은 분포를 따르며, 정규성을 만족하지 않음을 확인할 수 있다.

그러므로, spearman 순위상관계수를 이용하여 검정한 결과, p-value는 0.8916으로 귀무 가설을 기각하지 못하며, 상관계수는 약 -0.05454으로 선형적 상관관계가 없음을 확인할 수 있다. 개봉 전후의 긍정적인 트윗 개수의 증감률과 총 매출과의 분포는 다음 [Figure 7]과 같다.



[Figure 7] Distribution of positive tweets rate before/after opening and total sales

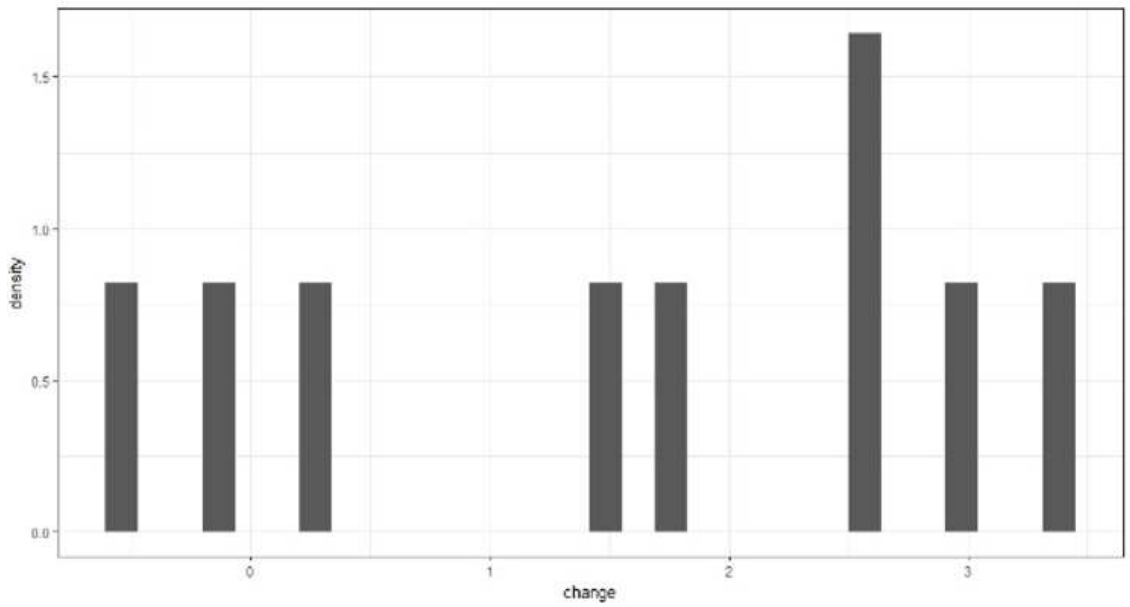
4.2.4 개봉 전후 부정적인 트윗 개수의 증감률

마찬가지로, 부정적인 트윗에 대하여 다음과 같은 가설을 세우고 검정을 실시하였다.

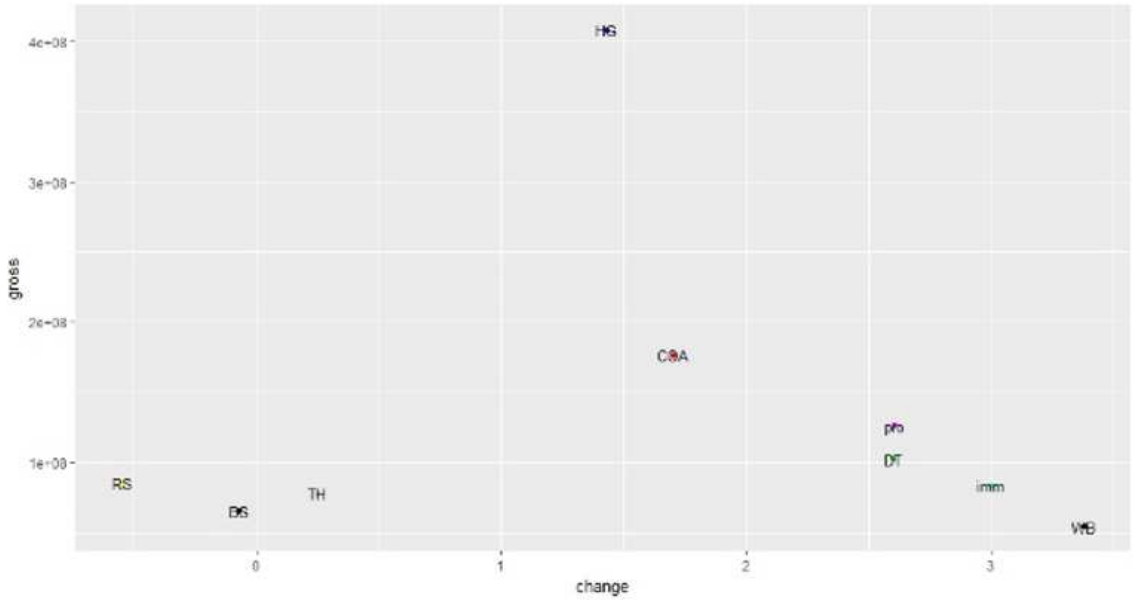
- H_0 : 개봉 전후의 부정적인 트윗 개수의 증감률과 총 매출은 특정한 선형관계가 없다. (r=0)
- H_1 : 개봉 전후의 부정적인 트윗 개수의 증감률과 총 매출은 특정한 선형관계가 있다. (r≠0)

이 때, 데이터 중 'The Iron lady' 영화의 경우 부정적인 트윗 반응이 개봉 전에 존재하지 않으므로, 부정적인 트윗 분석에 있어서는 제외 하였다. 총 9개의 영화의 대한 부정적인 트윗 개수의 증감률은 다음 [Figure 8]과 같은 분포를 따르며, 정규성을 만족하지 않음을 확인할 수 있다.

그러므로, spearman 순위상관계수를 이용하여 검정한 결과, p-value는 0.8432로 귀무가설을 기각하지 못하며, 상관계수는 약 -0.0833으로 선형적 상관관계가 없음을 확인할 수 있다. 개봉 전후의 부정적인 트윗 개수의 증감률과 총 매출과의 분포는 다음 [Figure 9]과 같다.



[Figure 8] Rate of increase / decrease of negative tweets before and after opening



[Figure 9] Distribution of negative tweets before and after opening and distribution of total sales

5. 결 론

본 연구에서는 영화의 매출과 개봉 전후의 트윗 개수 증감률과의 상관성을 살펴보았다. 구체적으로, 영화 매출에 SNS에서 나타난 사람들의 반응이 영향이 있음을 파악하고 그 증감률이 매출과 상관관계가 있는지 연구하였다. 이러한 연구의 목적을 달성하기 위해 먼저 영화 흥행 성과의 대부분이 개봉 후 3주 이내에 결정된다는 연구 결과에 따라 (Kwon, 2014) 사람들의 반응 변화 비율이 이를 따르는 지를 파악하였다. 그 후 공식 영화 편이 나온 후부터 개봉 전까지, 개봉 후부터 폐막까지의 트윗 증감률을 전체, 긍정, 부정적인 부분으로 나누어 매출과의 상관성을 파악하는 실험을 진행하였다.

본 연구의 결과를 요약하면 다음과 같다. 먼저 10편의 영화에 대한 트윗과 매출의 분포가 정규성을 따르는지 확인 후, 정규분포를 따르지 않음을 파악하여 비모수통계 스피어만 순위상관계수(Spearman Rank Correlation Coefficient)를 이용하였다. 개봉 전과 개봉 후 3주까지의 트윗 증감률과 매출과의 상관관계를 살펴본 결과 증감률과 매출 간의 상관성은 없음을 파악할 수 있었다. 이 연구 결과는 영화 흥행 성과에서 개봉 전후로의 반응 변화 비율은 매출에 큰 영향을 받지 않는다는 점을 말해준다. 이후 이 연구를 확장하여, 개봉일을 기준으로 개봉 전과 개봉 후 폐막일까지의 트윗 증감률과 매출의 상관성을 확인했고, 이 또한 상관성이 없음을 파악하였다. 이는 사람들의 영화 개봉 전후로 반응 변화 비율은 매출에 큰 영향이 없음을 다시 한 번 입증하며, 개봉 후 상영기간 동안 이 SNS 반응 변화 추이만으로 마케팅 전략을 바꿀 필요가 없음을 나타낸다.

또한, 감성 분석(sentiment analysis)을 통해 전체적인 트윗 증감률이 아닌, 긍정적인 반응과 부정적인 반응으로 나누어 각각의 경우, 개봉 전후로의 증감률과 매출과의 상관성이 없음을 확인하였다. 따라서 긍정적 혹은 부정적인 감정을 나타내는 반응들의 변화 비율 또한 매출과 상관관계가 없음을 파악하였고, 이는 특정 영화에 관한 사람들의 반응 변화 비율이 영화 흥행에 영향을 미치지 않는다고 할 수 있다.

본 연구는 다음과 같은 한계점이 있다. 첫째, 다양한 연령층의 반응을 살펴보지 못하였다. SNS의 특성상 대부분의 반응은 젊은 층의 반응들로 구성되었으며, 중장년층 이상은 SNS에 익숙하지 못하므로 전 계층이 즐길 수 있는 영화 산업의 현실을 완전히 반영하지 못했다는 한계점이 있다. 추후에는 더 넓은 계층을 대상으로 연구할 방안을 모색해보아야 한다. 둘째, 연구한 영화의 개수가 적었다는 한계가 있을 수 있다. 본 연구에서는 총 10편의 영화에 대한 트윗을 이용하여 연구를 진행하였지만, 영화의 개수를 좀 더 늘린다면 더욱 정확한 결과가 나올 가능성이 있다. 또한, 마지막으로 본 연구에서는 북미지역이 데이터를 이용 하였으므로, 국내 영화를 대상으로 진행한 연구인 영화 흥행 성과의 대부분이 개봉 후 3주 이내에 결정된다는 연구 결과(Kwon, 2014)와의 차이가 있을 수 있다.

6. Reference

[1] Sun Ju Kwon. (2014). Factors influencing Cinema Success: using News and Online Rates. Korean Association for Cultural Economics, 17(1), 35-56

[2] Byung-Do Kim, & Tae-Young Pyo. (2002). Forecasting Model for Box-Office Revenue of Motion Pictures. The joural of management, 36(1), 1-23.

[3] Byoung-Sun Kim. (2007). Comparison of Factors Predicting Theatrical Movie Success : Focusing on the Classification by the Release Type and the Length of Run . Korean Association For Communication And Information Studies, 53(1), 257-287.

[4] Yon Hyong Kim, & Jeong Han Hong. (2011). A Study for the Development of Motion Picture Box-office Prediction Model. Communications of the Korean Statistical Society, 18(6), 859-869.

[5] O-Joun Lee, Seung-Bo Park, Daul Chung, & Eun-Soon You. (2014). Movie Box-office Analysis using Social Big Data, The Korea Contents Society, 14(10), 527-538.

[6] Ho-young Lee, Hee-yeon Kim, Bu-yeon Jung, Duk-Jin Chang, & Ki-hoon Kim (2011). Growth of Social Media and Evolution of Online Social Relationships. Korea Information Society Development Institute.

[7] Seung Yeon Cho, Hyun-Koo Kim, Beom-soo Kim, & Hee-Woong Kim (2014). Predicting Movie Revenue by Online Review Mining: Using the Opening Week Online Review, The Korea Society of Management Information Systems, 16(3), 113-134

[8] Jae Yeob Jeong, & Hyeon-Cheol Kim.

(2016), How does twitter message affect the movie audience?: Focused on type of message source, type of message, and time of movie adoption, 27(6), 179-208.

[9] Min-hoe Heo, Sung-hoon Park, & Sung-jun Jo (2012), Decision Analysis for the Number of Screens of Opening Movies via Audience Evaluation , Korea Business Intelligence Data mining Society.

[10] Austin, B. (1989). Immediate seating: A look at movie audiences. Wadsworth Pub Co.

[11] Asur, S. and Huberman, B.(2010), Predicting the Future with Social Media, Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010.

[12] Bayus, B. L. (1985). Word of Mouth: The Indirect Effects of Marketing Efforts. Journal of Advertising Research, 25(3), 31.

[13] Chen, H., & Zimbra, D. (2010). AI and opinion mining. IEEE Intelligent Systems, 25(4), 72-79.

[14] Dellarocas, C. (2003)., The digitization of word of mouth: promise and challenges of online feedback me- chanisms, Management Science, 49(10), 2003, 1407-1424.

[15] De Vany, A., & Walls, W. D. (1999). Uncertainty in the Movie Industry : Does Star Power Reduce the Terror of the Box Office ? Journal of Cultural Economics, 23(4), 285-318.

[16] De Vany, A., & Walls, D. (1999). Uncertainty in the ovies: Can star power reduce the terror of the box office? Journal of Cultural Economics, 23, 285~318.

[17] Elberse, A., Eliashberg, J., & Leenders, M. A. A. M. (2006). The Motion Picture Industry: Critical Issues in Practice, Current Research, and New Research Directions. Marketing Science, 25(6), 638-661.

[18] Eliashberg, J., & Shugan, S. M. (1997). Film Critics: Influencers or Predictors? Journal of Marketing, 61(2), 68-78.

[19] Faber, R. J., & O' Guinn, T. C. (1984). Effect of Media Advertising and other Sources on Movie Selection. Journalism & Mass Communication Quarterly, 61(2), 371-377.

[20] Ghiassi, M., Lio, D., & Moon, B. (2015). Pre-production forecasting of movie revenues with a dynamic artificial neural network. Expert Systems with Applications, 42(6), 3176-3193.

[21] Gilbert, T. F. (2007), Human competence: Engineeering worthy performance (Tribute ed.), SanFrancisco: Jossey-Bass/Pfeiffer.

[22] Hennig-Thurau, T., Houston, M. B., & Walsh, G. (2007). Determinants of motion picture box office and profitability: an interrelationship approach. Review of Managerial Science, 1(1), 65-92.

[23] Katz, E., & Lazarsfeld, P. F. (1955). Personal Influence. New York, 792.

[24] Lica, L., & Tuta. M. (2011), Predicting Product Performance with Social Media, nformatics in education, 15(2), 46-56.

[25] Litman, B. and Kohl, L. (1989). Predicting Financial Success of Motion Pictures: The '80s Experience. Journal of Media Economics, Fall, 35-50.

[26] Liu, Y. (2006). Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. Journal of Marketing, 70(3), 74-89.

[27] Mishne, G. and Glance, N.(2006), Predicting Movie Sales from Blogger Sentiment, AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs.

[28] Pokomy, M., & Sedgwick, J. (1999). Movie stars and the distribution of financially successful films in the motion picture industry. Journal of cultural economics, 23, 319~323.

[29] Prag, J., & Casavant, J. (1994). An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry. Journal of Cultural Economics, 18, 217~235.

[30] Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46.

[31] Sawhney, M. and Eliashberg, J. (1996). A Parsimonious Model for Forecasting Gross

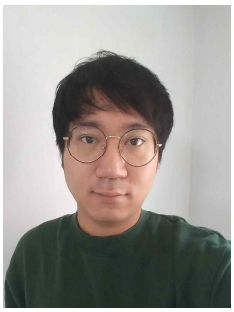
Box- Office Revenues of Motion Pictures. Marketing Science, 15 (2), 113-131.
[32] Shannon, C. E. (2001). A Mathematical Theory of Communication. Retrieved from http://lanethames.com/dataStore/ECE/InfoTheory/shannon.pdf
[33] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics, 37(2), 267-307.
[34] Tun Thura Thet, Na, J.-C., & Khoo, C. S. G. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science, 36(6), 823-848.
[35] Wiebe, J. (1994). Tracking point of view in narrative. Computational Linguistics, 20(2), 233-287. Retrieved from http://scholar.google.com/scholar?q=intitle:Tracking+Point+of+View+in+Narrative#0
[36] Zhang, W., & Skiena, S. (2009). Improving movie gross prediction through news analysis. Proceedings - 2009 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2009, 1, 301-304.

저 자 소 개

박 지 윤

Park, Ji yun is currently a student of Industrial & Management Engineering at School of Engineering, Inha University. She will graduate in Industrial and Manufacturing Engineering from Inha university in 2018. Her research interests include Data Mining, and etc. (sacas96@gmail.com)

유 인 혁



Yoo, In Hyeok is currently a master student majoring Industrial & Management Engineering at School of Engineering, Inha University. He graduated Industrial Management Engineering from Kangwon National University in 2016. His research interests include Data Mining, Sentiment Analysis, and etc. (mkultra1008@gmail.com)

강 성 우

Kang, Sung Woo is currently an Assistant Professor of Industrial & Management Engineering at School of Engineering, Inha University. He received his Ph.D. in Harold and Inge Marcus Department of Industrial and Manufacturing Engineering from The Pennsylvania State University in 2016. His research interests include Data Mining, Massive Data Processing, Product/ Service Design that covers intelligent information & knowledge management. Currently he has been converging traditional quality control and reliability engineering with massive data processing hereby suggesting robust product design models. (kangsungwoo@inha,ac,kr)