


Article

Movie Title Keywords: A Text Mining and Exploratory Factor Analysis of Popular Movies in the United States and China

Xingyao Xiao ¹, Yihong Cheng ² and Jong-Min Kim ^{3,*} ¹ Graduate School of Education, University of California, Berkeley, CA 94720, USA; xiaoxg@berkeley.edu² Lynch School of Education and Human Development, Boston College, Chestnut Hill, MA 02467, USA; chengyz@bc.edu³ Statistics Discipline, Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN 56267, USA

* Correspondence: kimw@augsborg.edu or jongmink@morris.umn.edu

Abstract: Unprecedented opportunities have been brought by advancements in machine learning in the prediction of the economic success of movies. The analysis of movie title keywords is one promising but rarely investigated direction of study. To address this gap, we performed a text mining and exploratory factor analysis (EFA) of the relationships between movie titles and their corresponding movies' levels of success. Specifically, intragroup and intergroup analyses of 217 top hit movies in the United States and 245 top hit movies in China showed that successful movies in these two major movie markets with outstanding total lifetime grosses featured titles with similar and different patterns of most frequently used words, revealing useful information about viewers' preferences in these countries. The findings of this study will serve to better inform the movie industry in giving more economically promising names to their products from a machine-learning perspective and inspire further studies.



Citation: Xiao, Xingyao, Yihong Cheng, and Jong-Min Kim. 2021. Movie Title Keywords: A Text Mining and Exploratory Factor Analysis of Popular Movies in the United States and China. *Journal of Risk and Financial Management* 14: 68. <https://doi.org/10.3390/jrfm14020068>

Academic Editor: Shigeyuki Hamori
Received: 19 January 2021
Accepted: 4 February 2021
Published: 6 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: text mining; exploratory factor analysis (EFA); movie title keywords

1. Introduction

The movie industry is a risky one, as there is no guarantee of even a self-sustaining return on investment for the majority of its players. In the U.S., where the movie market is the most developed and leading the rest of the world, the Disney empire alone claimed 33 percent of the total box office revenue in 2009, without taking titles under its newly purchased 21st Century Fox into account (McClintock 2019). In a country with an exponentially expanding movie market such as China, hundreds of low-budget movies are struggling to earn their budgets back while a handful of top hits sweep the lion's share of the total box office revenue into their pockets (Zhang 2010).

What on Earth, then, made the big names so successful and the small potatoes so powerless in the movie industry? Is it possible at all for the vast majority of producers of low-budget movies around the globe, who are more vulnerable to risks (Pikowicz and Zhang 2006), to know, in advance, the likelihood of at least winning their costs back? A scant number of studies have been conducted to propose potential predictors. For instance, Chang and Ki (2005) found that budgets, among many other conjectured factors, played a significant role in the performance of successful movies in the U.S. However, their study failed to explain the exact mechanism and conflicted with Feng and Sharma (2016) study, which showed, with quantile regressions, that high-budget films were not always successful in the Chinese movie market. Others have postulated that action movies and sequel movies are more likely to be high grossing, as evidenced by the fact that more than half of the highest-grossing movies belong to these genres, including *Toy Story*, *The Dark Knight*, *Harry Potter* and *Transformers* (Pangarker and Smit 2013). However, as mentioned

before, these highest-grossing movies only account for the tip of the iceberg and hardly represent the long tail of the less-famous movies unknown to the public.

Fortunately, a new line of hope is brought by the mass analysis of text. Thanks to the advancement of machine learning, which enables the processing of extremely large datasets of texts that would otherwise be virtually impossible to interpret with human intelligence, more economically promising decisions can be made. For example, [Legoux et al. \(2016\)](#) showed that the success of a movie could be largely predicted by the critical reviews it received from distribution intermediaries, or exhibitors, owing to their machine-learning-powered study of 165,000 weekly theatre-level exhibitor notes in the U.S. Furthermore, [Du et al. \(2014\)](#) built their machine-learning-based model to predict a movie's box office revenue by analyzing comments about it on Sina Weibo, the most popular microblog site in China.

As powerful as reviews from exhibitors and viewers are as predictors of a movie's success, they are, nevertheless, factors that could not be controlled during its production. For movie producers who wish to do whatever they can to predict the success of their products before they are finished, the focus has to be turned to the investigation of texts at a more fundamental level. Many researchers have therefore chosen to analyze movie titles, an impactful element that largely determines viewers' first impressions about a movie that movie producers can relatively easily adjust at a low cost. For example, [Sood and Drèze \(2006\)](#) found that informative movie titles significantly boosted consumers' purchasing behaviors. Additionally, recent research conducted by [Bae and Kim \(2019\)](#) showed that the sequels that consumers responded to the most in South Korea were those that differed very slightly in their titles from their successful first episodes. In the winner-take-all movie market, where a tiny shift in movie titles can lead to enormous economic gain or loss ([Elberse and Oberholzer-Gee 2007](#)), it is therefore necessary to build upon these findings and better understand what it takes to create a successful movie title that is likely to bring decent box office revenue.

To this end, text mining was employed in this study to visualize and compare popular movie title keywords in the United States and China. Exploratory factor analyses ([Henson and Roberts 2006](#)) were conducted using the movie title keywords of popular movies in these countries from 2015 to 2019 to explore and identify potential factors that underlie popular movie title keywords in the United States and China. Similarities and differences between extracted factors of successful movie title keywords in the United States and China are discussed in terms of their implications for the movie industry and further research. More specifically, we were dedicated to answering the following research questions:

1. What are the factors underlying popular movie title keywords that contributed to the success of their corresponding movies from 2015 to 2019 in the United States and China?
2. What differences, if any, exist among the extracted factors of movie title keywords in the United States and their counterparts in China?
3. What implications about the contribution of movie title keywords to the success of movies, if any, can be drawn from the similarities and differences among the extracted factors of movie title keywords in the United States and their counterparts in China?

2. Text Mining and Exploratory Factor Analysis

In this study, we employed two major research methods to answer the research questions listed above, namely, text mining and exploratory factor analysis. Text mining is a method widely used in information-science-related research based on extracting significant and meaningful patterns from unstructured text documents. It was a perfect fit for this study, as we were interested in identifying patterns underlying a huge set of movie titles. Furthermore, to estimate the number of factors and indicators related to these underlying dimensions, we employed a multivariate statistical analysis called exploratory factor analysis (EFA).

2.1. Text Mining

Text is an extensive and highly diverse medium that transfers information across cultures. Due to the gigantic volume of text in various formats, the processing of text necessitates techniques that efficiently and effectively uncover underlying patterns. Text mining, a tool frequently applied in the fields of big data analyses and better known as statistical text mining when used in combination with other statistical methods, has proven to be one of such techniques. The interdisciplinarity of statistical text mining has become more powerful today, as information science researchers cooperate increasingly closely with statisticians to combine more statistical methods with text mining. For example, Tan (1999) proposed that text mining could synergize exceptionally well with the following statistical disciplines, at least: content analysis, clustering analysis, factor analysis and data mining techniques; information extraction; and machine learning.

Tan also summarized the framework of text mining into two phases: text refining and knowledge distillation, as shown in Figure 1. Text refining transforms free-form text documents into one specific intermediate form, and knowledge distillation continues to infer knowledge or patterns revealed by the intermediate form.

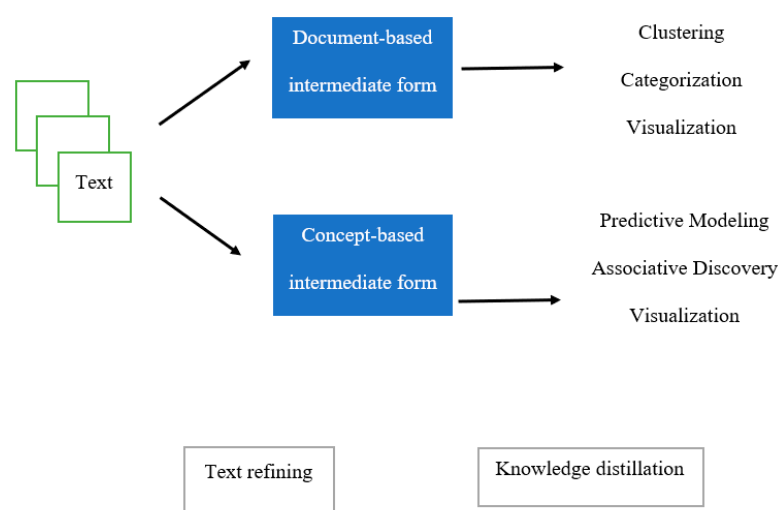


Figure 1. The framework of text mining.

2.2. Exploratory Factor Analysis

As computational power rapidly grows, factor analysis is becoming more and more popular as a powerful method of “reducing variable complexity to greater simplicity” (Kerlinger 1979). Factor analysis is a member of the family of statistical methods used to describe the relationships among many observed variables in terms of a few underlying but unobservable constructed factors. Researchers mainly focus on two factor analysis models: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Compared to CFA, EFA does not require a strong empirical or conceptual foundation to be established before the evaluation of models, as researchers do not have to specify the number of latent factors at the beginning (Rahmawati et al. 2017). A model of an exploratory factor analysis model is shown in Figure 2.

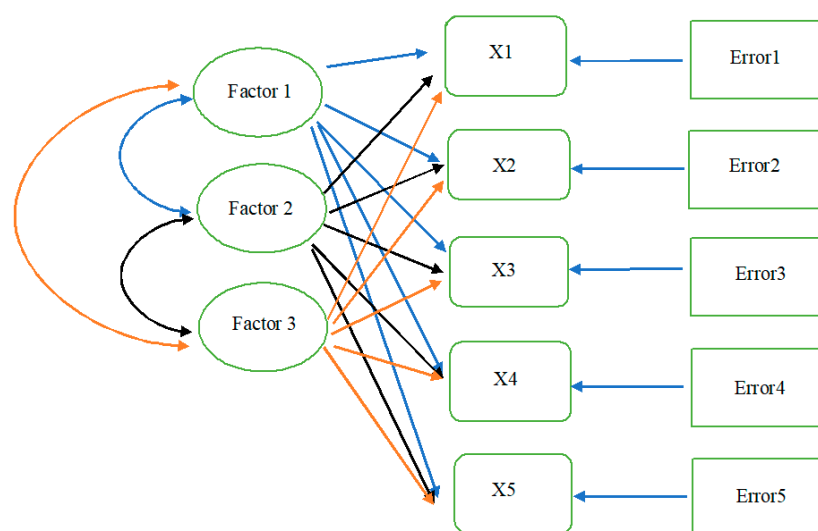


Figure 2. The exploratory factor analysis (EFA) model.

An EFA consists of the following four kinds of elements:

- *Measured/observed variables* are represented in the diagram by rounded rectangles.
- *Latent/unobserved factors/constructs* are represented in the diagram by ovals.
- *Error terms* are also represented by rectangles that are associated with the measured variables. They are latent variables as well. Each variable that is predicted (has a directional arrow pointing to it) must be associated with a residual variable.
- *Paths*, specifically, directional paths of influence or prediction, are represented in the diagram by lines with arrows pointing in a given direction.

Stapleton (1997) proposed that EFA is a useful tool for estimating the minimum number of latent hypothetical factors within a larger number of variables. Stevens (1996) also outlined the following tenets of the exploratory theory:

- Heuristics;
- Determine the number of factors;
- Determine whether the factors are correlated or uncorrelated;
- The variables are free to load on all of the factors.

Since the basic single factor model has very few real-world applications, we focus on the more commonly utilized k -factor analysis model, which can use any number of latent factors to describe relationships in a dataset. If we observe a set of observed variables represented by (x_1, x_2, \dots, x_q) and assume these are linked to a set of unobserved latent variables represented by (f_1, f_2, \dots, f_k) , where $k < q$ since the number of latent factors must be less than the total number of observed variables x , we can obtain a set of regression model equations where (x_1, x_2, \dots, x_q) are the observed variables linked to the unobserved latent variables (factors) (f_1, f_2, \dots, f_k) .

From a standard regression model perspective, we consider the λ values of the equations to act as regression coefficients for our manifest x variables. More specifically, however, from a factor analysis perspective, these λ s are called factor loadings and represent and show the nature and magnitude of the relationships between each x and each latent factor f (Everitt and Hothorn 2011).

3. Data Analysis and Results

In this study, we performed EFA to extract factors from the titles of popular movies in the United States and China released between 2015 and 2019 based on ratings provided by IMDb to help us to better identify and make sense of meaningful patterns in them. In this section, we present a series of charts that resulted from this process.

3.1. Descriptive Statistics

Firstly, Table 1 shows a summary of movie title keywords, their frequencies, and their corresponding movies' total lifetime grosses for the 50 most popular movies in the United States and China. In both countries, it turns out that “man” and “movie” are among the three most frequently appearing keywords. As the most frequently appearing movie title keyword in both countries, “man” is a major, most-favored keyword likely because of the immense popularity of superhero movies across the two countries. The second most frequently appearing keyword in the United States, “star”, undoubtedly reveals the success of the *Star Wars* series in terms of both its total lifetime gross and its fame in general, as marked by its relative success in the Chinese movie market despite the cultural differences. The third most frequently appearing keyword, “movie”, likely indicates the popularity of the common format of naming a movie as “xxx, the movie”.

Table 1. Descriptive statistics of the 50 most popular movies in the United States and China.

Rank	Movie Title Keywords	U.S.		China		
		Freq.	Total Lifetime Gross (US Dollars)	Movie Title Keywords	Freq.	Total Lifetime Gross (US Dollars)
1	man	11	1,503,705,642	man	12	1,286,977,547
2	star	11	3,093,464,851	love	11	351,111,366
3	movie	10	963,140,165	movie	11	301,770,391
4	home	9	1,050,200,471	mr	11	1,036,840,650
5	wars	9	2,723,848,650	big	8	318,827,177
6	day	7	357,568,154	detective	7	912,804,521
7	last	7	925,398,857	king	7	813,778,125
8	night	6	282,339,798	one	7	377,716,097
9	spider	6	914,974,535	star	7	349,263,730
10	war	6	1,392,994,823	legend	6	468,295,230
11	chapter	5	553,686,725	secret	6	364,634,367
12	death	5	202,382,769	wars	6	300,914,509
13	la	5	205,835,542	world	6	672,948,134
14	men	5	528,461,049	bears	5	371,029,415
15	big	4	172,822,602	boonie	5	371,029,415
16	black	4	869,138,374	doraemon	5	174,366,043
17	book	4	517,647,590	dragon	5	341,165,601
18	christmas	4	198,788,539	last	5	350,106,007
19	dark	4	247,747,451	lost	5	590,933,626
20	family	4	258,246,744	men	5	408,509,297
21	halloween	4	326,563,058	new	5	158,104,607
22	life	4	609,627,658	war	5	771,637,776
23	perfect	4	345,710,380	adventure	4	132,342,875
24	secret	4	698,987,488	battle	4	309,207,013
25	story	4	1,179,982,844	dad	4	88,345,839
26	age	3	565,698,652	Go	4	166,338,613
27	american	3	436,388,747	miss	4	112,922,275
28	angel	3	181,241,969	mission	4	378,389,513
29	avengers	3	1,996,194,350	part	4	137,159,151
30	breaks	3	230,839,314	son	4	82,829,886
31	captain	3	908,835,188	spider	4	378,178,062
32	creed	3	248,248,852	storm	4	394,405,176
33	daddy	3	252,148,518	story	4	129,291,307
34	dog	3	129,295,337	three	4	91,607,441
35	dragon	3	267,744,775	wild	4	97,283,162
36	evil	3	175,269,310	wolf	4	1,055,812,724
37	fallen	3	549,274,456	angel	3	241,283,819
38	fantastic	3	435,574,980	avengers	3	1,213,969,174
39	fifty	3	381,156,240	back	3	431,516,633
40	first	3	188,332,780	black	3	160,386,346

Table 1. Cont.

Rank	Movie Title Keywords	U.S.		China		
		Freq.	Total Lifetime Gross (US Dollars)	Movie Title Keywords	Freq.	Total Lifetime Gross (US Dollars)
41	furious	3	752,972,340	blue	3	208,408,638
42	go	3	208,972,108	book	3	238,475,673
43	hotel	3	370,335,684	captain	3	751,111,870
44	jumanji	3	598,471,374	chinese	3	97,643,298
45	jungle	3	770,377,961	comedy	3	129,133,564
46	king	3	693,313,247	Conan	3	63,568,995
47	lego	3	340,838,447	Dark	3	173,347,863
48	little	3	100,338,715	Fallen	3	322,947,935
49	madea	3	193,782,960	Fu	3	420,394,532
50	next	3	281,547,967	Furious	3	984,718,005

The bar charts of the top 10 movie title keywords in the United States (Figure 3) and the top 10 movie title keywords in China (Figure 4) further confirm that the success of “man” as a movie title keyword stems from the success of movies about superheroes such as *Ant-Man*, *Spider-Man*, and *Batman*. In addition, comparing Figure 3 with Figure 4, it is clear that there is a difference between the U.S. and China in terms of the frequency of the appearance of man in the most popular movie titles.

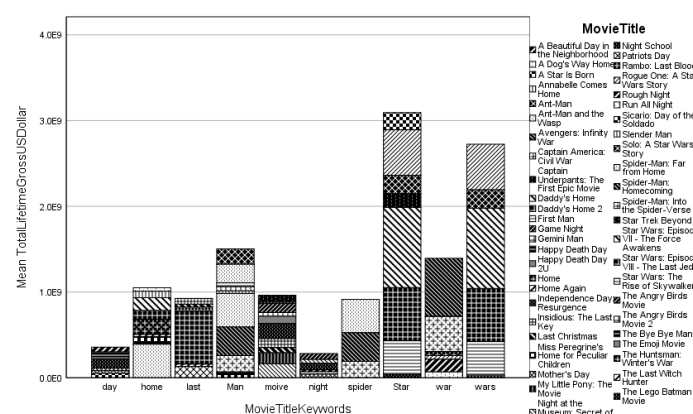


Figure 3. Top 10 most popular movie title keywords and their corresponding movie titles in the United States.

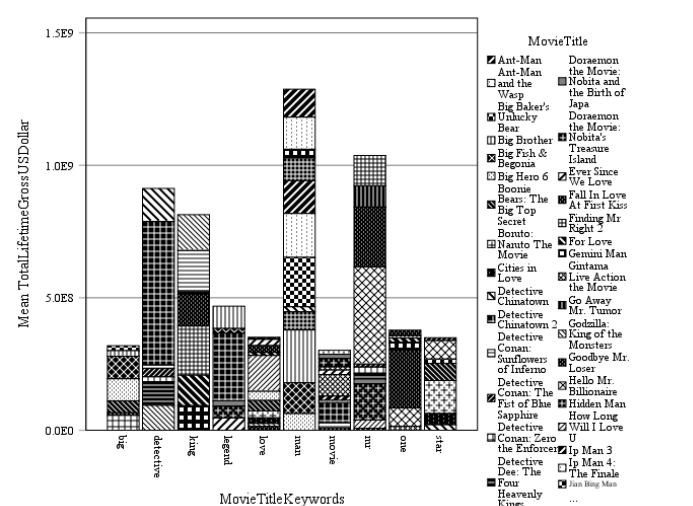


Figure 4. Top 10 most popular movie title keywords and their corresponding movie titles in China.

Next, we compared the total lifetime grosses of the movies released between 2015 and 2019 in the United States and China. Apparently, based on the 217 movies we observed that were released in the United States and the 245 movies we observed that were released in China, the movies from the United States outnumbered their Chinese counterparts according to all the five summary statistics we chose, including the mean, standard deviation, 25th percentile, 75th percentile, and interquartile range, as shown in Table 2.

Table 2. Summary statistics of the total lifetime grosses in the United States and China.

		Ob.	Mean	Std. Dev.	25th Perc.	75th Perc.	Interquartile Range (IQR)
Total	U.S.	217	144,610,410	154,714,862	45,184,072	173,257,905	128,073,833
Gross	China	245	81,974,624	103,354,089	19,572,917	108,402,274	88,829,357

3.2. Data Visualization

In order to offer a more visually direct representation of the movie title keywords, we present the word clouds formed by movie title keywords in the United States (Figure 5) and by movie title keywords in China (Figure 6). In both figures, the movie title keywords that appear more frequently in the dataset show up in bigger fonts. Our previous discussion still stands regarding the success of man, star, and movie as outstanding movie title keywords.



Figure 5. Word cloud of movie title keywords in the United States.



Figure 6. Word cloud of movie title keywords in China.

3.3. Exploratory Factor Analysis Result (U.S.)

An exploratory factor analysis of the top 50 popular movie title keywords (MTK) from 2015 to 2019 in the United States was performed on 217 movies. Prior to running the analysis with RStudio, the data were screened by examining, for each MTK, descriptive statistics, interitem correlations, and possible univariate and multivariate assumption violations. From this initial assessment, all the variables were found to be interval like, variable pairs appeared to follow bivariate normal distributions, and all the MTKs were independent of one another. Because our sample size of 217 movies in this data analysis was large, the MTK to movie ratio was deemed adequate. Prior to conducting the EFA, the adequacy of the input data was confirmed by Bartlett's sphericity test and the matrix determinant. The Bartlett test of sphericity result was significant ($\chi^2 = 1859.17, p < 0.001$), suggesting that the correlation matrix was significantly different from the identity matrix and that the variables were correlated, which supported the data reduction. The correlation matrix determinant of 0.00003 was greater than the necessary value of 0.00001. Hence, the data were adequate for the EFA, and multicollinearity was not a problem for these data.

Furthermore, EFA with principal axis factoring was used in extracting the factors to be retained. As indicated by the scree plot (Figure 7), three factors to be extracted had eigenvalues greater than one. Therefore, these factors were extracted in the first set of analyses.

Next, in order to improve the interpretability of the extracted factors, both Varimax and Promax Rotations were performed. The results were compared and indicated no significant differences. Therefore, to simplify the interpretation of the extracted factors (IBM Knowledge Center 2020), Varimax Rotation with Kaiser Normalization and the structure coefficients from the Varimax Rotation are presented in Table 3. As shown in the table, the structure coefficients are reasonable but not notably large in magnitude, owing to the relatively small amount of variance explained by this structure.

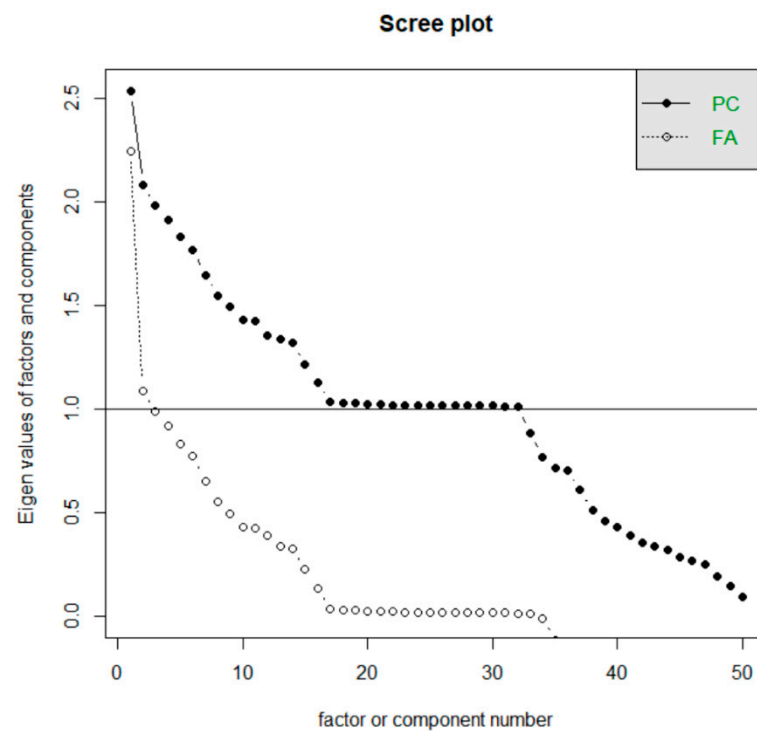


Figure 7. Scree plot indicating the number of retained factors for movies in the United States.

Table 3. Three-factor rotated matrix.

Movie Title Keyword	FA2	FA1	FA3
Man		−0.177	
Star	−0.484	0.65	−0.422
Movie	1.06	0.536	−0.179
Home		−0.187	
Wars	−0.505	0.69	−0.442
Day			0.197
Last	−0.129	0.117	−0.109
Night			
Spider		−0.159	
War			
chapter			
Death		0.212	0.563
La			
Men			0.119
Big			
Black		0.204	0.527
Book			
Christmas			
Dark			
Family			
Halloween			
Life			
Perfect			
Secret			
Story	−0.311	0.424	−0.27
Age			
American			
Angel		0.26	0.623

Table 3. Cont.

Movie Title Keyword	FA2	FA1	FA3
Avengers			
Breaks			
Captain	0.273		
Creed			
Daddy		−0.109	
Dog			
Dragon			
Evil			
Fallen			0.194
Fantastic			
Fifty			
First	0.256		
Furious			
Go			
Hotel			
Jumanji			
Jungle			
King			
Lego	0.391	0.193	
Little	0.136		
Madea		−0.103	
Next			

As shown in Table 4, the first factor extracted from the analysis (FA2) was “Family Movie”, which included the following MTKs: little, first, captain, lego, and movie. The corresponding movies can also be found in the table, such as *My Little Pony: The Movie*; *Captain Underpants: The First Epic Movie*; and *The Lego Batman Movie*.

Table 4. Correlations of extracted factors (U.S.).

	FA2	FA1	FA3
FA2	1	0.301196	−0.07575
FA1	0.301196	1	−0.20568
FA3	−0.07575	−0.20568	1

The second factor extracted from the analysis (FA1) was “Sequels”, which included the following MTKs: movie, star, story, and last. A few examples of their corresponding movies are *Star Wars: Episode VII—The Force Awakens*; *Toy Story 4*; and *Star Wars: Episode VIII—The Last Jedi*.

The third factor extracted from the analysis (FA 3) was “Horror and Thriller Movie”, which included the following MTKs: angel, death, black, day, fallen, and men. The corresponding movies are *The Woman in Black 2: Angel of Death*; *Happy Death Day*; and *X-Men: Apocalypse*.

Moreover, Table 5 shows the correlations between the extracted factors by using Promax Rotation. Two extracted factors, Family Movie and Sequels, were positively correlated with each other ($\text{corr}(\text{FA2}, \text{FA1}) = 0.30$). On the contrary, Horror and Thriller Movie was found to be negatively correlated with Family Movie and Sequel ($\text{corr}(\text{FA3}, \text{FA2}) = -0.08$ and $\text{corr}(\text{FA3}, \text{FA1}) = -0.21$). The three-factor varimax with a rotated structure mirrored the analysis of the top 50 popular MTKs in the United States. As shown in Table 3, the top 50 popular MTKs were distributed among the three factors in patterns that indicated distinct dimensions that could be used to further analyze the potential factors that lead to the success of movies in the United States. This will be discussed further in Section 4.

Table 5. Structure of the movie title keywords resulting from factor analysis in the U.S.

Factor	Movie Title Keyword	Movie
FA2: Family Movie	little	My Little Pony: The Movie Little Little Women
	first	Captain Underpants: The First Epic Movie The First Purge First Man
	captain	Captain Marvel Captain America: Civil War Captain Underpants: The First Epic Movie
	lego	The Lego Batman Movie The Lego Ninjago Movie The LEGO Movie 2: The Second Part
	movie	The Peanuts Movie The Angry Birds Movie The Lego Batman Movie The Lego Ninjago Movie The LEGO Movie 2: The Second Part The Emoji Movie Captain Underpants: The First Epic Movie My Little Pony: The Movie The Angry Birds Movie 2 Teen Titans GO! To the Movies The SpongeBob Movie: Sponge Out of Water
FA1: Sequels	wars	Star Wars: Episode VII—The Force Awakens Rogue One: A Star Wars Story Star Wars: Episode VIII—The Last Jedi Solo: A Star Wars Story Star Wars: The Rise of Skywalker Underworld: Blood Wars
	star	Star Wars: Episode VII—The Force Awakens Rogue One: A Star Wars Story Star Trek Beyond The Star Solo: A Star Wars Story A Star Is Born Star Wars: The Rise of Skywalker Star Wars: Episode VIII—The Last Jedi
	story	Solo: A Star Wars Story Rogue One: A Star Wars Story Toy Story 4
	last	Star Wars: Episode VIII—The Last Jedi Last Christmas The Last Witch Hunter Transformers: The Last Knight Insidious: The Last Key Rambo: Last Blood

Table 5. Cont.

Factor	Movie Title Keyword	Movie
FA3: Horror and Thriller Movie	angel	The Woman in Black 2: Angel of Death Alita: Battle Angel Angel Has Fallen
	death	Happy Death Day Happy Death Day 2U The Woman in Black 2: Angel of Death Maze Runner: The Death Cure Death Wish
	black	The Woman in Black 2: Angel of Death Men in Black: International Black Panther Black Mass
	day	Happy Death Day Patriots Day Sicario: Day of the Soldado A Beautiful Day in the Neighborhood Happy Death Day 2U Independence Day: Resurgence Mother's Day
	fallen	Angel Has Fallen London Has Fallen Jurassic World: Fallen Kingdom
	men	X-Men: Apocalypse Pirates of the Caribbean: Dead Men Tell No Tales Men in Black: International X-Men: Dark Phoenix What Men Want

3.4. Exploratory Factor Analysis Result (China)

Similarly, an EFA was carried out on 245 movies to determine the number of factors extracted from the MTKs of the 50 most popular movies released from 2015 to 2019 in China. Before conducting the EFA, the adequacy of the input data was confirmed by Bartlett's sphericity test and the matrix determinant, which indicated that the data were suitable for EFA and that there was a sufficient correlation between the variables to proceed with the analysis. A total of five factors with eigenvalues greater than 1.00 were extracted based on the scree plot (see Figure 8). To better interpret the meaning behind the extracted factors, both Varimax and Promax Rotations were performed and indicated no significant difference between the results. All the extraction methods yielded the same structure, and the results of the principal factor solution with Varimax Rotation are reported in Table 6.

The five-factor Varimax rotated structure mirrored the underlying factors of the MTKs in the U.S. The first factor extracted from the data (FA1) was "Family Movie", which included the following MTKs: new, dad, and son. Clearly, a family movie theme can be seen in these three MTKs. Moreover, by inspecting their corresponding movies, namely, *New Happy Dad and Son 2: The Instant Genius*; *Crazy New Year's Eve*; and *Dad, Where Are We Going 2*, the underlying factor becomes even more visible.

The second factor extracted (FA2) was "Comedy", including these MTKs: Boonie Bears, secret, and adventure. The corresponding movies were as follows: *Fantastica: A Boonie Bears Adventure*; *The Secret Life of Pets 2*; *New Happy Dad and Son 3: Adventure in Russia*. Clearly, these movies belong to the genre of comedy.

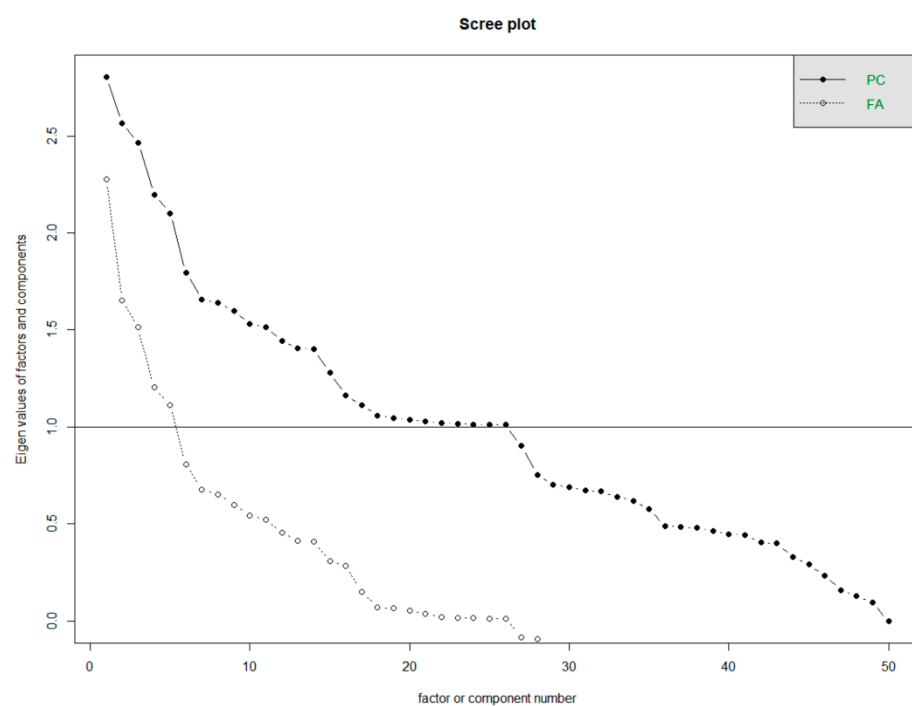


Figure 8. Scree plot indicating the number of correlated factors for movies in China.

Table 6. The structure of movie title keywords resulting from the factor analysis (China).

Factor	Movie Title Keyword	Movie
Family Movie	new	New Happy Dad and Son 2: The Instant Genius Crazy New Year's Eve New Happy Dad and Son 3: Adventure in Russia
	dad	New Happy Dad and Son 2: The Instant Genius Dad, Where Are We Going 2 New Happy Dad and Son 3: Adventure in Russia Dad's Holiday
	son	New Happy Dad and Son 2: The Instant Genius Father and Son Seventh Son (Action) New Happy Dad and Son 3: Adventure in Russia
Comedy	Boonie Bears	Fantastica: A Boonie Bears Adventure Boonie Bears 5 Boonie Bears: Blast into the Past Boonie Bears Winter Boonie Bears: The Big Top Secret
	Secret	The Secret Life of Pets 2 Secret Superstar The Secret Life of Pets Boonie Bears: The Big Top Secret Kingsman: The Secret Service Night at the Museum: Secret of the Tomb
	Adventure	New Happy Dad and Son 3: Adventure in Russia The Adventure of Afanti Doraemon: Great Adventure in the Antarctic Kachi Kochi Fantastica: A Boonie Bears Adventure

Table 6. Cont.

Factor	Movie Title Keyword	Movie
Detective	Detective	Detective Chinatown Detective Conan: Sunflowers of Inferno Detective Conan: The Fist of Blue Sapphire Detective Conan: Zero the Enforcer Detective Chinatown 2 Detective Dee: The Four Heavenly Kings More Than Blue (Romance)
		Detective Conan: The Fist of Blue Sapphire A Witness Out of the Blue Detective Conan: Zero the Enforcer
	Conan	Detective Conan: Sunflowers of Inferno Detective Conan: The Fist of Blue Sapphire The Hunger Games: Mockingjay—Part 1
Action/Crime	part	The Hunger Games: Mockingjay—Part 2 A Chinese Odyssey: Part Two—Cinderella A Chinese Odyssey: Part Three Three Seconds/The Informers
	three	A Chinese Odyssey: Part Three Three Three Billboards Outside Ebbing, Missouri Peppa Celebrates Chinese New Year
	Chinese	A Chinese Odyssey: Part Two—Cinderella A Chinese Odyssey: Part Three Fantastica: A Boonie Bears Adventure
Sequels	Boonie Bears	Boonie Bears 5 Boonie Bears: Blast into the Past Boonie Bears Winter Boonie Bears: The Big Top Secret Bride Wars
	wars	Star Wars: Episode VIII—The Last Jedi Rogue One: A Star Wars Story Solo: A Star Wars Story Star Wars: Episode VII—The Force Awakens Star Wars: The Rise of Skywalker Star Wars: Episode VII—The Force Awakens
	star	Star Wars: Episode VII—The Force Awakens Star Wars: Episode VII—The Force Awakens Star Trek Beyond Star Wars: Episode VIII—The Last Jedi Rogue One: A Star Wars Story Solo: A Star Wars Story Star Wars: Episode VIII—The Last Jedi Star Wars: The Rise of Skywalker The Star New Happy Dad and Son 2: The Instant Genius
	dad	New Happy Dad and Son 3: Adventure in Russia Dad, Where Are We Going 2 Dad's Holiday

The third factor extracted (FA3) was “Detective”, including the following MTKs: detective, blue, and Conan. Their corresponding movies were *Detective Chinatown*; *Detective Conan: The Fist of Blue Sapphire*; and *Detective Conan: Zero the Enforcer*.

The fourth factor extracted (FA4) was “Action/Crime”. The manifested MTKs were part, three, and Chinese. Our extraction was justifiable by the content of their corresponding

movies, including *The Hunger Games: Mockingjay—Part 1*; *Three Billboards Outside Ebbing, Missouri*; and *A Chinese Odyssey: Part Three*.

The fifth factor extracted (FA5) was “Sequels”. The observed MTKs were Bonnie Bears, wars, star, and dad. Movie sequels such as *Star Wars: Episode VII—The Force Awakens*; *Star Trek Beyond*; *Star Wars: Episode VIII—The Last Jedi*; *Rogue One: A Star Wars Story*; and *Solo: A Star Wars Story* are a good example of manifesting this factor.

The correlations between the extracted five factors based on Promax Rotation are summarized in Table 7. The implications underlying the extracted five factors are discussed in the next section.

Table 7. Correlations of extracted factors (China).

	FA1	FA2	FA4	FA5	FA3
FA1	1.00	−0.02	−0.01	0.02	−0.02
FA2	−0.02	1.00	0.01	−0.04	0.02
FA4	−0.01	0.01	1.00	−0.01	0.01
FA5	0.02	−0.04	−0.01	1.00	−0.03
FA3	−0.02	0.02	0.01	−0.03	1.00

4. Discussion

Our results indicate many interesting patterns of successful MTKs and their corresponding popular movies that we have not elaborated in detail so far. This section is dedicated to describing these patterns and discussing their implications. We begin by presenting our intragroup analyses of popular movies in the U.S. followed by intragroup analyses of popular movies in China and proceed to intergroup analyses of popular movies in both countries.

4.1. Intragroup Analysis of Movie Title Keywords in the United States

“Family Movie” is, unsurprisingly, an important factor of popular movies in the U.S. As a genre that can be enjoyed by viewers of all ages, family movies, or animated family movies to be specific, are innately attractive to a bigger population of consumers than other types of movie. Given the presence of the Motion Picture Association movie rating system in the U.S., family movies are especially more likely to be freely accessed by a greater population of consumers that most other rated movies with certain age restrictions cannot possibly compare to. As long as parents across the U.S. continue to take their children to the movies as a treat, movies made for the whole family will most likely continue to have a uniquely important role to play in the movie market. As such, for those who are planning to make a widely welcomed movie, it is never a bad idea to make something that can make a 6-year-old laugh alongside a 60-year-old.

More specifically, it is an interesting pattern that many of them share one structure of naming, or a concise combination of the names of the characters featured in the movie followed by a colon and “the Movie”, such as *My Little Pony: The Movie*. This structure is advantageous in many ways, as it allows the titles of family movies to efficiently convey to potential viewers who they can expect to see. In cases where the characters featured are adaptations from well-established iconic characters, such titles can also greatly attract the attention of audiences who love these characters by revealing, to a large extent, what a viewer can expect to see. Similarly, the success of *The Lego Batman Movie* indicates the importance of emphasizing the major character featured in the shortest length possible. In just four words, this title can distinguish itself from other, more-serious Batman movies for older viewers only, such as *The Dark Knight*, sending a clear and strong message to parents anxiously choosing from an array of movies that this one is a safe choice for the whole family.

“Sequels” is also not an unexpected factor extracted from the MTKs of popular movies in the U.S., largely because of the success of the Star Wars sequel. An all-time classic that sets a milestone of success challenged by few competitors in the entire history of the movie

industry, *The Star Wars* has been favored by Americans for decades. Additionally, while some viewers might prefer some episodes over others depending on when they were born, how acceptive they are of less-advanced visual effects, and the plots of the stories in general, *The Star Wars* carries a culture that stands the test of time and guarantees a sound return on investment. Meanwhile, *The Star Wars* sequels were noticeably not the only sequel movies among the 217 movies we analyzed. Although not as phenomenally well received as *The Star Wars*, sequels such as *Transformers*, *The Woman in Black*, *Underworld*, and *Happy Death Day* are still great examples of how impactful the titles of sequel movies can be in terms of establishing and preserving brand effects.

The third extracted factor, “Horror and Thriller Movie”, covers a more diverse range of movie titles. Although these movies cover a variety of topics, their titles are constantly made using religious elements. Frequently appearing keywords such as “angels” and “death” indicate that the movie producers know exactly how to create thrillingly popular movie titles in the U.S. by taking advantage of a stunning similarity shared by the majority of Americans. Given that nearly 8 in 10 Americans believe in the existence of angels despite differences in their religious beliefs ([The Associated Press 2011](#)), using the image of angels and death in the titles of horror and thriller movies proves to be the best way of entertaining the highly diverse viewers in the U.S. with a common factor that almost everyone can find some resonance with, albeit in slightly different ways.

4.2. Intragroup Analysis of Movie Title Keywords in China

Similar to in the U.S. market, “Family Movies” and “Sequels” are two extracted factors of popular movies in China as well, although the exact keywords involved differ with the exception of “star” and “wars”. Once again, almost all the movies under the “Sequels” category are *Star Wars* episodes, proving the extraordinary success of *The Star Wars* sequels at an international level. Two Chinese original sequels, *New Happy Dad and Son* and *Dad, Where Are We Going*, stood out without doubt, given the popularity of these icons as TV series. As its name suggests, *New Happy Dad and Son* depicts a dad and his son living happily together. Considering that the first *Happy Dad and Son* episodes were aired about 25 years ago, the *New Happy Dad and Son* movies are, indeed, dedicated to a new generation of dads and sons. On the other hand, *Dad, Where Are We Going* is based on a reality show of how a group of fathers and sons from big cities adapt to life in the countryside. Overall, the success of these movies indicates that Chinese audiences truly appreciate and value harmonious father–son relationships.

As a new factor extracted from popular movies in China, “Comedy” includes many animated sequel movies that may as well be categorized as “Family Movies” and “Sequels”. A salient example is *New Happy Dad and Son: Adventure in Russia*, the third episode of the *New Happy Dad and Son* sequel. Similarly, *Boonie Bears* is a Chinese original series favored by viewers of all ages that tells hilarious stories of two carefree bears casually defending the woods from a lone lumberman who always fails. Compared to the *Happy Dad and Son* sequels, the *Boonie Bears* carries a shorter history but has earned its popularity over the past eight years on TV screens and in cinemas. In addition, apart from these Chinese original series, audiences in China enjoy *The Secret Life of Pets*, an animated sequel produced by Universal Pictures, as well.

The third extracted factor, “Detective”, turns out to be contributed almost solely by *Detective Conan*, a Japanese detective series that portrays the adventures of a teenage detective solving one mysterious case after another in the hope of finding the solution to a poison that turned his appearance into a child’s. Originally distributed as manga and animations back in the 1990s, the series has gained immense popularity in China among generations of viewers, for whom the stories of the forever elementary schooler never get old. While the earlier episodes of the series were quite thrilling, the elements of horror have been fading away in more recent chapters of *Detective Conan*, making them good fits for family movies. Meanwhile, *Detective Chinatown* is a Chinese comedic detective series featuring some of the most popular comedy actors in China such as Baoqiang Wang. The

first episode of the series was aired in 2015, and more installments are under production. Similar to *Detective Conan*, *Detective Chinatown* contains minimal blood and gore to the extent that it can be enjoyed by family members of almost all ages.

The fourth extracted factor, “Action/Crime”, describes a diverse set of movies, ranging from the dystopian fiction *The Hunger Games* to the black comedy crime drama *Three Billboards Outside Ebbing, Missouri* and the slapstick comedy fantasy *A Chinese Odyssey*. As different as these movies are, Chinese audiences demonstrate a consistent preference for movies with comedic elements, though the nature of such elements can vary from dark to light.

4.3. Intergroup Analysis of Movie Title Keywords in the United States and China

Comparing the extracted factors of popular movies in the U.S. with the extracted factors of popular movies in China, there exist many interesting similarities. Movie producers may find some of these similarities inspiring as they create movie titles. First, audiences in both countries have a strong preference for family movies. In both countries, the majority of such movies tend to be animated movies with titles that primarily consist of the names of iconic characters such that consumers taking their families with them at the box office can quickly identify their perfect choices. One unique example in the case of China is *Dad, Where Are We Going*, as it is adapted from a popular reality show of the same name. In this case, the key element in its title that signals to potential viewers is, therefore, not the names of particular characters but the fame of the reality show itself. Either way, however, the titles of popular family movies never hesitate to take advantage of brand effects by using well-established content distributed in other formats to draw attention from viewers of all ages. Second, the popularity of *The Star Wars* sequels proves to be international, given that in both countries, “Sequels” is an extracted factor contributed almost entirely by various episodes of the series. It is, therefore, never a bad idea for a producer to work with George Lucas and create movies that depict the universe of *The Star Wars* to a greater extent.

The extracted factors also indicate that there is a salient difference between U.S. and Chinese viewers that movie producers need to keep in mind as they consider who their target audiences are. Specifically, Chinese viewers tend to enjoy comedies more than their American counterparts to the extent that many popular movies in China contain comedic elements even though they fit into more serious categories such as detective movies and action and crime movies. Patterns in the titles of such movies are difficult to detect, but they almost certainly belong to certain popular sequels that Chinese viewers are familiar with, as exemplified by the *Detective Conan* series, the *Detective Chinatown* series, and the *Chinese Odyssey* series. On the other hand, American viewers tend to appreciate elements of horror and thrill more. The titles of such movies are highly diverse as well, but most of them make use of keywords such as angels and death to cater to the religious nature of American culture.

5. Implications

In sum, inspired by the [Bae and Kim \(2019\)](#) study and the [Sood and Drèze \(2006\)](#) study, which revealed that informative movie title keywords play a critical role in predicting the economic success of their corresponding movies, this study discovered, with the use of the methods of statistical text mining and exploratory factor analysis, interesting patterns that underlie the titles of popular movies in the United States and China. These patterns have a range of implications for a variety of stakeholders in different ways.

For movie producers who have established their brands in the market, the message they can get is straightforward. With the success of sequel movies in both the U.S. and China, the strategy of building up brand effects proves to be quite effective in summoning loyal audiences who are willing to pay for the sake of the character or the name of the sequel that they see in the titles. There is no need for them to use any complex skills in the creation of successful movie titles, as simply adding “the movie” next to the name of the main character would work. This not only confirms the findings of [Bae and Kim \(2019\)](#)

that sequel movies with titles similar to their successful first episodes are more likely to be successful as well, but also explains the massive success achieved by the Disney empire as it promotes one episode after another under the name of the phenomenally successful *Star Wars*, as exemplified by their decision to produce *Rogue One*, a fairly independent story on its own. Had Disney forgotten to add “a Star Wars Story” in the title, the revenue achieved by *Rogue One* would probably be an entirely different story.

For producers who do have limited budgets, on the other hand, the strategy they should use based on our findings is more complicated. Clearly, as they do not have a successful first episode to benefit from the brand effects of, they are unlikely to achieve success via action sequel movies. Even though these genres describe the majority of the most successful movies, as Pangarker and Smit (2013) suggested, low-budget movie producers will inevitably find the budgets needed for visually impressive CGI and action stars unaffordable. Therefore, as discussed in the previous section, horror and thriller movies should be considered by low-budget movie producers in the U.S., while comedic movies should be considered by their counterparts in China. In both the U.S. and China, low-cost, animated family movies are a decent direction to look at, too. In the case of China, where the advantage of high-budget films is known to be less salient (Feng and Sharma 2016), low-budget film producers stand an especially better chance.

One major limitation of this study is that despite identifying important factors that largely describe the features of successful movie title keywords, we did not fully dive into analyzing and uncovering the relationships among these extracted factors. Another major limitation of this study is that our findings are highly time-sensitive. In years when the *Star Wars* franchise is not producing new episodes, our methods would not be capable of detecting its immense popularity, as words such as star, wars, and rogue as in *Rogue One* would not appear on our radar. Meanwhile, as COVID-19 continues to have devastating effects on the movie industry across the world, the representativeness of the movies we focused on was undoubtedly compromised, as they were all released between 2015 and 2019. Due to the space limit, we will continue to address these limitations in a future study in which we are going to quantify the relationships among the extracted factors with structural equation modeling techniques and qualitatively investigate the implications of the pandemic for the box office.

Author Contributions: Conceptualization, J.-M.K.; data curation, Y.C.; formal analysis, X.X.; investigation, Y.C.; methodology, J.-M.K.; project administration, X.X.; resources, X.X. and Y.C.; software, X.X.; supervision, J.-M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Bae, Giwoong, and Hye-jin Kim. 2019. The impact of movie titles on box office success. *Journal of Business Research* 103: 100–9. [CrossRef]
- Chang, Byeng-Hee, and Eyun-Jung Ki. 2005. Devising a practical model for predicting theatrical movie success: Focusing on the experience good property. *Journal of Media Economics* 18: 247–69. [CrossRef]
- Du, Jingfei, Hua Xu, and Xiaoqiu Huang. 2014. Box office prediction based on microblog. *Expert Systems with Applications* 41: 1680–89. [CrossRef]
- Elberse, Anita, and Felix Oberholzer-Gee. 2007. Superstars and underdogs: An examination of the long tail phenomenon in video sales. *MSI Reports Working Paper Series* 4: 49–72.
- Everitt, Brian, and Torsten Hothorn. 2011. *An Introduction to Applied Multivariate Analysis with R*. Berlin: Springer Science & Business Media.
- Feng, Fan, and Ravi Sharma. 2016. Modeling the main determinants of movie scales: An econometric study of Chinese marketplace. *Journal of Reviews on Global Economics* 5: 190–209. [CrossRef]
- Henson, Robin K., and J. Kyle Roberts. 2006. Use of exploratory factor analysis in published research. *Educational and Psychological Measurement* 393–416. [CrossRef]
- IBM Knowledge Center. 2020. *Factor Analysis Rotation*. Retrieved from IBM Knowledge Center. Available online: https://www.ibm.com/support/knowledgecenter/SSLVMB_sub/statistics_mainhelp_ddita/spss/base/idh_fact_rot.html (accessed on 20 December 2020).

- Kerlinger, Fred N. 1979. *Behavioral Research: A Conceptual Approach*. New York: Holt, Rinehart & Winston.
- Legoux, Renaud, Denis Larocque, Sandra Laporte, Soraya Belmati, and Thomas Boquet. 2016. The effect of critical reviews on exhibitors' decisions: Doreviews affect the survival of a movie on screen. *International Journal of Research in Marketing* 33: 357–74. [CrossRef]
- McClintock, Pamela. 2019. 2019 Global Box Office Revenue Hit Record \$42.5B Despite 4 Percent Dip in U.S. Retrieved from Billboard. Available online: <https://www.billboard.com/articles/news/8547827/2019-global-box-office-revenue-hit-record-425b-despite-4-percentdip-in-us> (accessed on 3 January 2021).
- Pangarker, N. A., and Eon Smit. 2013. The determinants of box office performance in the film industry revisited. *South African Journal of Business Management* 44: 47–58. [CrossRef]
- Pikowicz, Paul G., and Yingjin Zhang. 2006. *From Underground to Independent: Alternative Film Culture in Contemporary China*. Lanham: Rowman & Littlefield.
- Rahmawati, Sela, Jadi Suprijadi, and Zulhanif. 2017. Text mining factor analysis (TFA) in green tea patent data. *AIP Conference Proceedings* 1827: 020040.
- Sood, Sanjay, and Xavier Drèze. 2006. Brand extensions of experiential goods: Movie sequel evaluations. *Journal of Consumer Research* 33: 352–60. [CrossRef]
- Stapleton, Jennifer. 1997. *DSDM: Dynamic Systems Development Method: The Method in Practice*. Boston: Addison-Wesley Professional.
- Stevens, James. 1996. *Applied Multivariable Statistics for the Social Sciences*. Mahwah: Lawrence Erlbaum Associates.
- Tan, Ah-Hwee. 1999. Text mining: The state of the art and the challenges. Paper presented at PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, Beijing, China, April 26–28.
- The Associated Press. 2011. *Poll: Nearly 8 in 10 Americans Believe in Angels*. Retrieved from CBS News. Available online: <https://www.cbsnews.com/news/poll-nearly-8-in-10-americans-believe-in-angels/> (accessed on 4 January 2021).
- Zhang, Yingjin. 2010. Transnationalism and translocality in Chinese cinema. *Cinema Journal* 49: 135–39. [CrossRef]