http://dx.doi.org/10.15830/kjm.2021.36.3.1

Using Topic Modeling to Analyze Transition of Movie Themes in South Korea

토픽모델링을 이용한 영화 소재의 변천 연구

Giwoong Bae • 배기웅, Hye-jin Kim • 김혜진

ABSTRACT

In this study, we investigate how the themes of films released in South Korea (hereafter, Korea) have changed over time and how these different themes affect box office revenue. We used topic modeling and linear regression analysis to analyze films released in Korea between 2004 and 2018, and obtained the following results. First, we found 20 themes (i.e., topics) each in circulation for wide and limited release movies. Second, trend analysis showed that many themes followed increasing or decreasing patterns for both types of movies. However, depending on the year, different themes were seen to be more profitable, and this effect was more prevalent among limited release movies. At a managerial level, these findings can help film studios, distributors, and exhibitors identify and adopt specific themes that will contribute more to movie revenue at any point in time.

Keywords: Topic Modeling, Latent Dirichlet Allocation, Korean Film Industry, Film Themes, Wide Release Movies, Limited Release Movies

Giwoong Bae | Samsung SDS, First Author
Hye-jin Kim | Korea Advanced Institute of Science and Technology(hyejinkim@kaist.ac.kr), Corresponding Author

pISSN 1229-456x, eISSN 2713-9913

초 록

본 연구는 한국 영화에서의 소재(테마)가 어떻게 변화했는지, 그리고 소재가 박스오피스 수입에 어떻게 영향을 미쳤는지 실증적으로 분석한다. 본 연구는 2004년에서 2018년 사이 한국에서 개봉한 영화를 토픽 모델링과 회귀모형으로 분석하였으며, 주요 결과는 다음과 같다. 첫째, 영화의 소재는 wide release movie와 limited release movie 모두에서 20개의 토픽이 순환해온 것으로 밝혀졌다. 둘째, 트렌드 분석 결과 wide release movie와 limited release movie 양쪽에서 다양한 토픽들이 꾸준한 증가 혹은 감소의 패턴을 보이는 것으로 나타났다. 하지만 연도에 따라 각 토픽의 수익성은 다르게 나타났고, 이 효과는 limited release movie에서 더욱 분명하게 나타났다. 본 연구의 의의로, 영화 제작사와 배급사, 상영사는 이를 이용해 특정 시기에 흥행에 상대적으로 더 도움이 될 것으로 예측되는 소재를 차용할 수 있을 것으로 기대된다.

핵심주제어: 토픽모델링, 잠재 디리클레 할당, 한국 영화 산업, 영화 소재, Wide Release Movies, Limited Release Movies

배 기 웅 | 삼성 SDS, 제1저자
김 혜 진 | 한국과학기술원(hyejinkim@kaist.ac.kr), 교신저자

Ⅰ. Introduction

The central unifying concept of a movie is its theme, comprising the plot, dialog, cinematography, and music. Among these components, audiences consider the plot to be one of the most important factors in their movie-attending decisions. For example, in South Korea (hereafter, Korea), reports state that consumers deem the plot and theme of a film as the most important criteria for choosing an independent film(Choi and Chon 2010). However, there is limited research on the effect of plot on box office revenue, as plot text is in the form of unstructured data that is difficult to analyze.[1)] Therefore, while previous studies have attempted to explain the movie industry with variables that are easier to obtain and measure, such as genre and Motion Picture Association of America (MPAA) ratings, it is difficult to understand the themes exhibited in films and how they change over time, with this limited information. Manual text analysis or human coding of data can also be used to analyze text data (Black and Kelly 2009; Pavlou and Dimoka 2006). However, this approach is prone to subjective analysis and behavioral biases (Kahneman 2003).

It may be noted that recent developments in text mining methodologies make it possible to analyze text data better. One such method is topic modeling, including probabilistic latent semantic analysis (PSLA), hierarchical latent tree analysis (HLTA), and latent Dirichlet allocation (LDA), used to extract dominant topics from several documents, where each document is a mixture of latent topics (Blei and Lafferty 2009).

Here, we find themes (i.e., topics) in movies and examine how they have changed over time, analyzing movie plot text data using LDA. We then examine the effect of the extracted topics on their respective movies' box office success. We divided movies into wide and limited release movies to control for the effect of the opening scale (number of theaters). Moreover, we loosely related "commercial movies" to wide release movies and "independent movies" to limited release ones. While commercial films primarily aim to make profits, independent films are produced to show the creator's artistic intentions and expand the diversity of film culture (Korean Film Council, 2020). This study is investigates the adoption of themes and their effects on box office revenues for movies whose purpose is to make profits and for those whose purpose is to make money as well as evoke favorable reviews/responses. Our research questions are as follows

Research question 1: How are the themes (topics) different for wide and limited release movies?

Research question 2: Relative to all identified movie themes, how does the proportion of each theme change over time for wide and limited release movies?

Research question 3: How do the themes of movies affect the box office revenues of wide and limited release movies each year?

Ⅱ. Theoretical background

1. Variables influencing box office revenue

The film industry has been the focus of many studies (Ahn and Kim 2003; Kim, Kim, and Kim 2014) because the

1. In an exception, Chung and Eoh (2010) performed an empirical analysis to examine the effect of favorable measures of movie scenes on the film market performance, where the favorable measure was collected by participants who had seen the movie.

industry is sizable, the lifecycle of the products (movies) is short and clear, and the relevant data are publicly available. Numerous variables affecting box office revenue have been identified in the extant literature, including genre (Park and Park 2001), viewing ratings (e.g., MPAA ratings in the United States), electronic word of mouth (e-WOM), and the number of opening screens.[2)]

Among the variables, the number of opening screens of a movie is one of the most important variables influencing box office revenues. Movies are classified into wide, platform, and limited release movies, depending on the release scale across the number of opening screens (Einav 2007). Wide release is the most common strategy used by major distributors. These movies first play across thousands of screens with intensive advertising, and the number of screens gradually decreases over several weeks, until the lifecycle ends. A platform release movie starts across a small number of screens, and the number of screens increases after imprinting the movie to consumers. Finally, a limited release movie is released across a small number of screens without any predetermined plans to increase its screen size within its short lifecycle.[3)]

2. Application of text mining in business literature

Text mining refers to the process of deriving knowledge from text data (Aggarwal and Zhai, 2012). This section describes the business literature that apply text mining methodologies, specifically, general text mining methods, and text mining using LDA.

General text mining literature

Textual analysis, the most basic method for dealing with text data, quantitatively analyzes surface measures, such as the readability of the text and the number of characters. Studies have analyzed the impact of variables derived from textual analysis on online purchase conversion rate (Ludwig et al. 2013), the utility of reviews (Korfiatis, GarcíA-Bariocanal, and SáNchez-Alonso 2012), and the stock market performance of a firm (Tirunillai and Tellis 2012). In addition, Yoo and Gretzel (2009) used variables obtained from a textual analysis of hotel reviews to determine the difference between deceptive reviews (reviews not written by actual consumers) and actual reviews. Hu, Bose, Koh, and Liu (2012) proposed a method to detect deceptive reviews using a Wald-Wolfowitz test in an online book market. Ghose, Ipeirotis, and Li (2012) proposed a method to match the preferences of consumers and hotels by combining the variables obtained from the textual analysis and the location of the hotel. For more advanced textual analysis, some studies have applied clustering algorithms to words. For example, Ghose et al. (2012) clustered nouns and noun clauses of reviews and used the clusters as independent variables in a regression analysis. Lee and Bradlow (2011) proposed a method to evaluate the attributes of digital cameras by applying k-means clustering to online reviews.

Topic modeling has advantages over text mining, as it typically exploits more information, such as assuming that each article can have multiple topics simultaneously and that the same word can be used for different topics (Blei and Lafferty 2009). The following section describes the application of topic modeling using LDA in business literature.

2. For a general review of variables that affect box office revenue, see Eliashberg, Elberse, and Leenders (2006) and Hadida (2009).
3. It may be noted that movies can be categorized into two types, wide release and limited release. In this case, the platform release and limited release in the previous classification are categorized into limited release.

Text mining literature using LDA

Topic modeling, proposed by Papadimitriou, Raghavan, Tamaki, and Vempala (2000) and Hofmann (1999), is a statistical model that finds topics in a class of documents. It clusters combinations of words rather than each appearance of a word. For example, if the word dog appears in documents with father, mother, brother, and me, the topic is likely to be about *family*, and if dog appears with wolf, cat, and pigeon, the topic is likely to be about *animals*. Topic modeling assumes that multiple topics can appear simultaneously in different proportions in each document. It is widely used in a variety of disciplines, including computer science and business, and includes various algorithms such as PLSA and LDA. LDA, one of the most well-known topic modeling approaches, assumes a Dirichlet prior for the distribution of document-topic and topic-word (Blei and Lafferty 2009). This is based on the observation that in most cases, topics comprise only a small number of words.

Researchers have analyzed consumer reviews in various domains using topic modeling. Mankad, Han, Goh, and Gavirneni (2016) analyzed 5,380 consumer reviews of all hotels in Moscow using LDA. On the one hand, they found that negative reviews tended to be longer, have fewer topics, and were more influential, with higher volatility of sentiment compared to positive reviews. On the other hand, positive reviews tended to be shorter, had more topics, and were less influential, with lower volatility of sentiment. Calheiros, Moro, and Rita (2017) classified consumer sentiment and analyzed how topics derived from LDA were related to the sentiment. Lim and Lee (2019) used LDA to extract the characteristics of airline consumer reviews of full-service carriers (FSCs) and low-cost carriers (LCCs), associated them with the SERVQUAL model, and found that the most important attributes for consumer evaluation in FSCs and LCCs are tangibility and reliability, respectively. Kim and Lee (2019) performed a segmentation analysis of consumer review text in the US film industry using LDA.

Ⅲ. Data

The data in this study includes information about all movies released in Korea between 2004 and 2018 that were exhibited for more than one week after release. The sample totaled 6,660 movies. We collected the plot of every movie from the website, NAVER movies (movie.naver.com). Fifteen of the 6,660 movies were excluded due to missing plot information, and the final dataset comprised 6,645 movies. On an average, each plot was described in 99.39 words.

1. Morphological analysis and noun extraction

The collected plots were preprocessed using the Python KoNLPy package, a Korean natural language processing library. KoNLPy includes Hannanum, Kkma, Komoran, Mecab, and Twitter[4] classes, and each class has different characteristics. For example, Hannanum handles compound nouns well (Park and Oh 2017), and Mecab is more accurate than other classes for analyzing news articles (Jin, Hong, Lee, and Joo 2019). In a study with online review data, Kim et al. (2018) found that Kkma, Twitter, and Mecab were less sensitive to spacing errors. Furthermore, compared to Kkma,

4. The Twitter class name has been changed to the Okt class, but for the sake of consistency with previous literature, this study describes it as the Twitter class.

Twitter extracted more sentiment words and provided word normalization functions. We used the Twitter class to extract nouns from movie plots, on account of Twitter's popularity (Ko and Yang 2018), its speed in processing large data, and suitability for noun extraction (Choi and Choi 2018). After removing all morphemes, we were left with a total of 512,536 nouns with 29,785 unique nouns.

In the analysis of the Korean language, morpheme analysis is relatively inaccurate when there are typographical errors (typos). However, given that the plot data used in this study are refined text, officially provided by the studios, it is unlikely that there are typos. Nevertheless, to remove all inaccuracies, we manually checked and removed morphemes that were classified as nouns but were not actually nouns. In addition, there were expression related problems for words from foreign languages (e.g., Korea *vs*. Corea), and there were some cases in which words were extracted in a non-noun form (e.g., captured *vs*. capture) because of the incompleteness of the library. All corresponding words were manually modified to their original form. In contrast, we preserved all the compound nouns (e.g., painting ability, working time, writing study, and blond beauty[5]) because this study expected nouns to have additional meaning when compounded.

Then, we created themes consisting of words that could be used regardless of the domain by removing proper nouns.[6] Table 1 lists the nouns that were removed.

The literature on topic modeling indicates that words that appear too often or too infrequently can be removed (Boyd-Graber, Mimno, and Newman 2014; Mankad et al. 2016). For example, Mankad et al. (2016) removed the five most frequent words and the 10 least frequent words. However, to

〈Table 1〉 Types of nouns that were removed.

Nouns that were removed	Remark
The names of characters in creative works (e.g., Avengers and Shrek), names of historical figures (e.g., Mozart and Beatles), and common names (e.g., James and Michael).	
Country names (e.g., US) and words derived from specific countries (e.g., US citizen and US military).	We also removed related words even if they did not contain a country name (e.g., Dollar, Yen)
The names of geographical entities (e.g., Boston and Alaska), universities (e.g., Harvard and Yale), and companies or brands (e.g., Disney, Gucci, and Chanel).	
Words indicating specific events in the past (e.g., American Revolutionary War).	
Words associated with specific religions (e.g., Islam, Catholicism, and Christianity).	However, words that do not belong to a specific religion (e.g., heaven and God) were retained.
Colors (e.g., gray and yellow), numbers and related words (e.g., one, second, and teenager), directions (e.g., front, north, and east), and days (e.g., Monday and Tuesday).	
Words that cannot be used alone.	For example, *dump* is always used with *truck* in Korea, making *dump truck*.
Foreign words that are not used in everyday life in Korea.	In contrast, foreign words that are commonly used in Korea (e.g., couple, mission, and score) were preserved.

5. Compound words are one word in the Korean language.

6. However, we preserved words widely used in everyday life (e.g., Gulliver and Pandora).

minimize information loss due to word removal, we deleted words that appeared only once. On an average, after preprocessing, each plot contained 44.78 nouns.

2. Classification of movies by release strategy

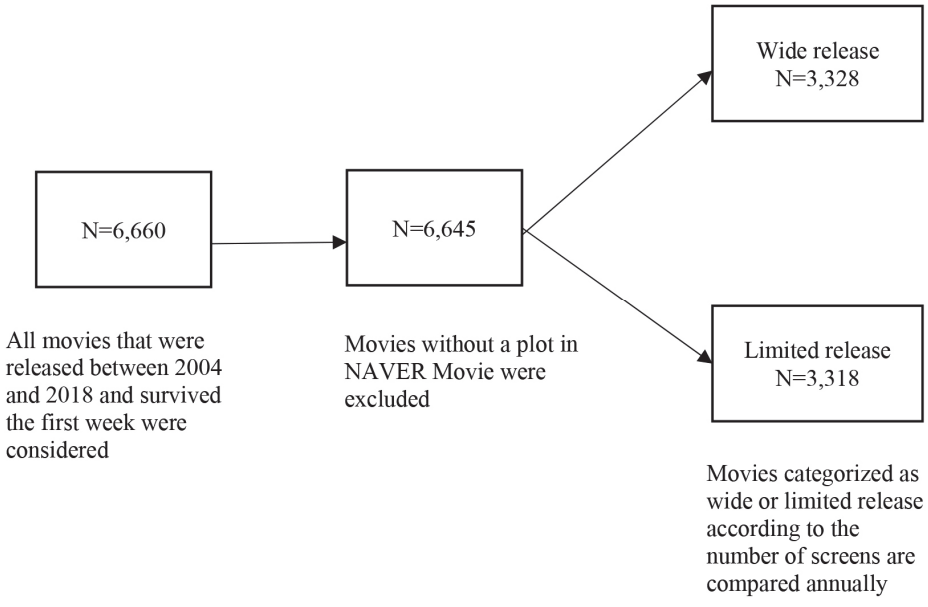
The movie release strategy can be classified based on the relative size of the number of opening screens. For example, Chen, Chen, and Weinberg (2003) investigated the distribution of the number of opening screens, and then classified each movie as wide release if the release was larger than the median number of screens, and as platform release otherwise.[7)]

This study investigated the distributions of the number of first-week screens for each year and defined the movies whose number of screens were more than the median for the year as wide release, and the others as limited release. Figure 1 presents the overall process of sample construction.

Table 2 shows the statistics of variables for wide and limited release movies.

According to Table 2, the average number of first-week screens is 18.64 times higher for wide release movies. By comparing the statistics of wide release movies and limited release movies, the following are observed. First, wide release movies earned 2.6 billion South Korean won (KRWs) on an average in the first week, a value 58 times higher than the earnings of limited release movies. US movies accounted for 41% of wide release movies, but only 19% of limited release movies. With respect to viewing ratings, ages of 18 and over had a proportion of 18% in wide release movies but 28% in limited release movies. Regarding the genre, wide release movies had a high proportion of action, science fiction (SF), adventure, and animation, while limited release movies had a high proportion of drama, documentary, and adult.

〈Figure 1〉 Sample construction process.



7. Limited release films were excluded in the study.

The above comparison can be used to infer the average image of wide release movies and limited release movies. Wide release movies are relatively more popular, likely to be produced in the US, and often consist of genres such as action, adventure, and SF, all of which require huge capital outlays. In contrast, limited release movies tend to be produced for relatively niche markets.

〈Table 2〉 Statistics of variables for wide and limited release movies.

Variable	Wide release movies	Limited release movies
Observations	3,328	3,318
Average first-week revenue (in millions of South Korean won (KRW))	2,637.94	45.38
Average Number of opening-week screens	1,985.05	106.48
Average prerelease e-WOM volume	275.87	37.99
Production country	Korea (32%), US (41%), the others (27%)	Korea (28%), US (19%), the others (53%)
Viewing ratings	G (22%), ages of 12 and over (24%), ages of 15 and over (36%), ages of 18 and over (18%)	G (16%), ages of 12 and over (20%), ages of 15 and over, ages of 18 and over (28%)
Average of genre dummy variables		
Action	0.24	0.13
Comedy	0.21	0.17
Romance	0.14	0.17
Drama	0.37	0.54
SF	0.07	0.03
Family	0.06	0.03
Horror	0.07	0.04
Fantasy	0.08	0.05
History	0.02	0.01
Documentary	0.02	0.11
War	0.02	0.02
Crime	0.08	0.05
Thriller	0.18	0.12
Mystery	0.05	0.03
Adventure	0.13	0.03
Animation	0.18	0.06
Musical	0.01	0.01
Western	0.0024	0.0027
Performance	0.0021	0.0087
Adult	0.0003	0.0018
Others	0.0009	0.0036

Ⅳ. Topic modeling

LDA is a generative probabilistic model of a corpus. The LDA algorithm assumes that each document consists of a set of *k* latent topics, and that each topic *k* has a specific distribution over words. We implemented topic modeling using the gensim library of Python, where gensim is based on Hoffman, Blei, and Bach (2010), who implemented the online training.

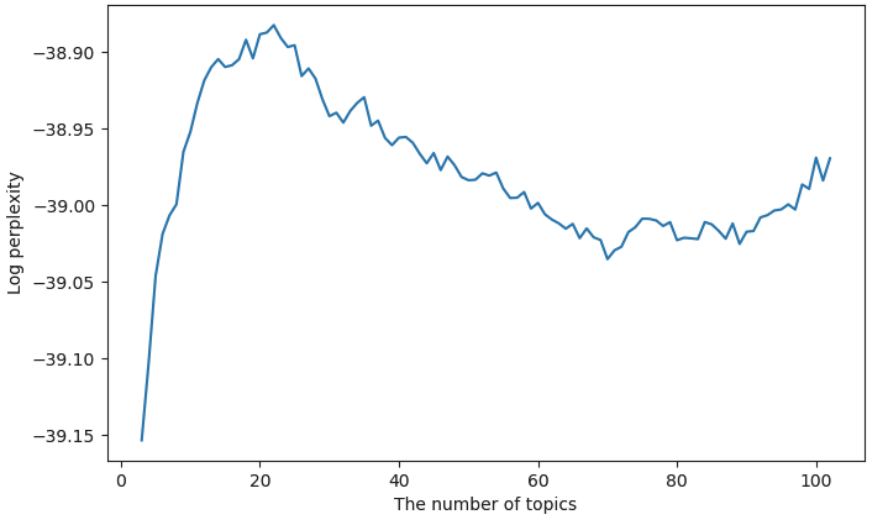
In this study, three hyperparameters, α, η and k were set in the LDA algorithm. α, the hyperparameter of the Dirichlet prior distribution of documents, was set to $\frac{50}{k}$; it varies according to k. The hyperparameter η is the Dirichlet prior distribution hyperparameter over the words in each topic, and was set to 0.01 (Griffiths and Steyvers 2004; Wei and Croft 2006). The number of topics, k was determined by the log perplexity graph, where perplexity is considered a rule of thumb that indicates how well a probability model predicts an actual observation point. Perplexity is defined using information entropy.

The perplexity of a probability model (Bao and Datta, 2014; DiMaggio, Nag, and Blei 2013; Griffiths and Steyvers, 2004) and the number of topics can be determined from the minimal local maximums of a log perplexity graph (Rosen-Zvi, Griffiths, Steyvers, and Smyth 2004; Lim and Lee, 2019). Figure 2 and 3 show the graphs of k and log perplexity for wide release movies and limited release movies, respectively.

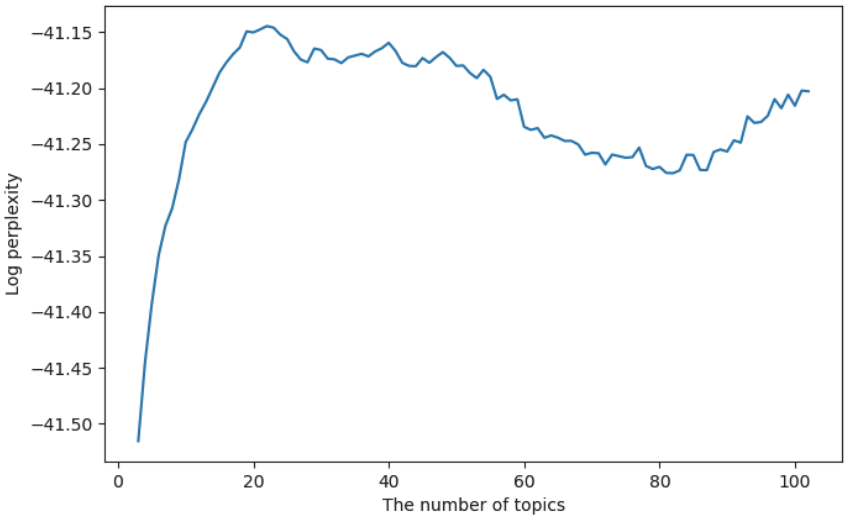
In both the graphs, we observed that the first local maximum of log perplexity occurred at k=20. Therefore, this study determined 20 topics each for wide release movies and limited release movies. The top 20 topic words in the wide release movies and the limited release movies are shown in Appendix A, and the representative movies for each theme in wide and limited release movies are shown in Appendix B.

An examination of the specific themes for both types of movies (see Appendix A) showed that themes such as “Alien Invasion (Theme 2),” “Rise of a Hero (Theme 3),” “Adventure (Theme 12),” and “War (Theme 18)” are often found in typical blockbuster action movies. Furthermore, themes such as “Political Conspiracy (Theme 5),” “Risky game (Theme 6),” and “Vengeance (Theme 9),” may be

〈Figure 2〉 Log perplexity graph by the number of topics for wide release movies



〈Figure 3〉 Log perplexity graph by the number of topics for limited release movies



seen in thriller movies. Lastly, there are also softer themes such as “Brotherhood (Theme 8),” “Pursuit of Hope (Theme 14),” and “Romance (Theme 15).” In contrast, in limited release movies, there are not only similar topics such as “Thriller (Theme 1),” “Romance (Theme 16),” and “War (Theme 10),” but also themes that are distinguishable from wide release movies such as “Passion for Art (Theme 2),” “Aspiration and Friendship (Theme 8),” “Journey (Theme 9),” “Slice of Life (Theme 17),” and “Coming of Age (Theme 20).” Additionally, we found that the “Internationally Renowned (Theme 5)” aspect is expressed in the plots of the movies, emphasizing the fact that a movie has received an international award or is directed by a famous director. Thus, we see some similarities and differences in the identified themes between wide and limited release movies.

Ⅴ. Trend analysis

We conducted a trend analysis to investigate whether the topics obtained in section IV have an increasing or decreasing pattern over time. To this end, regression analysis was performed with the year as the independent variable and the topic weight as the dependent variable. If the obtained regression coefficient is positive, the topic is considered a hot topic, for which interest increases over time, and if it is negative, the topic is a cold topic, for which interest decreases with time (Kim, Choi, and Kwahk, 2017).

1. Trend analysis for wide release movie topics

Base on the trend analysis for the wide release movie group, the hot topics were “The Rise of a Hero (Theme 3),” “Dude with a Problem (Theme 7),” “Gangster (Theme 10),” “Rites of Passage (Theme 13),” “Romance (Theme 15),” and “Horror (Theme 19),” while the cold topics were “Shocking Discovery (Theme 1),” “Aspiration (Theme 4),” “Political Conspiracy (Theme 5),” “Fighter (Theme 11),” and “Humanity (Theme 17).” We found that the neutral topics were “Alien Invasion (Theme 2),” “Risky Game (Theme 6),” “Brotherhood (Theme 8),” “Vengeance (Theme 9),”

〈Table 3〉 Trend analysis results

Wide Release			
Theme	Coefficient	p-value	Hot / Cold
Theme 1: Shocking Discovery	-0.00038	0.018	Cold
Theme 2: Alien Invasion	-6.2E-05	0.446	-
Theme 3: The Rise of a Hero	0.000609	0.000	Hot
Theme 4: Aspiration	-0.00022	0.024	Cold
Theme 5: Political Conspiracy	-0.00159	0.000	Cold
Theme 6: Risky Game	7.35E-06	0.944	-
Theme 7: Dude with a Problem	0.000718	0.000	Hot
Theme 8: Brotherhood	-1.5E-05	0.855	-
Theme 9: Vengeance	-0.00016	0.089	-
Theme 10: Gangster	0.000228	0.013	Hot
Theme 11: Fighter	-0.00017	0.037	Cold
Theme 12: Adventure	0.000165	0.212	-
Theme 13: Rites of Passage	0.000234	0.007	Hot
Theme 14: Pursuit of Hope	0.000118	0.119	-
Theme 15: Romance	0.000273	0.002	Hot
Theme 16: Supernatural	5.88E-05	0.52	-
Theme 17: Humanity	-0.00037	0.000	Cold
Theme 18: War	0.000215	0.062	-
Theme 19: Horror	0.000214	0.004	Hot
Theme 20: Family Drama	0.000119	0.149	-
Limited Release			
Theme	Coefficient	p-value	Hot / Cold
Theme 1: Thriller	0.000427	0.000	Hot
Theme 2: Passion for Art	0.000277	0.000	Hot
Theme 3: Crime Team	0.000344	0.000	Hot
Theme 4: Family	0.0004	0.000	Hot
Theme 5: Internationally Renowned	-0.00024	0.003	Cold
Theme 6: Adulterous Affair	0.000192	0.062	-
Theme 7: Escape from Reality	9.88E-05	0.321	-
Theme 8: Aspiration and Friendship	0.000115	0.220	-
Theme 9: Journey	-6.7E-05	0.415	-
Theme 10: War	-0.00032	0.000	Cold
Theme 11: Disaster	-8.8E-05	0.236	-
Theme 12: Child Growing up	-0.00087	0.000	Cold
Theme 13: Action Thriller	0.000244	0.006	Hot
Theme 14: Crime	0.000505	0.000	Hot
Theme 15: Mystery	-0.00011	0.187	-
Theme 16: Romance	9.78E-05	0.217	-
Theme 17: Slice of Life	-0.00076	0.000	Cold
Theme 18: Revenge	0.000354	0.001	Hot
Theme 19: Love Affair	-0.00013	0.147	-
Theme 20: Coming of Age	-0.00046	0.000	Cold

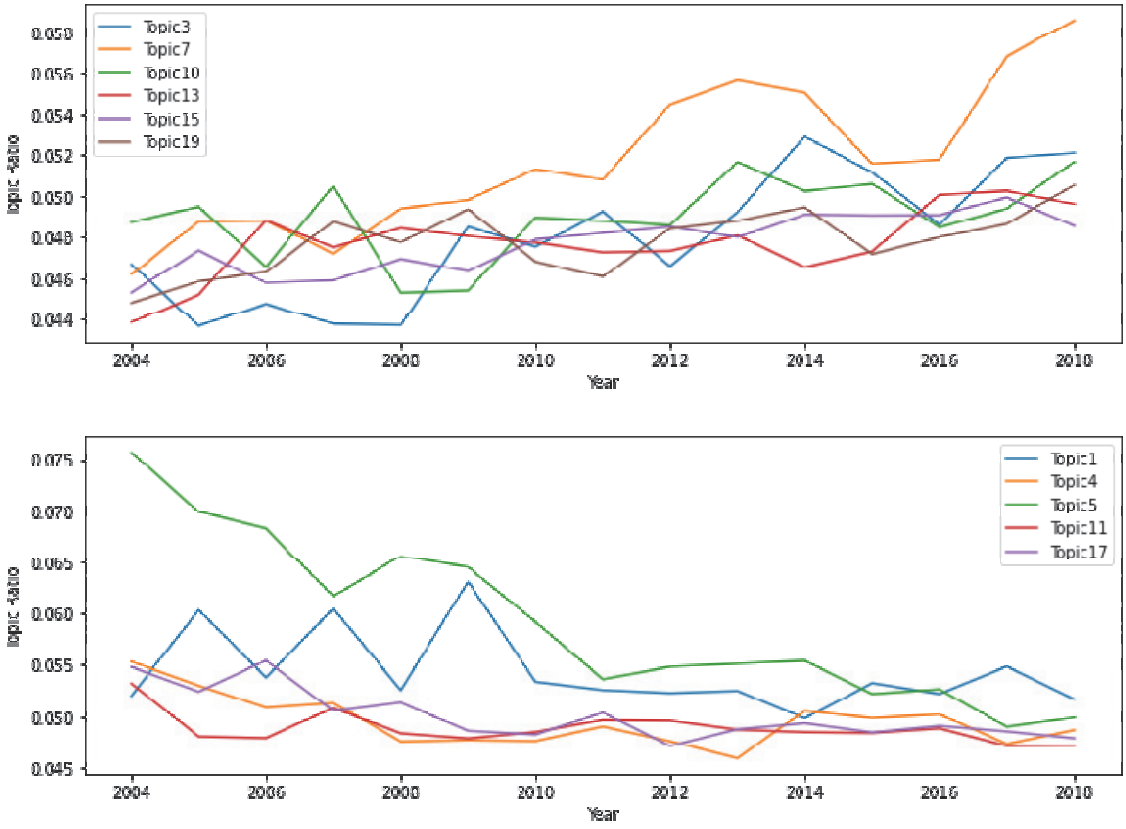
"Adventure (Theme 12)," "Pursuit of Hope (Theme 14)," "Supernatural (Theme 16)," "War (Theme 18)" and "Family Drama (Theme 20)." Graphs depicting changes in the weight of hot and cold topics by year are presented in Figure 4. Trend analysis shows time trends in how different themes were applied in films.

2. Trend analysis for limited release movie topics

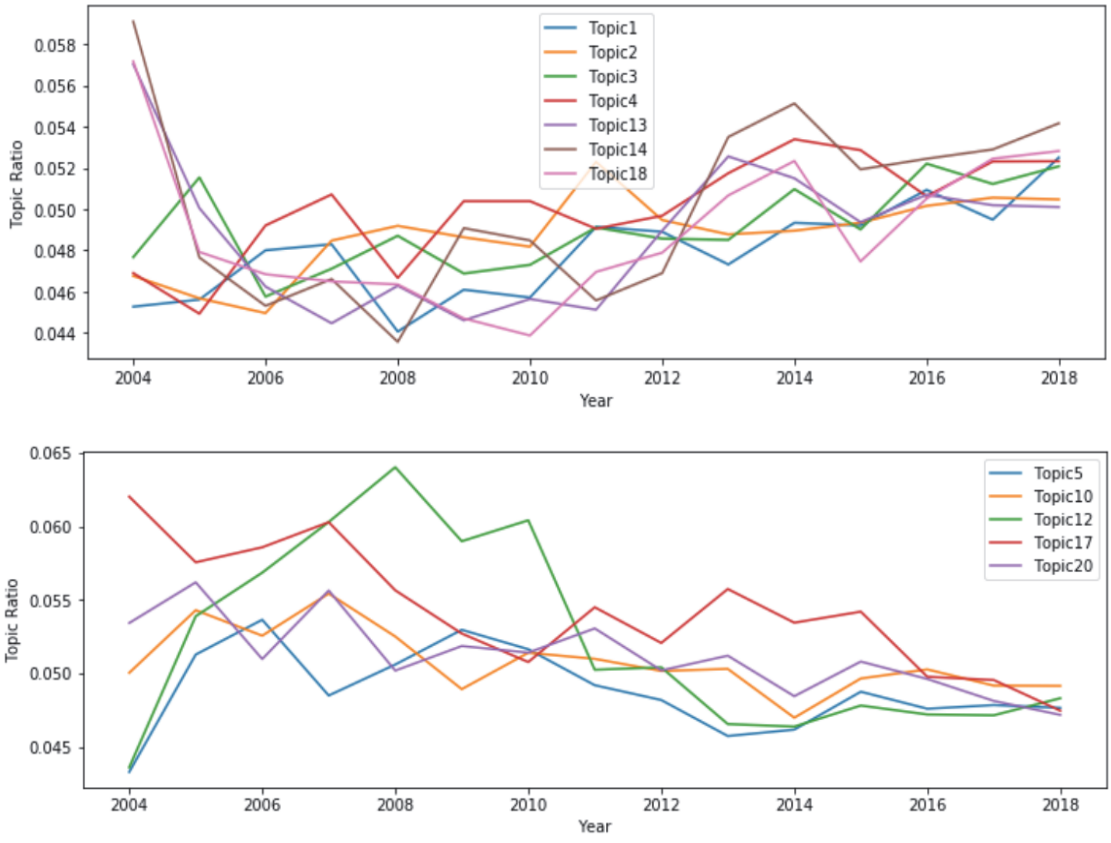
The trend analysis results for the limited release movie were as follows. The hot topics were "Thriller (Theme 1)," "Passion for Art (Theme 2)," "Crime Team (Theme 3)," "Family (Theme 4)," "Action Thriller (Theme 13)," "Crime (Theme 14)," and "Revenge (Theme 18)," while the cold topics were "Internationally Renowned (Theme 5)," "War (Theme 10)," "Child Growing up (Theme 12)," "Slice of Life (Theme 17)," and "Coming of Age (Theme 20)." The neutral topics were "Adulterous Affair (Theme 16)," "Escape from Reality (Theme 7)," "Aspiration and Friendship (Theme 8)," "Journey (Theme 9)," "Disaster (Theme 11)," "Mystery (Theme 15)," "Romance (Theme 16)," and "Love Affair (Theme 19)." Figure 5 shows the trends of hot and cold topics for limited release movies.

Overall, we found that both wide and the limited release movies had their own hot and cold topics. Specifically, wide release movies had six hot topics and five cold topics, and limited release movies had seven hot topics and five cold

〈Figure 4〉 Hot (above) and Cold (below) topics for wide release movies



〈Figure 5〉 Hot (above) and Cold (below) topics for limited release movies



topics. We interpret that limited release movies have slightly more diversity in themes, compared to wide release movies. This may imply that limited release movies or independent films have more room to explore different themes, whereas wide release movies are constrained within themes that have the potential to make money. However, further research is needed to confirm this observation scientifically.

Ⅵ. Regression analysis

To examine the effect of each theme on box office revenue by year, we conducted a regression analysis. The variables used in the analysis are described in the following sections.

1. Variables

lnRev, the dependent variable, is defined as the natural log of the first-week revenue of each movie (in 10,000 KRWs), and *Topic1,⋯,Topic20*, the independent variables, are defined as the proportions of themes in each movie obtained from topic modeling analysis.

The model also included the following variables to control for factors that affect box office revenue. First, the number of opening screens, one of the most important variables affecting box office revenue, was included. We defined the natural log of the number of opening screens as *lnScreen*. Second, to account for the increasing film industry every year, we included and defined *Year*, representing the year of

release of the film. For example, the *Year 2004* is defined as 2004, and the *Year 2018* is defined as 2018. Third, the volume of prerelease electronic word of mouth (e-WOM) was included. The volume of e-WOM has been known to significantly affect movie revenue (Dellarocas, Zhang, and Awad 2007; Duan et al. 2008a; Liu 2006). To control for this effect, 1 was added[8] to the number of prerelease comments in NAVER movies, and the natural log was used to define *lnPreWOM*. Fourth, the country of production was included. To control for the revenue difference depending on the production country, we defined *Country_Korea* for movies made in Korea and *Country_US* for movies made in the United States. Movies made in other countries were used as the baseline. Fifth, viewing ratings (equivalent to MPAA ratings in the US) were included, because the total number of consumers who could watch the movie decreased as the viewing rating increased (i.e., as it was more age-restrictive). We defined the ages of 12 years and over, 15 years and over, and 18 years and over as *Rating_12*, *Rating_15*, and *Rating_18*, respectively, and set G (no restriction) as the baseline. Sixth, we included a variable named *Focus*, to measure the degree to which a movie includes many topics. Mankad et al. (2016) compared the ratings of reviews with a small number of topics and reviews on various topics using the Herfindal-Hershman index (Rhoades, 1993). *Focus* was used to control for the effects of theme diversity. The *Focus* of movie i is defined as:

$$Focus_i = \sum_{k=1}^{k} P(k|i)^2,$$

where is the proportion of theme k in movie i.

We did not include genre in the study. This study collected the genres defined by the Korean Film Council (http://www.kofic.or.kr), that defines 21 genres, such as action, comedy, and romance. However, multicollinearity was expected to occur when genre was included as a control variable due to the similarity of themes and genres. Therefore, this study did not use genre variables in the analysis. Second, we did not measure any sentiment in the plot. Although sentiment analysis can be used with topic modeling (Mankad et al., 2016), we assumed that most of the movie plots were written without any explicit sentiment. Third, we did not include the valence of prerelease e-WOM, because valence information was not available for many limited release movies. Moreover, the effect of valence on sales is insignificant according to some studies (Liu 2006; Duan et al. 2008a; Duan et al. 2008b)

2. Effects of topics on movie revenue

We conducted two types of regression. First, to identify the general influencers of movie revenue, we ran a regression model without topics. Second, to determine which topics affected movie revenue each year, we ran separate regression models for each year, including the proportions of each topic in a movie as independent variables. The analyses were conducted separately for wide and limited release movies.

The model for pooled regression is as follows:

$$\begin{aligned} lnRev_i = {} & \beta_0 + \beta_1 lnScreen_i + \beta_2 Year_i + \beta_3 lnPreWom_i \\ & + \beta_4 CountryKorea_i + + \beta_5 CountryUS_i \\ & + \beta_6 Rating12_i + \beta_7 Rating15_i + \beta_8 Rating18_i \\ & + \beta_9 Focus_i + \epsilon_i. \end{aligned}$$

The model for yearly regression analysis including proportion of topics is as follows:

8. 1 is added in the variable definition to account for the situation when the e-WOM volume is 0.

$$
\begin{aligned}
lnRev_i^k = {} & \beta_0 + \sum_{j=1}^{20} \beta_1^{(j)} Topic_{ji} + \beta_3 lnPrewom_i \\
& + \beta_4 CountryKorea_i + \beta_4 CountryUS_i \\
& + \beta_6 Rating12_i + \beta_7 Rating15_i + \beta_8 Rating18_i \\
& + \beta_9 Focus_i + \epsilon_i \text{ for each year} \\
& (2004 \leq \text{year} \leq 2018),
\end{aligned}
$$

where i is the observation and $Topic_{ji}$ is the standardized topic ratio for the topic's j and ith observations. Standardization was conducted for each topic to prevent multicollinearity.

3. Regression result for wide release movies

The pooled regression results for wide release movies for the entire period (excluding the topics) are presented in Table 3.

The number of opening-week screens and prerelease e-WOM volume positively affects revenue, as reported previously (Liu, 2006). *Focus* was insignificant, implying that the diversity of movie themes in wide release movies does not affect the revenue.

Table 4 reports the results of yearly regression analysis, presenting significant themes and the corresponding year. For the cases where a particular topic appeared to be significant in more than one year (i.e., two or more); if all values are positive, the pattern is defined as positive, negative if all values are negative, and circular if positive and negative values are mixed. Complete results of the regression analysis are provided in Appendix C.

On examining the results, seven of the 20 themes were found to be *Positive*, *Negative*, or *Circular*. Thus, we find that about 40% of the 20 themes influence movie revenue with specific patterns. Specifically, the "Supernatural (Theme 16)" theme had a positive effect on movie box office performance in 2004 and 2014. In contrast, "Political Conspiracy (Theme 5)" consistently had a negative effect on movie box office performance in 2005, 2013, 2015, 2016, and 2018, and "Romance (Theme 15)" had a negative effect in 2005 and 2012.

4. Regression results for limited release movies

The pooled regression results for limited release movies for the entire period (excluding the topics) are presented in

〈Table 3〉 Pooled regression result for wide release movies

Variable	Coefficient	Standard error
lnScreen	1.80***	0.016
Year	-0.18***	0.0029
lnPreWom	0.068***	0.0088
Country_Korea	0.081*	0.032
Country_US	0.13***	0.029
Rating_12	-0.0077	0.033
Rating_15	-0.029	0.031
Rating_18	-0.061	0.036
Focus	0.82	1.18
Constant	356.74***	5.80

Note: *=0.05, ***= p〈0.001

Table 5.

As with the wide release movies, the number of opening-week screens and prerelease e-WOM volume positively affected movie revenue, in line with previous literature. High viewing ratings had negative effects, consistent with intuition. Conversely, *Year* negatively affected revenue as it did for wide release movies, contrary to expectations. Finally, *Focus* was insignificant, implying that the thematic diversity of limited release movies does not affect revenues, same as that for wide release films. Table 6 presents the results of yearly regression analysis with the significant themes and their corresponding years. The complete results are provided in Appendix C.

We found that 15 of the 20 themes were *Positive*,

〈Table 4〉 Significant themes and year for wide release movies

Theme	Significance	Pattern
Theme 1: Shocking Discovery		
Theme 2: Alien Invasion	- (2005), + (2006)	*Circular*
Theme 3: The Rise of a Hero		
Theme 4: Aspiration	+ (2008), - (2009)	*Circular*
Theme 5: Political Conspiracy	- (2005), - (2013), - (2015), - (2016), - (2018)	*Negative*
Theme 6: Risky Game	- (20045, - (2011), - (2013), + (2014), - (2018)	*Circular*
Theme 7: "Dude with a Problem"		
Theme 8: Brotherhood		
Theme 9: Vengeance	+ (2004)	
Theme 10: Gangster	- (2005)	
Theme 11: Fighter		
Theme 12: Adventure	- (2018)	
Theme 13: Rites of Passage	+ (2005)	
Theme 14: Pursuit of Hope	+ (2013), - (2018)	*Circular*
Theme 15: Romance	- (2005), - (2012)	*Negative*
Theme 16: Supernatural	+ (2004), + (2014)	*Positive*
Theme 17: Humanity	- (2018)	
Theme 18: War		
Theme 19: Horror		
Theme 20: Family Drama	+ (2004)	

〈Table 5〉 Pooled regression result for limited release movies

Variable	Coefficient	Standard error
lnScreen	1.17***	0.017
Year	-0.24***	0.0049
lnPreWom	0.12***	0.015
Country_Korea	-0.33***	0.044
Country_US	0.049	0.051
Rating_12	-0.26***	0.063
Rating_15	-0.31***	0.058
Rating_18	-0.14*	0.060
Focus	0.89	2.60
Constant	479.87***	9.83

Note: *= p〈0.05, ***=p〈0.001

〈Table 6〉 Significant themes and year for limited release movies

Theme	Significance	Pattern
Theme 1: Thriller	- (2014), + (2017)	*Circular*
Theme 2: Passion for Art	+ (2009), + (2011)	*Positive*
Theme 3: Crime Team	- (2005), - (2017)	*Negative*
Theme 4: Family	+ (2014), + (2015), - (2017)	*Circular*
Theme 5: Internationally Renowned	+ (2004), + (2011)	*Positive*
Theme 6: Adulterous Affair	+ (2005), + (2011), + (2014)	*Positive*
Theme 7: Escape from Reality	+ (2015)	
Theme 8: Aspiration and Friendship	- (2005), + (2009), - (2017)	*Circular*
Theme 9: Journey	+ (2009)	
Theme 10: War	+ (2004), - (2013)	*Circular*
Theme 11: Disaster	- (2008), - (2013), - (2018)	*Negative*
Theme 12: Child Growing up	- (2005)	
Theme 13: Action Thriller	+ (2014), - (2017)	*Circular*
Theme 14: Crime	- (2017)	
Theme 15: Mystery	+ (2007), - (2013)	*Circular*
Theme 16: Romance	+ (2004), - (2005)	
Theme 17: Slice of Life	- (2007), + (2014), - (2017)	*Circular*
Theme 18: Revenge	- (2005), + (2014), - (2017)	*Circular*
Theme 19: Love Affair	+ (2010), + (2014)	*Positive*
Theme 20: Coming of Age	+ (2009), - (2013),	*Circular*

Negative, or *Circular*, with four *Positive*, two *Negative*, and nine *Circular*. It may be noted that 75% of the notable patterns among the 20 limited release movie themes is greater than 40% of the significant patterns of the wide release movie themes. Thus, the revenues of limited release movies may be more influenced by trends in themes of wide release movies.

Ⅶ. Conclusion

This study examines changes in the themes of movies that were released between 2004 and 2018, and how the themes affected revenues by using topic modeling and regression analysis. The following results were obtained. First, during the 15 years under consideration by this study, there were 20 different movie themes each for wide release movies and limited release movies. Second, the trend analysis showed that many topics revealed increasing or decreasing patterns for both wide and limited release movies. However, different topics became more profitable depending on the year, and this effect was more salient for limited release movies. In other words, the revenues of both wide and limited release movies varied depending on their themes, and the revenues of the limited release movies were more likely to be influenced by the theme trend. This study can potentially help studios, distributors, and exhibitors identify and use themes that are more helpful in increasing movie revenues at any point in time.

The academic contributions of this study are as follows. We proposed a framework for applying topic modeling to the film industry to investigate movie theme trends from the perspective of both the supply side (movie studios) and the demand side (revenue). The framework is not limited to the

Korean film industry, but also applicable to any movie studio, regardless of country, that provides refined plot texts for their releases.

Here, we list out some limitations of the study and avenues for future research. First, in this study, various types of proper nouns were removed, resulting in information loss. Thus, discussions about preserving and analyzing proper nouns can be included in future works. Second, this study examined how movie themes affected box office revenue. Future research should investigate how the themes of these films affect the volume and valence of prerelease e-WOM. Third, we ran yearly regression to examine the effect of topics on movie revenue, noting the availability of in less data and outlier influence susceptibility, i.e., for movies that experienced a huge success. We expected this effect to be more salient for wide release movies. However, when we examined the top movies each year to identify outliers, we noted that this effect did not appear to be critical. We are positive that these research contributions successfully add to the existing literature on the Korean movie industry.

〈Received October 12. 2020〉
〈1st Revised April 30. 2021〉
〈Accepted May 17. 2021〉

References

Ahn, Sung Ah, and Tae Joon Kim (2003), "The Determinants of Opening Share and Decay Rate in Motion Pictures," *Korean Journal of Marketing*, 18(3), 1-17.

Aggarwal, Charu and ChengXiang Zhai (2012), "An Introduction to Text Mining," *Mining Text Data*, 1-10, Boston, MA: Springer.

Bao, Yang, and Anindya Datta (2014), "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures," *Management Science*, 60(6), 1371-1391.

Black, Hulda G., and Scott W. Kelley (2009), "A Storytelling Perspective on Online Customer Reviews Reporting Service Failure and Recovery," *Journal of Travel & Tourism Marketing*, 26(2), 169-179.

Blei, David M., and John D. Lafferty (2009), "Topic Models," in A. N. Srivastava, & M. Sahami (Eds.), *Text Mining: Classification, Clustering, and Applications*, 71-94, CRC Press.

Boyd-Graber, Jordan, David Mimno, and David Newman (2014), "Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements," in Airoldi, E. M., Blei, D., Erosheva, E. A., & Fienberg, S. E. (Eds.), *Handbook of Mixed Membership Models and Their Applications*, 225-254, CRC Press.

Calheiros, Ana Catarina, Sérgio Moro, and Paulo Rita (2017), "Sentiment Classification of Consumer-generated Online Reviews Using Topic Modeling," *Journal of Hospitality Marketing & Management*, 26(7), 675-693.

Chen, Xinlei, Yuxin Chen, and Charles B. Weinberg (2013), "Learning about Movies: The Impact of Movie Release Types on the Nationwide Box Office," *Journal of Cultural Economics*, 37(3), 359-386.

Choi, Garam, and Sung-Pil Choi (2018), "A Study on the Deduction of Social Issues Applying Word Embedding: With an Emphasis on News Articles Related to the Disables," *Korea Society for Information Management*, 35 (1), 231-250.

Choi, M. E., and B. S. Chon (2010), "A Study on Determinant Factors of Behavior in Watching Independent Film," *Korean Journal of Broadcasting and Telecommunication Studies*, 24(5) 503-543.

Chung, Jai Hak, and Ji Yeon Eoh (2010), "A Marketability Forecasting Model for Story-based Content," *Korean Journal of Marketing*, 25(2), 65-88.

Dellarocas, Chrysanthos, Xiaoquan Michael Zhang, and Neveen F. Awad (2007), "Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures," *Journal of Interactive Marketing*, 21(4), 23-45.

DiMaggio, Paul, Manish Nag, and David Blei (2013), "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of US Government Arts Funding," *Poetics*, 41(6), 570-606.

Duan, Wenjing, Bin Gu, and Andrew B. Whinston (2008a), "Do Online Reviews Matter?—An Empirical Investigation of Panel Data," *Decision Support Systems*, 45(4), 1007-1016.

Duan, Wenjing, Bin Gu, and Andrew B. Whinston (2008b), "The Dynamics of Online Word-of-mouth and Product Sales—An Empirical Investigation of the Movie Industry," *Journal of Retailing*, 84(2), 233-242.

Einav, Liran (2007), "Seasonality in the US Motion Picture Industry," *The Rand Journal of Economics*, 38(1), 127-145.

Eliashberg, Jehoshua, Anita Elberse, and Mark AAM Leenders (2006), "The Motion Picture Industry: Critical Issues in Practice, Current Research, and New Research Directions," *Marketing Science*, 25(6), 638-661.

Ghose, Anindya, Panagiotis G. Ipeirotis, and Beibei Li (2012), "Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-generated and Crowdsourced Content," *Marketing Science*, 31(3), 493-520.

Griffiths, Thomas L., and Mark Steyvers (2004), "Finding Scientific Topics," in *Proceedings of the National Academy of Sciences of the United States of America*, 101 (Supplement 1), 5228-5235

Hadida, Allègre L. (2009), "Motion Picture Performance: A Review and Research Agenda," *International Journal of Management Reviews*, 11(3), 297-335.

Hofmann, T. (1999), "Probabilistic Latent Semantic Analysis," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 289-296, Morgan Kaufmann Publishers Inc.

Hoffman, Matthew, Francis R. Bach, and David M. Blei (2010), "Online Learning for Latent Dirichlet Allocation," in *Advances in Neural Information Processing Systems*, 856-864.

Hu, Nan, Indranil Bose, Noi Sian Koh, and Ling Liu (2012), "Manipulation of Online Reviews: An Analysis of Ratings, Readability, and Sentiments," *Decision Support Systems*, 52 (3), 674-684.

Jin, Hoon, Jeoung-Pyo Hong, Kang-Ho Lee, and Dong-Won Joo (2019), "Diagnosis of Corporate Insolvency Using Massive News Articles for Credit Management," in *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 4849-4854, IEEE.

Kahneman, Daniel (2003), "Maps of Bounded Rationality: Psychology for Behavioral Economics," *American Economic Review*, 93(5), 1449-1475.

Kim, Chang-Sik, Su-Jung Choi, and Kee-Young Kwahk (2017), "Investigation of Research Trends in Information Systems Domain Using Topic Modeling and Time Series Regression Analysis," *Journal of Digital Contents Society*, 18(6), 1143-1150.

Kim, Jason, Min Kyoung Kim, Yeoeun Park, Eomji Kim, Junhee Lee, Dongho Kim, and Seonho Kim (2018), "Sentiment Analysis of Korean Teenagers' Language Based on Sentiment Dictionary Construction," in *International Conference on Frontier Computing*, 541-550, Springer.

Kim, Jongdae, and Youseok Lee (2019), "Market Segmentation Approach Using Topic Model Analysis: Focusing on Movie Market," *Korean Journal of Marketing*, 34(4), 53-72.

Kim, Youngju, Dong Soo Kim, and Jae Hwan Kim (2014), "Non-compensatory Decision Making for Movie Choice: Role of Genre and Online Word of Mouth," *Korean Journal of Marketing*, 29(1), 1-20.

Ko, Dong-Woo, and Jung-Jin Yang (2018), "Korean Natural Language Processing and Analysis Using KoNLPy and Word2Vec", *Journal of Korea Information Science Society*, 2140-2142.

Korea Film Council (2020), "What Are the Independent Films?", retrieved from https://www.kofic.or.kr/kofic/business/guid/introGuideKorMovie.do

Korfiatis, Nikolaos, Elena GarcíA-Bariocanal, and Salvador SáNchez-Alonso (2012), "Evaluating Content Quality and Helpfulness of Online Product Reviews: The Interplay of Review Helpfulness vs. Review Content," *Electronic Commerce Research and Applications*, 11(3), 205-217.

Lee, Thomas Y., and Eric T. Bradlow (2011), "Automated Marketing Research Using Online Customer Reviews," *Journal of Marketing Research*, 48(5), 881-894.

Lim, Juhwan, and Hyun Cheol Lee (2019), "Comparisons of Service Quality Perceptions between Full Service Carriers and Low Cost Carriers in Airline Travel," *Current Issues in Tourism*, 1-16.

Liu, Yong (2006), "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of Marketing*, 70(3), 74-89.

Ludwig, Stephan, Ko de Ruyter, Mike Friedman, Elisabeth C. Brüggen, Martin Wetzels, and Gerard Pfann (2013), "More than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates," *Journal of Marketing*, 77(1), 87-103.

Mankad, Shawn, Hyunjeong S. Han, Joel Goh, and Srinagesh Gavirneni (2016), "Understanding Online Hotel Reviews through Automated Text Analysis," *Service Science*, 8(2), 124-138.

Papadimitriou, Christos H., Prabhakar Raghavan, Hisao Tamaki, and SantoshVempala (2000), "Latent Semantic Indexing: A Probabilistic Analysis," *Journal of Computer and System Sciences*, 61(2), 217-235.

Park, JunHyeong, and Hyo-Jung Oh (2017), "Comparison of Topic Modeling Methods for Analyzing Research Trends of Archives Management in Korea: Focused on LDA and HDP," *Journal of Korean Library and Information Science Society*, 48(4), 235-258.

Park, Hyung Hyun, and and Chan Su Park (2001), "The Relationship between Critical Reviews and Performance of Movies: Does It Hold in the Internet Era?," *Korean Journal of Marketing*, 16(4), 71-85.

Pavlou, Paul A., and Angelika Dimoka (2006), "The Mature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation," *Information Systems Research*, 17(4), 392-414.

Rhoades, Stephen A. (1993), "The Herfindahl-hirschman Index," *Federal Reserve Bulletin*, 79(3), 188.

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth (2004), "The Author-topic Model for Authors and Documents," in *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence*, 487-494, AUAI Press.

Tirunillai, Seshadri, and Gerard J. Tellis (2012), "Does Chatter Really Matter? Dynamics of User-generated Content and Stock Performance," *Marketing Science*, 31(2), 198-215.

Yoo, Kyung Hyan, and Ulrike Gretzel (2009), "Comparison of Deceptive and Truthful Travel Reviews," in Hopken, Wolfram, Ulrike Gretzel, Rob Law (Eds.), *Information and Communication Technologies in Tourism 2009*, 37-47, Springer-Verlag, Netherlands.

Wei, Xing, and W. Bruce Croft (2006), "LDA-based Document Models for Ad-hoc Retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 178-185, ACM.

〈Appendix B〉

〈Table B1〉 Representative movie for each theme in wide release movies.

Theme	Proportion	Title	Plot
1	0.18	Genome Hazard	My first birthday after the marriage, my wife died | Ishigami leads a happy life with his beloved wife. He and his wife go home together for the first time together after marriage to celebrate his birthday. Ishigami returns to his home happily, but he finds his wife's cold body, and he falls in shock. *(the rest omitted)*
2	0.27	Transformers	The transformers, extraterrestrial creatures with much higher intelligence and power than humankind, are divided into Autobots that defend justice and Decepticons that represent evil. They have been fighting for a long time to possess the ultimate energy source 'the Cube'. *(the rest omitted)*
3	0.20	Elena and the Secret of Avalor	Princess Sophia left for the kingdom of Avalor to save princess Elena, but there was no happiness and music in the kingdom due to the evil wizard Shuriki. Could princess Sophia save the Princess Elena and the kingdom against Shuriki?
4	0.21	The Sound of a Flower	A girl whose destiny is the sound was born in the era that women could not play Pansori! | At the end of Joseon dynasty, it was prohibited to break taboos. 'Chaeseon Jin' (Suji Bae) appeared to 'Jaehyo Shin' (Seungryong Ryu), the head of the first Pansori school in Joseon. *(the rest omitted)*
5	0.26	City of Damnation	Warranty! Laughter is guaranteed! They are coming! | Traffic police 'Chungdong Jang' joins to a special investigation team because he is not known outside. He takes on a special role of infiltration as a new member of a gangster group. *(the rest omitted)*
6	0.27	A Cruel Attendance	Comic suspense | The embarrassing situation of a humane kidnapper | An amateur kidnapper, his own daughter was kidnapped! | A kind and ordinary office worker Dongchul (Suro Kim) faces the biggest crisis of his life with investment failure and a large amount of interest on the bond! *(the rest omitted)*
7	0.23	Hearty Paws 2	Maumi 2 returns to the family | Maumi that echoed the whole country comes back to the family! | Dongwook (Jungki Song), who moved to his third high school already, does not study at all though he is soon to graduate. Maumi, which was a gift from his dead father, is his only friend. *(the rest omitted)*
8	0.24	My Brother...	In the late 1990s, brothers whose ages are within one year are attending the same class at a high school. The younger brother, Jonghyun (Bin Won), is good at fighting and has a good-looking face, and his older brother, Seonghyun (Hakyoon Shin), is good at studying and very kind. *(the rest omitted)*
9	0.19	Musudan	Special agents were dispatched to find out the unidentified incident at DMZ in Korea! | Everyone dies if they cannot come back within 24 hours! | After the consecutive and unidentified deaths and disappearances in DMZ, the army dispatched special agent Jinho Cho (Minjoon Kim) and Yuwha Shin (Jia Lee). *(the rest omitted)*
10	0.33	The Flock	Now, the girls are disappearing... | A teenage girl's disappearance occurred in the jurisdiction of the Department of Public Safety agent Erroll, but the police concluded it was a simple runaway. On the first day of tracking, the criminal deliberately left an article about the missing girl for Erroll to see. *(the rest omitted)*
11	0.26	Riverbank Legends	You are over if you are pushed out of battle! | All legends begin at 18:1 battle | Kyungro Yoo (MC Mong) is a chat-fighter who bluffs and causes confusion. Sunghyun Ki (Chunhee Lee) always seems to stop the fight but he suddenly beats the opposition with his fist. *(the rest omitted)*
12	0.24	Elias and the Treasure of the Sea	Defend the treasures of the sea against the evil 'Arctic Queen'! | Elias' rescue adventure begins to save the village! | Cute rescue boat 'Elias' is busy with keeping fishing boats in a harbor town. *(the rest omitted)*
13	0.16	Hector and the Search for Happiness	Hector, a London psychiatrist who meets crying people every day, wonders what the true happiness is, and he goes on a journey to find happiness. *(the rest omitted)*
14	0.21	Glory for Everyone	"I want to be a football player!" | Six years drama of Korea's first regional children center youth football club 'Hope FC'! | The striker Sunghoon who cannot score in the match, an eternal candidate Byeonghoon, a timid striker Gyuan, … *(the rest omitted)*
15	0.35	My Wife Got Married	Again… my wife got married! | Are you confident? | Dukhoon loves only Ina in his life. She has a cute appearance with an intellectual aspect that loves a book, and a knowledge and passion for football that is as good as a man. The more he meets her, the more he loves into her. *(the rest omitted)*
16	0.20	Isle of Dogs	Today, all the dogs in the world disappeared! | As a dog pandemic threatening human spreads, all the dogs in the world are banished to the garbage island, and a boy who lost his beloved dog leaves for the island to find the dog. The boy meets five special dogs there. *(the rest omitted)*
17	0.19	The Little Ghost	The wish of a little ghost living in an old castle is to watch the world. One day, his little wish came true like a miracle! But with a little glad moment, the ghost's body that reached the bright sunlight turned black! *(the rest omitted)*
18	0.23	Robin Hood	A huge spectacle reverses the world! | 13th century England, Robin Hood, who is a commoner but has excellent ability, went to the war for king Richard. Robin was trusted by the king in the French battle, but the king was killed in battle. *(the rest omitted)*
19	0.19	Final Destination 2	Still alive? | Death is closer than you might imagine! | Do you want fear? | Scale is fear! | Kimberly, who travels with her friends on a weekend trip, sees an illusion that many people including herself die horribly in a highway chain crash. *(the rest omitted)*
20	0.24	Wedding Dress	A wedding dress, a heartfelt farewell gift | I do not want to let my mother go yet | Mom, could you just stay with me like now? I'm sorry, my daughter... I'm so sorry that I left you. You are the most precious one in the world, and there are so many things that I want to do for you. *(the rest omitted)*

〈Table B2〉 Theme-wise Representative movie for each theme in limited release movies

Theme	Proportion	Title	Plot
1	0.19	The Whole Ten Yards	Killer comedy in 2005 | I'll kill them for you | Oz (Matthew Perry) is a dentist who is struggling with nagging wife and greedy mother-in-law. One day, an unusual man moves to his neighbor. *(the rest omitted)*
2	0.20	Sister Smile	Belgium in the late 1950s. Jeanine wants to get a free life through music, but she feels irritated by her mother who always says that "it is best for a woman to meet a good man and get married," so she enters the nunnery. Even in the nunnery, she could not give up her passion for music. *(the rest omitted)*
3	0.27	Venice Underground	A drug-related murder takes place on the shores of Venice, Los Angeles, and the mayor urges the police chief to resolve the case within 48 hours. The chief selects an academy candidate who knows drugs well to create an informal special drug investigation team. *(the rest omitted)*
4	0.26	Dear Pyongyang	Father, now I want to tell you my story. I was born in Osaka, a city called 'shrine of Korean residents in Japan,' and grew up as the youngest sister of three brothers. My father left Jeju island, his hometown, at the age of 15 and came to Japan. After liberation, he chose North Korea as his 'motherland'. *(the rest omitted)*
5	0.30	Who's Camus Anyway?	A fresh youthful movie reverses with breathtaking tension and excitement. | What would I do if I kissed him... How would I feel if I killed a person... | A lively university campus in the city center. Students in the 'Visual Workshop' of Literature Department produce films as a part of their curriculum. *(the rest omitted)*
6	0.24	Desire	I want to take it! | A couple indulges in a man at the same time | One day she knew her husband was having an affair, and then her new life began. *(the rest omitted)*
7	0.21	Little Ashes	The 18-year-old Salvador Dalí, who entered college in Madrid, Spain, meets Federico García Lorca and Luis Buñuel, who will later become Spanish famous poet and film directors, respectively. Dalí's genius and uniqueness attracts their attention they share a friendship together. *(the rest omitted)*
8	0.19	Nodame Kantabire: Saishuu-gakushou-Kouhen	Their last love concerto! | Nodame (Juri Ueno), who is an unusual but lovely genius pianist, and Chiaki (Hiroshi Tamaki), an orchestra's permanent conductor, has growing love and dreams for music through Paris, Vienna and Prague. *(the rest omitted)*
9	0.39	Brother Bear	Ten years after *Lion King*... Disney's new release of life drama | There are three brothers when a giant mammoth lived in North America. Kenai is the youngest among three brothers, and is receiving a totem ritual, which is performed by a tribal God of North America so that Kenai can take good care of his life. *(the rest omitted)*
10	0.23	Last Scout	In 2065, the last survivors who escaped the earth which was destroyed by the nuclear war leave for a new planet where human can settle. The crew of the ship Pegasus also begin their long journey to find hope, but life threatening beings are waiting for them! Now they begin the last war of the universe for the survival of mankind.
11	0.22	Battle Royale II	Absolute war survivor! | All participation of the third grade B class! | Removal of anti-BR law organization, time limit 3 days… rules were changed! | The new century education reform act, the so-called 'BR Law', has been launched, which forces people to kill each other until the last one is left, among a randomly chosen class in a middle school. *(the rest omitted)*
12	0.43	JSA Inter-Korean Elementary School	South and North Korea establish an Inter-Korean elementary school to prepare the unification. A peaceful Daeseongdong free town. The village heads talk with the villagers. Jonghak's mother and Eunbyeol's mother care for pepper and talk about a nerdy old man. *(the rest omitted)*
13	0.16	The Devil's Double	Latif Yahia, an alumni of Saddam Hussein's son Uday Hussein, was suddenly ordered to come to the Palace of Saddam Hussein one day in 1987. *(the rest omitted)*
14	0.29	Mad Detective	Detective Bun left the police station a few years ago with the stigma of 'mad detective.' Meanwhile, detective Wong who was tracing a suspect in the forest disappeared, and his peer, Chi-Wai, only returns safely. After 18 months of the disappearance of Wong, a series of questionable murders occurred in the city downtown. *(the rest omitted)*
15	0.24	Laputa: Castle in The Sky	A legendary castle is revealed beyond the clouds when the mysterious necklace shines brightly! | An airship floats in the quiet night sky. The skull mark on the tail wing shows that it is a pirate ship. People suddenly run. *(the rest omitted)*
16	0.24	Paris, I Love You	Eighteen colors serenade of love | Paris, wherever you go you fall in love | Fog light love 'Montmartre' (Director: Bruno Podalydès): Man parking in the narrow alley of Montmartre met the women of his destiny | Milky love 'Quais de Seine' (Gurinder Chadha): A French boy looking for a girl in la Seine fell in love with a Muslim girl! *(the rest omitted)*
17	0.19	Tatsumi	"I hope to naturally draw the picture with my spirit more than making a story." Tears, empathy, hope, affection, comfort... Five stories of living characters. *(the rest omitted)*
18	0.18	Escape Plan 2: Hades	Breslin, the world's best escapee, escape the worst prison Hades armed with state-of-the-art systems! | Breslin, one of the best escape experts on the planet, goes to Hades, one of the world's best dungeons, to save his colleagues trapped by Kimbral, the betrayer. *(the rest omitted)*
19	0.19	Lost Flower Eo Woo-dong	The woman who is beloved by the king of Joseon dynasty becomes the best flower! | "I am Eo Woo-dong, the flower in the flower!" | Hyein is a well-known woman who has a beautiful appearance and a superior academic ability. One day, 'Dong Lee,' relative of the king, approaches Hyein. *(the rest omitted)*
20	0.23	Action Boys	Even you don't remember, (we are action actors) | They struggle for the best cut! | Sejin had a lot of debt owing to his tiger tattoo on his back because of the word of fortune teller. Jinseok is a boxer who became a hairdresser because he liked Winona Ryder in *Edward Scissorhands*. *(the rest omitted)*

〈Appendix C〉

〈Table C.1〉 Regression results by themes and years for wide release movies

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Topic1															
Topic2		-0.912^{+}	0.121*												
Topic3															
Topic4					0.090^{+}	-0.140**									
Topic5		-0.140*								-0.128*		-0.134*	-0.138*		-0.168*
Topic6		-0.103^{+}						-0.082^{+}		-0.130**	0.094^{+}				-0.139*
Topic7															
Topic8															
Topic9	0.128^{+}														
Topic10		-0.099^{+}													
Topic11															
Topic12															-0.101^{+}
Topic13		-0.091^{+}													
Topic14										0.082^{+}					-0.105^{+}
Topic15		-0.091*							-0.157**						
Topic16	0.124*										0.092*				
Topic17															-0.096^{+}
Topic18															
Topic19															
Topic20	0.209*														

Note: $^{+}$= p〈0.1, *= p〈0.05, **= p〈0.01, ***=p〈0.001. Insignificant results are not shown in the table.

〈Table C.2〉 Regression results by themes and years for limited release movies

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Topic1											0.155*			-0.167+	
Topic2						0.184+		0.148+							
Topic3		-0.168+												-0.295**	
Topic4											0.108+	0.094+		-0.184*	
Topic5	0.282*							0.222*							
Topic6				0.237+				0.241**			0.196*				
Topic7												0.103+			
Topic8		-0.235*				0.226*								-0.234*	
Topic9						0.168+									
Topic10	0.182+									-0.191*					
Topic11					-0.167+					-0.137*					-0.204+
Topic12		-0.280+													
Topic13											0.158*			-0.230*	
Topic14														-0.358**	
Topic15				0.224*						-0.112+					
Topic16	0.174+	-0.206+													
Topic17				0.201+							0.127+			-0.340*	
Topic18		-0.278*									0.195**			-0.267*	
Topic19							0.199+				0.174**				
Topic20						0.210*				-0.165*					

Note: + = p〈0.1, * = p〈0.05, **=p〈0.01. Insignificant results are not shown in the table.