Master's Thesis of Hye Jee Yang

The Effect of Similarity to the Original Novel on the Box Office and the Reason

원작 소설과의 유사도가 영화 흥행에 미치는 영향과 원인

August 2023

College of Business Administration
Seoul National University
Marketing Major

Hye Jee Yang

The Effect of Similarity to the Original Novel on the Box Office and the Reason

Hyejee Yang

Submitting a master's thesis of
School of business

April 2023

College of Business Administration
Seoul National University
Marketing Major

Hyejee Yang

Confirming the master's thesis written by
Hyejee Yang
July 2023

Chair	김준범	(Seal)
Vice Chair	이경미	(Seal)
Examiner	김상훈	(Seal)

Abstract

As the OTT market develops and the film market grows, film production companies are trying to secure narratives that can be scripted by acquiring intellectual property rights such as books and webtoons. This paper investigates the effect of similarity with the original, which is expected to be a success factor of the best-selling original film using various approaches. To investigate the effect of similarity between the film and the original on box office performance, 77 films based on novels released between 2015 and 2022 were collected through web crawling. In Study 1, it was confirmed that similarity to the original positively affected box office performance but had a negative correlation between action and mystery films. In Study 2, it was confirmed whether the audience who watched the movie considered it similar to the original in deciding to watch by adding a reference to the original in the movie review to the model. However, the effect of the amount of buzz about the original did not replace the effect of similarity. It was confirmed that the similarity with the original was used to reduce the risk of failure. This study presents a theoretical basis for brand expansion in cultural products to expand the research area and suggests that producers consider the probability and genre characteristics of the story when producing them.

Keywords: Natural language processing, Brand extension, Book adaptation, Motion picture industry, Box office revenue
Student Number: 2021-23235

Table of Contents

Chapter 1. Introduction

The Asian Content Film Market (ACFM) opened with the Busan International Film Festival in October and had publishers' booths. It was an event for exchanges between production officials and story providers, reflecting the current market situation in which various novels or webtoons are being adapted into movies. Here, publishers selected major works suitable for visualization and presented them to film producers. There were 1,027 meetings held at ACFM from October 8 to 11 (김은형, 2022).

Adaptation from novels is a traditional strategy to reduce the risk of introducing new movies. Many movies using the adaptation strategy, such as Harry Potter, showed successful performances at the box office. However, some movies failed in the film market, even though they are based on best-sellers.

Based on brand extension theories (Aaker & Keller, 1990; Chun, Park, Eisingerich, and MacInnis, 2015; Miniard, Alvarez, and Mohammed, 2020), this paper seeks whether the similarity between book and film outperforms as a determinant of the performances of book-to-film adaptation. First, the paper examines the motion pictures that have more similarities with the original works. Then the paper will investigate the impact of book adaptation by analyzing the user reviews and checking the impact of books on film revenue.

Therefore, the study confirmed the effect of similarities between movies and original novels on box office performance, including MPAA rate, running time, director power, and actor power used in previous studies. In addition, by confirming the moderating effect of the genre and pandemic situation, it was suggested which factors play a significant role in the success of the adaptation film. The results of this

research are expected to confirm how significant the degree of adaptation from the original novel is in predicting the film's success while enhancing the understanding of academia and the public of the transformative strategy of the recently increased importance.

Chapter 2. Related Literature

2.1. Movie Revenue Prediction

Previous works of literature on movie revenue prediction usually focused on finding the movie properties that affect the revenues. For example, Star power can influence box office revenue. Many studies consider the casting power as the covariate in the revenue forecasting model. However, the effect of celebrity participation has diverged. Some researchers, such as De Vany and Walls (1999) and Ravid (1999), showed that whether the movie star participates in a film does not significantly affect the movie revenue. On the other hand, Liu, Liu, and Mazumdar (2014) confirmed the impact of star power on box office performance.

Another factor that could increase or decrease the films' revenue is the review of the critics. As critics are predictors and influencers in the motion picture industry, the critics' reviews of the film correlated with the film's performance (Basuroy, Chatterjee, & Ravid, 2003; Eliashberg & Shugan, 1997). Not only is the review's valence critical, but the number of reviews is also a decisive factor in predicting movie revenue (Boatwright, Basuroy, & Kamakura, 2007).

The effect of the audience review is also an essential factor that could be used to predict movie revenue (Berger, Soensen, & Rasmussen, 2010; Chintagunta, Gopinath, & Venkataraman, 2010; Duan, Gu, & Whinston, 2008; Liu, 2006; Song,

Huang, Tan & Yu, 2019; Zhu & Zhang, 2010). Studies involving the relationship between eWOM and film performance usually focus on the valence and the volume of eWOM. According to Chintagunta et al. (2010), the valence of online reviews positively affects box office performance. The result of Song et al. (2019) supports that the number of microblogging posts positively affected revenue. The content of the user review is also essential. Ryoo, Wang, and Lu (2020) showed that spoilers in user reviews generate a positive effect on box-office scores.

Recently, the content of the movies has been introduced as a predictor of the box office score. Eliashberg, Hui, and Zhang (2007) identifies the textual content of movie scripts to predict the commercial success of the movies. Extended works, such as Eliashberg, Hui, and Zhang (2014), Kim, Lee, and Song (2021), and Moon, Jalali, and Song (2022), used machine learning approaches to predict box office revenue. In this context, our research would also use a machine-learning method to analyze the original books' textual data and the film's textual information.

2.2. Movie Adoption as Brand Extension

Adoption from the book is one of the traditional strategies to take over a previously successful brand image. For example, Joshi and Mao (2012) examined that the aspects of the original book, such as its ranking and book recency, positively affect the adapted movie's opening revenue. Knapp, Henning-Thurau, and Mathys (2014) also suggested that there are spill-over effects in the context of movie adoption. These researchers view the adoption of novels as an example of brand extension. In particular, Joshi and Mao (2012) insisted that the similarity between the original textual and the adopted motion films showed a significant effect on box

office performance. The researchers pointed out that the process of movie adoption is relevant to the process of brand extension, but the direction of extension may be revealed differently with a movie sequel.

The successful brand extension involves the parent-brand characteristics. These characteristics include the quality of the parent brand (Joshi & Mao, 2012; Smith & Park, 1992), parent-brand conviction (Kirmani, Sood, & Bridges, 1999), and parent-brand experience (Swaminathan, Fox, & Reddy, 2001; Völckner & Sattler, 2006). Notably, the fit between the parent brand and the extension product is an important measure to predict the success of brand extension (Aaker & Keller, 1990; Chun, Park, Eisingerich, & MacInnis, 2015; Miniard, Alvarez, & Mohammed, 2020). The fit matters to the successful brand extension because it could lead to the high accessibility of the parent brand (Miniard et al., 2020). Meanwhile, the fit, or similarity between the original and extended products, can cause satiation (the drop in enjoyment due to the repetition of consumption), which can be differed by the categories (Lasaleta & Redden, 2018). In this context, the genre can enhance or damage the effect of similarity on the performance of the motion picture.

H1a: The similarity between the original novel and the film will positively affect the opening revenue.

H1b: The similarity between the original novel and the film will positively affect gross revenue.

H2a: The genre will increase (or diminish) the effect of the similarity between the original novel and the film on the opening revenue.

H2b: The genre will increase (or diminish) the effect of the similarity between

the original novel and the film on gross revenue.

On the other hand, the simple fact that the original novels and adopted films are similar cannot demonstrate that only the similarities could generate positive synergy. This is because not all consumers are sensitive to the similarity or fit with the parent brand. Only when consumers have previous experience and are highly associated with the brand do consumers become keen to changes in the adoption process (Broniarczyk & Alba, 1994; Hem & Iversen, 2009; Menon & Raghubir, 2003; Park, Milberg, & Lawson, 1991; Paul & Datta, 2013). However, when people have low background information, extension judgment is merely influenced by brand effect or brand awareness. So, checking the effect of the amount of buzz related to the originals would help analyze whether the similarity between books and motion pictures is effective in the box-office scores or whether the solid storyline that has been proven to succeed in the market is effective in the box-office scores.

H3a: The amount of buzz will substitute the effect of the similarity between the original novel and the film on the opening revenue.

H3b: The amount of buzz will substitute the effect of the similarity between the original novel and the film on gross revenue.

Chapter 3. Data and Variables

3.1. Data

For this research, the list of movies released between January 2015 and August 2022 was obtained. From the list, 77 films with opening and gross revenue comprise

the study's sample. The selected movies are under conditions that are not sequels of any other films and can be matched to a book. The selected movies are under the conditions that (1) the movie can be matched with a book, (2) the movie has no prequels, and (3) the movie summary can be obtained through SuperSummary.

The study uses each movie's movie and each book's summary for similarity measures. As a data source, this research used book summaries from SuperSummary (www.supersummary.com) because summaries on the websites are prepared by masters and Ph.D. students in English literature majors and have evident consistencies in the format. Also, movie summaries are scrapped from moviespoiler.com (www.themoviespoiler.com). The website provides summaries of the movies from the cinephiles. This website was used as a data source from various studies (e.g., Eliashberg et al., 2007; Kim et al., 2021). Other data related to the study, such as Opening, Gross, Genre, Actor, Director, and R, were collected from IMDB (www.imdb.com), which has the most prominent movie database. Lastly, since the pandemic caused enormous damage to the entire theatre market, it is reasonable to include whether the social distance policy is implemented, confirmed by the timeline offered by the Centers for Disease Control and Prevention (CDC).

3.2. Variables

The variables included in the research model are in **<Table 1>,** and the descriptive summary of them is shown in **<Table 2>**. As dependent variables, the paper used opening revenues and gross revenues of the focal films.

The *Similarity* was calculated through the bidirectional encoder representations from the transformation model, also known as the BERT. The BERT model was

introduced in 2019 by Devlin, Chang, Lee, and Toutanova in Google AI Language Lab. This model is pretrained through Toronto Book Corpus and Wikipedia. Unlike word2vec or GloVe, BERT does not need to label words by generating embeddings. Also, unlike other models, such as biLMs or GPT-1, using left-to-right or right-to-left, BERT considers the context of the text from both directions (right and left). Also, the BERT model can be modified for each research purpose. This paper used BERT to create the embeddings from summaries and calculate their similarity.

Mention variable was measured through pre-trained topic modeling methods. Mainly, using topic search, the study distinguished the relevant topics to "Book," "Novel," and "Read." The article's topics were divided into 30 categories and selected after the topic extraction.

Genre variables were measured as category variables. When the movie is included in a particular genre, the category variable is encoded as 1; otherwise, the variable is encoded as 0.

Some other factors are introduced as control variables. The *Pandemic* variable was encoded as 1 in the movie that was released after the occurrence of COVID-19. The *Actor* variable is used to estimate the effect of the casting power, as is the *Director* variable. Both variables are the sum of the box-office revenues of five recent movies which the actor or director's prior movies.

<Table 1> Variables and Description

Variables	**Description**	**Data Source**
Opening	Opening revenue of the focal movie in U.S market.	IMDB (www.imdb.com)
Gross	Gross revenue of the focal movie in U.S market.	

Similarity	Similarity between the original novel and the movie calculated through embeddings from BERT and c-TF-IDF formula	MovieSpoilers (www.themoviespoilers.com), SuperSummary (www.supersummary.com)
Genre	Action, Fantasy, History, Horror, Mystery, Sci-Fi, Romance (Base Category)	IMDB (www.imdb.com)
Pandemic	Movies released after social distance policy = 1, Others = 0	CDC (https://www.cdc.gov/museum/timeline/covid19.html)
Mention	The probability of topic buzz related to "book", "novel", and "read" in user reviews	IMDB (www.imdb.com)
Actor	Sum of box-office revenues of 5 recent movies which the focal movie's actors were cast members prior to the release of the focal movie ($)	
Director	Sum of box-office revenues of 5 recent movies which directed by the focal movie's director prior to the release of the focal movie ($)	
R	R = 1, Others (G, PG, and PG13) = 0	

<Table 2> Descriptive Statistics and Correlation

Variable	Mean	S.D.	Min.	Max.	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Opening	15.73	21.99	0.0064	123.40	1.00												
2. Gross	53.02	69.36	0.0099	364.00	0.902	1.000											
3. Similarity	0.65	0.15	0.1649	0.8602	0.203	0.235	1.000										
4. Action	0.27	0.45	0	1	-0.024	-0.005	-0.249	1.000									
5. Fantasy	0.20	0.41	0	1	0.091	0.074	-0.020	0.333	1.000								
6. Horror	0.13	0.34	0	1	0.178	0.092	-0.098	0.024	-0.103	1.000							
7. Mystery	0.38	0.37	0	1	0.001	-0.040	-0.269	-0.120	-0.145	0.238	1.000						
8. Romance	0.48	0.48	0	1	-0.140	-0.153	0.059	-0.206	-0.108	-0.203	-0.186	1.000					
9. Sci-Fi	0.40	0.39	0	1	0.103	0.071	-0.126	0.361	0.071	0.298	0.041	-0.293	1.000				
10. Pandemic	0.23	0.43	0	1	-0.122	-0.141	0.000	-0.063	-0.056	0.152	0.079	-0.213	-0.039	1.000			
11. Actor	855.71	653.46	79.15	3356.68	-0.102	-0.097	-0.006	0.117	0.182	-0.134	-0.052	-0.092	0.071	0.332	1.000		
12. Director	162.15	267.68	0	1358.63	0.314	0.338	-0.159	0.346	0.109	-0.102	-0.068	-0.174	0.020	0.073	0.037	1.000	
13 R	0.22	0.42	0	1	0.032	-0.057	-0.117	0.026	-0.273	0.260	0.178	-0.326	-0.025	0.002	-0.087	0.019	1.000

Chapter 4. Study 1

4.1. Method

To measure the plot similarity, the study used the BERT model to make embeddings, find the correlation between embeddings and calculate cosine similarity. Cosine similarity is generally used to measure the similarity between documents in natural language processing (Huang, 2008). Huang explains the cosine similarity with the formula below:

$$SIM_c(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{|\overrightarrow{t_a}| \times |\overrightarrow{t_b}|}$$

Given two documents $\overrightarrow{t_a}$, and $\overrightarrow{t_b}$ with m-dimensional vectors over the term set T = $\{t_1 \ldots t_m\}$, cosine similarity is calculated with the described formula. It results in a number that is non-negative and a value between [0,1]. The major advantage of cosine similarity is that it does not adhere to the document lengths. The numbers from the BERT model and cosine similarity are used as variable *Similarity*.

4.2. Model

Study 1 suggests the ordinary least square models that use both the opening revenue and the gross revenue of the focal movies as dependent variables. To find out the effect of similarity between the original work and the film, the suggested models would be:

$$Opening = \beta_{00} + \beta_{01}Similarity + \beta_{02}Pandemic + \beta_{03}R + \beta_{04}Action + \beta_{05}Fantasy + \beta_{06}Horror + \beta_{07}Mystery + \beta_{08}SciFi + \beta_{09}Actor + \beta_{010}Director + \beta_{011}R + \beta_{012}Action * Similarity +$$

$\beta_{013}Fantasy * Similarity + \beta_{014}Horror * Similarity + \beta_{015}Mystery * Similarity + \beta_{016}SciFi * Similarity$ (1)

$Gross = \beta_{10} + \beta_{11}Similarity + \beta_{12}Pandemic + \beta_{13}R + \beta_{14}Action + \beta_{15}Fantasy + \beta_{16}Horror + \beta_{17}Mystery + \beta_{18}SciFi + \beta_{19}Actor + \beta_{110}Director + \beta_{111}R + \beta_{112}Action * Similarity + \beta_{113}Fantasy * Similarity + \beta_{114}Horror * Similarity + \beta_{115}Mystery * Similarity + \beta_{116}SciFi * Similarity$ (2)

Formula (1) used the opening score of the film as a dependent variable, and formula (2) used the gross revenue of the film as a dependent variable.

4.3. Results

This study was conducted to verify the effect of the similarity between the story of the movie and the original story on the movie's box office performance. The formula (1) results are summarized in **<Table 3>.** The paper included the model without interaction effects on the left side of **<Table 3>,** as Model 1. The right side of **<Table 3>** shows the results of formula (1) with interaction effects.

The reported result shows the preliminary evidence that the relationship between the similarity and the box office revenue is positive ($\beta_{01} = 40.0312$, p-value = 0.016). As predicted, the effect of the similarity differs by some genres. Compared with the impact of similarity in the base category (Romance), the effect of the similarity lessened when the movies were Action films or Mystery films (β_{012} =-102.7748, p-value = 0.040; β_{015} =-83.8446, p-value = 0.067). On the

other hand, in the case of Horror films, the effect of similarity is enhanced (β_{014} =184.0078, p-value = 0.004). Estimation results for the control variables show that the director's power positively impacts the box office opening scores. The loss due to the pandemic is confirmed ($\beta_{03} = -7.4323$, p-value = 0.054; β_{010} =0.0261, p-value = 0.012). As a result of Model 2, we can confirm that H1a is accepted and that part of H2a was accepted.

<Table 3> Results for formula (1)

Variables	**Model 1**			**Model 2**		
	Estimate (S.E.)	Pr(>|z|)	Sig.[1)]	Estimate (S.E.)	Pr(>|z|)	Sig.[1)]
Constant	-13.0423 (11.269)	0.251		-17.4423 (15.072)	0.252	
Similarity	40.0312 (16.235)	0.016	**	47.5499 (22.711)	0.041	**
Action	-9.9148 (8.07)	0.129		44.6365 (27.451)	0.109	
Fantasy	7.0794 (6.382)	0.271		-17.5348 (20.987)	0.407	
History	-8.0933 (12.360)	0.515		11.8078 (61.880)	0.849	
Horror	15.7589 (7.833)	0.048	*	-84.3194 (35.254)	0.020	*
Mystery	2.4364 (6.643)	0.715		46.1748 (23.905)	0.058	+
Sci-Fi	5.8387 (6.650)	0.383	0.439	37.4019 (35.928)	0.302	
Pandemic	-8.95 (6.037)	0.092	+	-7.4323 (5.993)	0.054	+
Actor	-0.0004 (0.004)	0.927		0.00002 (0.004)	0.661	
Director	0.0367 (0.009)	0.000	**	0.0261 (0.01)	0.012	*
R	1.7079 (5.997)	0.777		-4.3995 (6.158)	0.478	

Action* Similarity				-102.7748 (48.900)	0.040	*
Fantasy* Similarity				46.9614 (35.604)	0.192	
History*Similarity				-27.4355 (104.404)	0.794	
Horror* Similarity				184.0078 (61.570)	0.004	**
Mystery* Similarity				-83.8446 (45.001)	0.067	+
Sci-Fi* Similarity				-52.0128 (62.948)	0.412	
Summary	R-Square: 0.305 Adj.R-Square: 0.187 F-Stat: 2.587(p-value = 0.00862)			R-Square: 0.457 Adj.R-Square: 0.300 F-Stat: 2.916(p-value = 0.00121)		

1) Significance: **p<0.01, *p<0.05

Formula (2) results are described in **<Table 4>**. Model 3 also confirmed that the similarity substantially impacts the gross revenue of films (β_{11} =144.1410, p-value = 0.059). However, in Model 3, the genre effect is significant only in Horror movies (β_{114} =436.3005, p-value = 0.036). Other genres (Action, Fantasy, History, Mystery, SciFi) have no significant interaction effects. Outcomes of the estimation show that H1b and partially H2b were accepted.

<Table 4> Results for formula (2) compared with Model 2

Variables	**Model 2**			**Model 3**		
	Estimate (S.E.)	Pr(>|z|)	Sig.[1)]	Estimate (S.E.)	Pr(>|z|)	Sig.[1)]
Constant	-17.4423 (15.072)	0.252		-43.9854 (49.772)	0.380	
Similarity	47.5499 (22.711)	0.041	**	144.1410 (74.999)	0.059	+
Action	44.6365 (27.451)	0.109		99.2396 (90.653)	0.278	
Fantasy	-17.5348 (20.987)	0.407		-71.2093 (69.306)	0.308	

History	11.8078 (61.880)	0.849		60.5629 (204.344)	0.768	
Horror	-84.3194 (35.254)	0.020	*	-84.3194 (35.254)	0.099	+
Mystery	46.1748 (23.905)	0.058		133.2390 (78.940)	0.097	+
Sci-Fi	37.4019 (35.928)	0.302		32.8378 (118.645)	0.783	
Pandemic	-7.4323 (5.993)	0.054		-25.3248 (19.792)	0.206	
Actor	0.00002 (0.004)	0.661		-0.0042 (0.013)	0.744	
Director	0.0261 (0.01)	0.012	*	0.1005 (0.033)	0.004	**
R	-4.3995 (6.158)	0.478		-25.7630 (20.335)	0.210	
Action* Similarity	-102.7748 (48.900)	0.040	*	-233.7591 (161.480)	0.153	
Fantasy* Similarity	46.9614 (35.604)	0.192		157.9635 (117.573)	0.184	
History*Similarity	-27.4355 (104.404)	0.794		-27.6978 (344.772)	0.936	
Horror* Similarity	184.0078 (61.570)	0.004	**	436.3005 (203.320)	0.036	*
Mystery* Similarity	-83.8446 (45.001)	0.067	+	-236.7754 (148.605)	0.116	
Sci-Fi* Similarity	-52.0128 (62.948)	0.412		-17.8447 (207.873)	0.932	
Summary	R-Square: 0.457 Adj.R-Square: 0.300 F-Stat: 2.916(p-value = 0.00121)			R-Square: 0.405 Adj.R-Square: 0.233 F-Stat: 2.358 (p-value = 0.00787)		

1) Significance: **p<0.01, *p<0.05, +p<0.1

Study 1 suggests the ordinary least square models that use both the opening revenue and the gross revenue of the focal movies as dependent variables. The suggested finding is the effect of similarity between the original work and the film. However, the similarity between original novels and motion pictures can be

interpreted in two ways: (1) fans or people who have the potential to become fans show their passion by towing the movie's box office performance that is similar to the original work and (2) the excellent storytelling of the original work is cogent between the audience. To explain why the movie is similar to the best-seller, original work, Study 2 would be needed.

Chapter 5. Study 2

In Study 1, the impact of plot similarity on the box office performance is verified through the direct correlation. However, there may be an alternative explanation for this correlation: the audience of the people might care how much the plot is well constructed. The original novel is selected to be adopted because its storyline is confirmed by many readers and tested by critics.

To distinguish the underlying mechanism, this paper conducted Study 2 by adding the buzz in user reviews. The fundamental idea of distinguishing the mechanism is that the fit with parent and extended brands concerns when consumers have enough knowledge about the parental brand (Broniarczyk & Alba, 1994). Miniard and his colleagues (2020) also pointed out that the fit between parent brands and extended brands counts when the association between them is highly accessible. So, if the customer power who knows the original work is large enough, the amount of buzz could alter the effect of similarity.

5.1. Methods

To figure out the intensity of the book's relativeness within the movie’s user reviews, the study recounted the method of calculating *Spoiler Intensity* suggested

by Ryoo et al. (2021). In Ryoo et al. (2021), the researchers used CTM (Blei & Lafferty, 2005) to uncover the topics in the user reviews and calculate the topic probability among the focal user reviews. On the other hand, as our documents (user reviews from IMDB) were analyzed by the pre-trained language model BERT, it is impossible to follow the same step to measure topic probability. Instead of using topic probabilities from the CTM model, the study used the BERTopic method (Grootendorst, 2022) to calculate the topic probability in the documents.

BERTopic is recently suggested by Grootendorst. The model combines transformers (BERT) to detect the topics in the documents and create embeddings, UMAP to reduce the dimensionality of the embeddings, Hierarchical DBSCAN (HDBSCAN) to cluster the topics, and calculate probability through class-based TF-IDF. Grootensorst used UMAP to maintain the overall structure of vectors. Then he used HDBCAN, which could allow the most stable clusters and deduce clusters by density. Then, the model used c-TF-IDF to analyze the probability. The formula for c-TF-IDF is introduced below:

$$c - TF - IDF_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_j^n t_j}$$

In the formula, t stands for the frequency of each word from each class i and is divided by w (the total number of words for each class i). Then, the total number of documents m is divided by the sum of frequency t (when the number of classes is n). Then, the extracted representation words are described in **<Table 5>**:

<Table 5> Topic List

Topic	Representation	Topic	Representation
1	dog, dogs, the, and, of, movie, it, this, to, was	16	circle, the, to, is, of, and, watson, that, mae, hanks
2	simon, gay, and, to, the, is, love, that, out, of	17	monster, conor, the, his, is, and, to, calls, of, that
3	out, movie, of, this, it, you, love, and,	18	the, tris, allegiant, divergent, series,

	watch, cry		and, to, of, in, that
4	tower, the, dark, books, to, of, and, it, this, is	19	stella, feet, cf, and, apart, fibrosis, the, cystic, to, in
5	poirot, the, branagh, murder, and, of, to, on, is, orient	20	frankenstein, victor, igor, the, and, of, mcavoy, to, radcliffe, his
6	emma, and, the, mr, of, to, jane, her, is, was	21	book, the, read, out, movie, of, was, this, and, it
7	martian, the, mars, to, and, of, damon, is, in, matt	22	shark, meg, the, statham, of, to, and, jaws, jason, it
8	asian, asians, rich, the, and, of, in, is, to, singapore	23	single, rebel, to, wilson, johnson, dakota, her, and, be, she
9	out, of, movie, it, this, and, was, the, great, is	24	wave, the, cassie, 5th, and, to, aliens, of, is, in
10	hanks, saudi, tom, the, of, to, in, his, is, and	25	bernadette, her, cate, blanchett, she, the, is, and, of, to
11	the, player, spielberg, oasis, of, and, to, in, ready, is	26	room, jack, and, the, larson, is, to, of, ma, in
12	clint, eastwood, his, macho, and, to, the, of, is, he	27	book, read, the, movie, out, it, of, was, and, to
13	close, wife, glenn, joan, the, her, is, and, joe, of	28	invisible, moss, man, the, cecilia, her, to, and, is, of
14	dune, the, and, villeneuve, is, of, to, in, it, that	29	zombies, zombie, pride, prejudice, and, the, of, darcy, to, it
15	the, rachel, train, her, she, and, girl, blunt, emily, to	30	billy, war, the, halftime, ang, of, lee, to, in, is

The model provides the search tool for a particular keyword. The study selected the keywords "Book," "Novel," and "Read." Then, use the topic probability from the overlapped topics such as topic 21(book, the, read, out, movie, of, was, this, and, it), topic 27(book, read, the, movie, out, it, of, was, and, to) and other topics highlighted above. Then, add the topic probability in the OLS formula as a variable *Mention*. The formula will be described below:

$$Opening = \beta_{20} + \beta_{21}Similarity + \beta_{22}Pandemic + \beta_{23}R + \beta_{24}Action + \beta_{25}Fantasy + \beta_{26}Horror + \beta_{27}Mystery + \beta_{28}SciFi + \beta_{29}Mention + \beta_{210}Actor + \beta_{211}Director + \beta_{212}R + \beta_{213}Action * Similarity + \beta_{214}Fantasy * Similarity + \beta_{215}Horror * Similarity + \beta_{216}Mystery * Similarity + \beta_{217}SciFi * Similarity + \beta_{218}Mention * Similarity \quad (3)$$

$$Gross = \beta_{30} + \beta_{31}Similarity + \beta_{32}Pandemic + \beta_{33}R + \beta_{34}Action + \beta_{35}Fantasy + \beta_{36}Horror + \beta_{37}Mystery + \beta_{38}SciFi + \beta_{39}Mention + \beta_{310}Actor + \beta_{311}Director + \beta_{312}R + \beta_{313}Action * Similarity + \beta_{314}Fantasy * Similarity + \beta_{315}Horror * Similarity + \beta_{316}Mystery * Similarity + \beta_{317}SciFi * Similarity + \beta_{318}Mention * Similarity \quad (4)$$

5.2. Results

In **<Table 6>**, the result of formula (3) is summarized and compared with the result of formula (1). Both models have opening scores as dependent variables. However, although *Mention* was joined in the OLS formula (formula (1)), the model's tendency does not seem to alter. Still, *Similarity* affects the movie's opening performance (β_{21}=40.4165, p-value = 0.097). Also, the interaction effect between movie genres and Similarity remains comparable. In the *Action* genre, the effect of Similarity shows a negative correlation with the opening scores (β_{213}=-103.6597, p-value = 0.040). Also, it does in the *Mystery* genre (β_{216}=-91.1199, p-value = 0.052), compared to the *Romance* genre.

Most importantly, there was no significant effect of *Mention* or interaction on the opening score (β_{29}=-91.1199, p-value = 0.052; β_{218}=-91.1199, p-value = 0.052). This shows that buzz does not influence the effect of the similarity between the original novel and the film on the opening revenue. With the results, H3a is rejected.

<Table 6> Results for formula (3) comparison with Model 2

Variables	**Model 2**			**Model 4**		
	Estimate (S.E.)	Pr(>|z|)	Sig.[1)]	Estimate (S.E.)	Pr(>|z|)	Sig.[1)]
Constant	-17.4423 (15.072)	0.252		--13.5203 (15.675)	0.392	

Similarity	47.5499 (22.711)	0.041	**	40.4165 (23.941)	0.097	+
Mention				-0.1014 (22.830)	0.449	
Action	99.2396 (90.653)	0.278		44.7843 (27.694)	0.111	
Fantasy	-71.2093 (69.306)	0.308		-19.6455 (21.253)	0.359	
History	60.5629 (204.344)	0.768		10.6643 (62.409)	0.865	
Horror	-84.3194 (35.254)	0.099	+	-82.5396 (35.611)	0.045	*
Mystery	133.2390 (78.940)	0.097	+	49.9954 (24.384)	0.045	
Sci-Fi	32.8378 (118.645)	0.783		30.3066 (36.895)	0.315	
Pandemic	-25.3248 (19.792)	0.206		-9.5293 (6.503)	0.148	
Actor	-0.0042 (0.013)	0.744		0.0003 (0.004)	0.938	
Director	0.1005 (0.033)	0.004	**	0.0323 (0.012)	0.008	**
R	-25.7630 (20.335)	0.210		-2.7782 (6.419)	0.667	
Action* Similarity	-102.7748 (48.900)	0.040	*	-103.6597 (49.901)	0.040	*
Fantasy* Similarity	46.9614 (35.604)	0.192		47.8670 (36.161)	0.169	
History*Similarity	-27.4355 (104.404)	0.794		-26.9063 (105.287)	0.799	
Horror* Similarity	184.0078 (61.570)	0.004	**	181.1901 (62.123)	0.005	**
Mystery* Similarity	-83.8446 (45.001)	0.067	+	-91.1199 (45.901)	0.052	+
Sci-Fi* Similarity	-52.0128 (62.948)	0.412		-41.0589 (64.467)	0.527	

Mention*Similarity				125.0587 (129.828)	0.375	
Summary	R-Square: 0.457 Adj.R-Square: 0.300 F-Stat: 2.916(p-value = 0.00121)			R-Square: 0.467 Adj.R-Square: 0.289 F-Stat: 2.625 (p-value = 0.00260)		

1) Significance: **p<0.01, *p<0.05, +p<0.1

On the other hand, the result of the formula (4), described in **<Table 7>,** shows that when the *Mention* variable is joined in the formula (2), the effect of *Similarity* vanishes, and the model loses its explanatory power (β_{31}=116.9643, p-value = 0.143). Also, *Mention* does not have any significance to explain the gross revenue of movies power (β_{39}=-292.8835, p-value = 0.258). Still, *Mention* could not alter *Similarity*. With the results, H3b is rejected.

<Table 7> Results for formula (4) and comparison with Model 3

Variables	**Model 3**			**Model 5**		
	Estimate (S.E.)	Pr(>|z|)	Sig.[1)]	Estimate (S.E.)	Pr(>|z|)	Sig.[1)]
Constant	-43.9854 (49.772)	0.380		-27.8248 (51.524)	0.591	
Similarity	144.1410 (74.999)	0.059	+	116.9643 (78.695)	0.143	
Mention				-292.8835 (256.351)	0.258	
Action	99.2396 (90.653)	0.278		98.9422 (91.029)	0.282	
Fantasy	-71.2093 (69.306)	0.308		-79.3100 (69.859)	0.261	
History	60.5629 (204.344)	0.768	+	57.5184 (205.139)	0.780	
Horror	-84.3194 (35.254)	0.099	+	-189.3876 (117.054)	0.111	
Mystery	133.2390 (78.940)	0.097		147.8058 (80.151)	0.070	
Sci-Fi	32.8378 (118.645)	0.783		3.3509 (121.276)	0.978	

Pandemic	-25.3248 (19.792)	0.206		-32.6472 (21.375)	0.132	
Actor	-0.0042 (0.013)	0.744	**	-0.0039 (0.013)	0.774	
Director	0.1005 (0.033)	0.004		0.1243 (0.039)	0.002	**
R	-25.7630 (20.335)	0.210		-27.3222 (21.090)	0.200	
Action* Similarity	-233.7591 (161.480)	0.153		-234.7021 (162.416)	0.154	
Fantasy* Similarity	157.9635 (117.573)	0.184		169.3702 (118.577)	0.159	
History*Similarity	-27.6978 (344.772)	0.936		-28.2199 (346.081)	0.935	
Horror* Similarity	436.3005 (203.320)	0.036	*	425.9206 (204.201)	0.041	*
Mystery* Similarity	-236.7754 (148.605)	0.116		-264.9473 (150.877)	0.084	+
Sci-Fi* Similarity	-17.8447 (207.873)	0.932		-57.8417 (64.188)	0.371	
Mention*Similarity				442.0647 (459.618)	0.340	
Summary	R-Square: 0.405 Adj.R-Square: 0.233 F-Stat: 2.358 (p-value = 0.00787)			R-Square: 0.421 Adj.R-Square: 0.228 F-Stat: 2.181 (p-value = 0.0123)		

Chapter 6. Conclusion

6.1. Discussion

The paper illustrated that similarity between adopted films and original novels has a positive effect on both the film's opening and gross scores. To figure the similarity, the paper collected data through web crawling and calculated cosine similarities through the pre-trained dataset by SentenceTransformers, known as the BERT model. Accorded to the expectation, the effect of similarity on movie

performance was significant. In addition, the paper divided genre variables to scan the effect of similarity diverged by the movie genre. Especially, Action movies and Mystery movies showed a negative effect of similarity on their opening revenues, compared to Romance Movies. On the other hand, the effect of similarity enhanced in Horror movies in both opening and gross revenues. The results correspond with previous literature about a brand extension that the fit between parent brands and extended brands matters (Broniarczyk & Alba, 1994; Hem & Iversen, 2009; Park, Milberg, & Lawson, 1991; Paul & Datta, 2013).

The paper also revealed the underlying mechanism by using the amount of mention related to the original works. In Study 2, the effect of customer buzz does not seem relative to movie performance (both opening performance and gross performance). Based on the prior academic findings (Broniarczyk & Alba, 1994; Menon & Raghubir, 2003; Miniard et al., 2020), the result may imply that most of the audience is focusing on the well-made story structure, not on comparability with the original works. This may also explain the reason that the effect of the director is significant ($\beta_{010} = 0.0261$, p-value = 0.012). People are more interested in movies that have intelligible storylines.

Unlike past research (Eliashberg et al.,2007; Eliashberg et al.,2014; Moon et al., 2022; Ryoo et al., 2021) that used the narratives' likeness to previous box office hits to greenlight the script's future success, the paper shifted the focus of the study to compare the original novel and movies to observe the success of adopted movies. As the book-to-film strategy has been maintained since the movie industry was established, the paper broadened the substantive topic of research that uses textual data to figure out the success factor. In addition, adding to Joshi and Mao, 2012, the paper focused on the book adaptation and used the similarity between the original

and the extension as a possible indicator of box-office success. The paper showed the corresponding result with Joshi and Mao (2012), but unlike the previous literature, it used quantitative measures to illustrate the effect of similarity. It also introduced genre factors to investigate the discrepancies that movies generically have. Lastly, the paper discovers the explanation behind the phenomenon. Referring to the study result and prior literature, it is possible to conclude that the general audience seeks the robustness of story structure that many readers confirm or have a significant reputation for choosing the movie. The approaches used in the paper specify the model by using real-market data and quantifying the similarity with natural language processing.

On the other hand, the paper also has some managerial implications. First, production companies should consider the properties of genres when they translate novels into movies. According to the results, although most genres maintained positive relationships with the effect of similarity, Action movies and Mystery movies showed negative synergy when the film version was provided in the theatre. In those genres, the directors and producers may twist the original plot and plant surprising factors to satisfy the audience. Second, modifications with a reasonable story structure may be allowed. Although the fans of original works might be disappointed when the plot's content is altered, the public without experiencing books regard the original as an indicator to avoid the risk of their choices.

6.2. Limitations and Directions for Future Research

The paper possesses limitations, but it can still suggest directions for future studies. To begin with, the research used summarized documents to figure out the

similarities. However, even though experts inspect the summarized texts with sufficient knowledge, some substances in the stories may need to be included or misinterpreted from full-length resources. It was due to low accessibility to original texts. Different from the movie scripts, it took work to get licenses for books. Future studies should use textual data reflecting the full content to complement shortcomings.

Next, the research did not consider the properties of the original novels, such as rank data or reviews of original books. According to Joshi and Mao (2012), book equity is one of the essential factors in estimating success. As the studies in the paper mainly focus on the similarity between films and books, future research may include the parent brand's awareness and compare its effect with the effect of fit.

Furthermore, the model can have limitations on applying other media industries. Recently, diverse kinds of intellectual properties have been adopted into movies or series. For example, Marvel transformed their original comics into movies and series in Disney+. As the distribution strategies and featured characteristics diverge from book-to-film cases, the delicate application of the model would be needed. Researchers may concentrate on the difference in each format for their next studies.

Finally, the research indirectly suggested that the effect of original fandom is mere. As the records that can indicate fandom power are insufficient, the research chooses to test the effect of buzz related to the original works. However, if there are any direct ways to estimate fandom's activities related to movie adoption, future studies should verify the effect more directly.

Bibliography

김은형 (2022, October 13), "이 책, 영화 만들면 재밌겠죠?"…'부산국제영화제' 간 출판사들, 한겨레, https://www.hani.co.kr/arti/culture/movie/1062442.html

Aaker, D. A., & Keller, K. L. (1990). Consumer evaluations of brand extensions. *Journal of Marketing*, *54*(1), 27-41.

Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of Marketing*, *67*(4), 103–117.

Berger, J., Sorensen, A. T., & Rasmussen, S. J. (2010). Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science*, *29*(5), 815-827.

Lafferty, J., & Blei, D. (2005). Correlated topic models. *Advances in neural information processing systems*, 18.

Boatwright, P., Basuroy, S., & Kamakura, W. (2007). Reviewing the reviewers: The impact of individual film critics on box office performance. *Quantitative marketing and economics*, *5*(4), 401-425.

Broniarczyk, S. M., & Alba, J. W. (1994). The importance of the brand in brand extension. *Journal of marketing research*, *31*(2), 214-228.

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, *29*(5), 944-957.

Chun, H. H., Park, C. W., Eisingerich, A. B., & MacInnis, D. J. (2015). Strategic benefits of low fit brand extensions: When and why?. *Journal of Consumer Psychology*, *25*(4), 577-595.

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.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter?—An empirical investigation of panel data. *Decision support systems*, *45*(4), 1007–1016.

Eliashberg, J., & Shugan, S. M. (1997). Film critics: Influencers or predictors?. *Journal of Marketing*, *61*(2), 68-78.

Eliashberg, J., Hui, S. K., & Zhang, Z. J. (2007). From story line to box office: A new approach for green-lighting movie scripts. *Management Science*, *53*(6), 881-893.

Eliashberg, J., Hui, S. K., & Zhang, Z. J. (2014). Assessing box office performance using movie scripts: A kernel-based approach. *IEEE Transactions on Knowledge and Data Engineering*, *26*(11), 2639-2648.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Hem, L. E., & Iversen, N. M. (2009). Effects of different types of perceived similarity and subjective knowledge in evaluations of brand extensions. *International Journal of Market Research*, *51*(6), 1-19.

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* (Vol. 4, pp. 9-56).

Joshi, A., & Mao, H. (2012). Adapting to succeed? Leveraging the brand equity of best sellers to succeed at the box office. *Journal of the Academy of Marketing Science*, *40*(4), 558–571.

Kim, J., Lee, Y., & Song, I. (2021). From intuition to intelligence: a text mining–based approach for movies' green-lighting process. *Internet Research*.

Kirmani, A., Sood, S., & Bridges, S. (1999). The ownership effect in consumer responses to brand line stretches. *Journal of Marketing*, *63*(1), 88–101.

Knapp, A. K., Hennig-Thurau, T., & Mathys, J. (2014). The importance of reciprocal spillover effects for the valuation of bestseller brands: introducing and testing a contingency model. *Journal of the Academy of Marketing Science*, *42*(2), 205-221.

Lasaleta, J. D., & Redden, J. P. (2018). When promoting similarity slows satiation: The relationship of variety, categorization, similarity, and satiation. *Journal of Marketing Research*, *55*(3), 446-457.

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, *70*(3), 74–89.

Liu, A., Liu, Y., & Mazumdar, T. (2014). Star power in the eye of the beholder: A study of the influence of stars in the movie industry. *Marketing Letters*, *25*(4), 385-396.

Menon, G., & Raghubir, P. (2003). Ease-of-retrieval as an automatic input in judgments: a mere-accessibility framework?. *Journal of Consumer Research*, *30*(2), 230-243.

Miniard, P. W., Alvarez, C. M., & Mohammed, S. M. (2020). Consumer acceptance of brand extensions: Is parental fit preeminent?. *Journal of Business Research*, pp. *118*, 335–345.

Moon, S., Jalali, N., & Song, R. (2022). Green-lighting scripts in the movie pre-production stage: An application of consumption experience carryover theory. Journal of Business Research, 140, 332-345.

Park, C. W., Milberg, S., & Lawson, R. (1991). Evaluation of brand extensions: The role of product feature similarity and brand concept consistency. Journal of Consumer Research, 18(2), 185–193.

Paul, S., & Datta, S. K. (2013). An empirical study of the effects of consumer knowledge on fit perception in brand extension success. *IUP Journal of Brand Management*, *10*(1), 37.

Ravid, S. A. (1999). Information, blockbusters, and stars: A study of the film industry. *The Journal of Business*, *72*(4), 463–492.

Ryoo, J. H., Wang, X., & Lu, S. (2021). Do spoilers really spoil? Using topic modeling to measure the effect of spoiler reviews on box office revenue. *Journal of Marketing*, *85*(2), 70-88.

Song, T., Huang, J., Tan, Y., & Yu, Y. (2019). Using user-and marketer-generated content for box office revenue prediction: Differences between microblogging and third-party platforms. *Information Systems Research*, *30*(1), 191-203.

Swaminathan, V., Fox, R. J., & Reddy, S. K. (2001). The impact of brand extension introduction on choice. *Journal of Marketing*, *65*(4), 1–15.

Völckner, F., & Sattler, H. (2006). Drivers of brand extension success. *Journal of Marketing*, *70*(2), 18-34.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, *74*(2), 133–148.

국문초록

OTT 시장이 발달하고 영화시장 규모가 커짐에 따라 영화 제작사들은 책이나 웹툰 등의 지식 재산권 인수를 통해 대본이 될수 있는 서사를 확보하고자 하고 있다. 본 논문은 다양한 접근법을 이용하여 베스트셀러 원작 영화의 성공요인으로 예상되는 원작과의 유사성의 영향을 조사한다. 이를 위해 2015년과 2022년 사이에 개봉한 소설 원작 영화 77편을 대상으로, 영화와 원작과의 유사성이 흥행 성적에 미치는 영향을 확인하였다. 연구 1에서는 원작과의 유사성이 흥행성적에 대체로 긍정적인 영향을 미치지만, 액션 영화와 미스터리 영화에서 부정적인 상관관계를 갖는 것을 확인하였다. 연구 2에서는 영화 리뷰 내 원작에 대한 언급을 OLS 모델에 추가하여 영화를 보는 관객층이 영화 관람 결정 과정에서 원작과의 유사성을 고려하는지 여부를 확인하였다. 그러나 영화 리뷰 내 언급량의 효과는 유사성의 효과를 대체하지 못하였고, 이를 통해 원작과의 유사성을 관람 시 실패 위험을 줄이기 위한 기준으로 사용하는 것을 확인하였다. 본 연구는 문화 상품에서의 브랜드 확장에 대한 학문적 근거를 제시하여 연구 영역을 넓히고, 제작사들에게 제작 시 스토리의 개연성과 장르 특성을 고려할 것을 제언한다.

키워드: 자연어처리, 브랜드확장, 책기반영화, 영화산업, 영화흥행

학번: **2021-23235**