HOLLYWOOD MOVIE DATA ANALYSIS BY SOCIAL NETWORK ANALYSIS AND TEXT MINING

Jong-Min Kim
University of Minnesota-Morris
jongmink@morris.umn.edu

Xingyao Xiao
Boston College
xiaoxg@bc.edu

Iksuk Kim
California State University
ikim@calstatela.edu

ABSTRACT

Analyzing the success of movies has always been a popular research topic in marketing. Movie title can be significant enough to attract audience attention. We investigate the relationship between the most frequently used movie title keywords and Return on Investment (ROI) together with movie performance data.

Keywords: Text Mining, Social Network Analysis, Movie Marketing Strategy.

1. INTRODUCTION

The continued growth of the movie industry has been a global phenomenon. According to the annual report from the Motion Picture Association of America, the global box-office market approached $38.3 billion in 2015. The expansion of the movie market encourages the production of research with various approaches. [1] showed that a movie with excellent reviews has a high chance of staying longer in a theater. Industrial decision-makers require the establishment of a highly accurate model to predict the success of a movie. These decision-makers aim to reduce the probability of making false decisions in the green-lighting process, the process of formally approving a movie production. [2] investigated the probability that an individual-level decrease in preference over time is due to the well-known decrease in a movie’s revenue after opening. Machine learning is a well-employed method and has been repeatedly used to build prediction models in previous studies (i.e., [3]; Lee, [4]). Machine learning can provide systematic support for decision-making. The previous research has concentrated on building new algorithms and methods of classification rather than focusing on the interpretation of findings. Thus, this study will analyze the relationship between the most frequently used movie title keywords and Return on Investment (ROI) together with other movie performance data.

In Section 2, we do a literature review and rationale of research and describe the Hollywood data we collected. Section 3 describes text mining and social network
analysis. In Section 4, we describe the nonparametric test of median and quantities. Section 5 performs data analysis based on text mining, social network analysis, and tests of median and quantiles to investigate the relationship between the most frequently used movie title keywords and ROI. In Section 6, concluding remarks are presented.

2. LITERATURE REVIEW AND RATIONALE OF RESEARCH

Audience movie reviews used as a forecasting tool to provide a “fingerpost” for film companies. [5] used a text mining technique to explore “movie reviews including word of mouth (WOM) factors (i.e., movie content, positive, negative, and promotion) and related factors (i.e., time, rating, and the number of ratings) for the box office.” To be specific, according to the frequency that words made an appearance and identified the most suitable cluster classification, [5] made a framework to indicate how movie review feeling (i.e., promotion, negative, positive content) and movie types affect box office sales. Then they used the k-mean algorithm to partition the observed keywords into exactly k clusters. Forecasting of movie success is not easy because the movie industry often depends on complex issues such as social and economic factors. Therefore, the previous research employed various methods for film producers and distributors to predict the economic success of a film. [4] used an ensemble approach to predict box office performance. [6] compared the performance of various machine learning methods by taking movie ratings as an example of high-dimensional data. [7] suggested a decision support system to help movie investment decisions at the early stage of movie productions by using social network analysis and text mining techniques extracting several sets of features such as “who”, “what”, “when”, and “hybrid” features that match “who” with “what” and “when” with “what” for predicting movie profitability. Table 1 shows the previous literature of movies with employed methods. [8] explored the internal and external factors that influenced box offices in China. They discussed the difficulty of predicting box office revenues accurately. By calculating the correlation coefficients of different periods and using linear regression with the stepwise method, they proved that movie views of 1 week before releasing on Youku represent the market performance, and they indicated how powerfully influential users control box offices. In other words, trailers integrate the virtue of timing and content, which is the best choice. [8] also made a correlation analysis in a fixed period based on consistent and representative online data (Sina Weibo). [9] found out the effect of Tweets on movie sales using machine learning algorithms. The findings of this research indicated the relationship between positive/negative Twitter WOM and higher/lower movie sales and revealed the potential values of monitoring people’s intentions.

In order to perform the analysis, we rely mainly on information concerning 2010-2015 movie titles and genres collected from IMDb. We retrieved box office performance, critics’ reviews, and production budgets from Boxofficemojo and Metacritic. The complete data set uses a total of 723 movies categorized under 24 distinct film genres. The descriptions of the employed variables are as follows:

- Audience: Total number of audience members in the U.S. for a particular movie.
- BoxOffice: The total revenue of a particular movie from U.S. domestic theaters.
- Budget: The total production cost of a particular movie.
- MetaScore: A weighted average score of published critic reviews of a particular movie.
- Number of Theaters: The total number of theaters screening a particular movie.
- Running Weeks: The length of a theater run for a particular movie, given in weeks.
- Return on Investment (ROI): \( ((\text{BoxOffice} - \text{Budget})/ \text{Budget}) \times 100 \)

3. TEXT MINING AND SOCIAL NETWORK ANALYSIS (SNA)

We use Text Mining and Social Network Analysis (SNA) for finding meaningful results from Hollywood movie data.

3.1. Text Mining

Text mining deals with helping computers understand the “meaning” of the text. Some of the standard text mining applications include sentiment analysis. This process includes data import, corpus handling, preprocessing, metadata management, and creation of term-document matrices.

First, we import data from an external source (IMDb) so that we can then use Analytics to organize and analyze all of our data in ways that better reflect our goal. Second, we do corpus handling by using text mining R package ‘tm’ commands TermDocumentMatrix and DocumentTermMatrix employing sparse matrices for corpora.

3.2. SNA

The methods of social network analysis have applied to our society, especially the social science community, in recent decades. In order to extract some meaningful information from Hollywood movie title keywords, we apply social network analysis R package sna to Hollywood movie title keywords. The set \( N \) contains \( g \) movie title keywords, which we will denote by \( N = \{n_1, n_2, \ldots, n_g\} \). Suppose the relation is unweighted and directional. Thus \( v_i \) either relates to \( v_j \) or does not, which means we do not consider the strength of the interaction. Ties can be represented graphically by drawing a line from the first movie title keyword in the element to the second. Such a graph is refereed as a directed graph. Suppose we have a single relation on one set of \( n \) movie title keywords in \( V \). We define \( X \) as the matrix. We let \( x_{ij} \) be the value of the tie from the \( i \)th to the \( j \)th keyword on the single relation. Since there are \( n \) keywords, the matrix is of size \( n \times n \). The value of the tie from \( v_j \) and \( v_i \) is placed into the \( (i, j) \)th element of \( X \). Since the relation is unweighted, the values for the tie are simply 0 and 1. Since we do not allow self-choices, the main diagonal would be 0.
4. TEST OF MEDIANS AND QUANTILES

Quantiles serve good help in the role that summarizes a frequency distribution that relates to the rank order of values. For example, the middle location value of the sample data (50\textsuperscript{th} percentile) is called a median, which also means 50\% of the probability distribution. The 75\textsuperscript{th} percentile (upper quartile) is the third quartile of the rank order of value that has an ascending trend. In our research, we use the median test to check whether two independent groups (movies that contain popular movie keywords and movies that not contain popular movie keywords) differ in central tendency.

So we use the Wilcoxon rank sum test to achieve our goal because the movie data do not follow a normal distribution assumption. The Wilcoxon rank sum test is a nonparametric approach test that can be used to determine whether we selected two dependent samples from populations having the same distribution. In terms of the marketing viewpoint to the film industry, define ROI (Rate on Investment) with Box-office and Budget variables as follows:

\[
ROI = \frac{(\text{Boxoffice} - \text{Budget})}{\text{Budget}} \times 100.
\]

(1)

The higher value of the ROI is, the more profitable a movie is. Our research goal is to compare the rate of return between the main popular movie title keyword and non-popular movie title keywords. We select an independent SRS of size \(n_1\) from one population and select an independent SRS of size \(n_2\) from another population. Suppose that there are \(N\) observations in all, where \(N = n_1 + n_2\). We rank all \(N\) observations such that the sum \(W\) of the ranks for the sample is the Wilcoxon rank sum statistic. If the two populations have the same continuous distribution, then \(W\) has mean

\[
\mu_W = \frac{n_1(N + 1)}{2}
\]

and its standard deviation is

\[
\sigma_W = \sqrt{\frac{n_1n_2(N+1)}{12}}.
\]

(3)

The Wilcoxon rank sum test rejects the hypothesis that the two populations have identical distributions when the rank sum \(W\) is far from its mean.

When the distribution may not be normal, we state the hypotheses in terms of population medians as follows: \(H_0:\text{median}_1 = \text{median}_2\) vs.\(H_a:\text{median}_1 \neq \text{median}_2\). The Wilcoxon rank sum test will test the hypotheses only if both populations must have distributions of the same shape. [10] propose to compare two independent groups via the lower and upper quantiles. We test \(H_0:\theta_{q1} = \theta_{q2}\), where
\(\theta_{q_j}\) is the \(q^{th}\) quantile corresponding to the \(j^{th}\) group \((j = 1, 2)\). The Harrell–Davis estimate of \(\theta_{q}\), the \(q^{th}\) quantile, is \(\hat{\theta}_q = \sum_{i=1}^{n} \frac{X_{i}}{W_i}\). Let \(X_{ij}\) be a random sample from the \(j^{th}\) group \((i = 1, \ldots, n_j)\). We generate a bootstrap sample from the \(j^{th}\) group by resampling with replacement \(n_j\) observations from group \(j\). Let \(\hat{\theta}^*_j\) be the estimate of the \(q^{th}\) quantile for group \(j\) based on this bootstrap sample. Let \(d^*_j = \hat{\theta}^*_1 - \hat{\theta}^*_2\). We repeat this process \(B\) times yielding \(d^*_b, b = 1, \ldots, B\). We set \(B = 2000\). Let \(l = \alpha B/2\) be rounded to the nearest integer, and let \(u = B - l\). Letting \(d^*_1 \leq \cdots \leq d^*_B\) denote the ascending order \(B\) bootstrap estimates, an approximate \(1-\alpha\) confidence interval for \(\theta_1 - \theta_2\) is \((d^*_{(l+1)} - d^*_u)\).

5. DATA ANALYSIS

We used text mining in R (2017) ‘tm’ and R (2017) ‘sna’ packages for our data analysis. It provides a graphical representation of word frequency; the more frequent the word is in the document, the larger the word is in the visual. First, we create a vector source with movie titles. Second, we perform text corpus data analysis with full support for international text, functions for reading data from newline-delimited 'JSON' files, for normalizing and tokenizing text, for searching for term occurrences, and for computing term occurrence frequencies, including n-grams. These functions create or convert another object to a corpus object. A corpus object is just a data frame with special functions for printing. Last, we use the command “DocumentTermMatrix” to get the word frequency.

To compare the difference of two different groups when data has non-normal distribution, we usually employ the Wilcoxon rank sum test which is a nonparametric test of the medians of two different groups. The following table is the result of a Wilcoxon rank sum test with \(n=50, 100, 200\) top popular movie title keywords obtained from the text mining method. We found that when \(n=200\), statistical significance for median difference exists for two different groups (popular movie keywords and non-popular movie keywords) at the 5% significance level.

| Table 1: Wilcoxon Rank Sum Test |
|---|
| **n** | **Test Statistic** | **P-Value** | **Statistical significance** |
| 50 | 60456 | 0.7088 | No |
| 100 | 62459 | 0.5832 | No |
| 200 | 71500 | 0.02822 | Yes* |

*significant at 0.05. Note. \(n=\) the number of top popular movie title keywords
Table 2: Normality Test of Shapiro-Wilk

| Group | Test Statistic | P-value | Statistical significance |
|-------|----------------|---------|--------------------------|
| P     | 0.12191        | 0.0000  | Yes                      |
| N     | 0.12191        | 0.0000  | Yes                      |

Note. P = movies that contain 200 popular keywords
N = movies that do not contain 200 popular keywords

We did a normality test of Shapiro-Wilk in Table 2, which is the test designed to detect all departures from normality. Table 2 indicates that the P-values are smaller than the significant level \( \alpha=0.05 \); the test rejects \( H_0 \): data has a normal distribution at the 0.05 significance level. It shows that the data is not distributed as a normally distributed population. In other words, the data of group P and group N are non-normal. When our outcome is not normally distributed, a nonparametric test is appropriate. That is the reason why we apply the nonparametric test of medians (Wilcoxon Rank Sum Test) of two different groups (Popular movie keywords and non-popular movie keywords).

![Histograms and Boxplots of Top 200 Popular Movie Title Keywords](image)

Figure 1. Histograms and Boxplots of Top 200 Popular Movie Title Keywords

By using histograms and boxplots with the top 200 popular movie keywords and non-popular movie keywords to visualize ROIs data, we confirmed that the movie data has a skewed to the right distribution (non-normal distribution). A whole list of the most popular 200 keywords in movie titles are listed in the Appendix. Figure 2 represents the word cloud that better visualizes the frequency of Hollywood movie keywords. To be specific, it provides a graphical representation of word frequency; the more frequent the word is in the document, the larger the word is in the visual. In figure 2, if the words have the same color and size, they have almost the same
frequency. For example, the three words: Movie, Man, and Part have the same size, and the color is black. Specifically, they are the top 3 popular movie keywords that we can find in the Appendix, and they have the same frequency (15466). Then, four words are in yellow (day, life, love, one), but the size of the word “day” (11248 frequency) is smaller than the other three words (15466 frequency).

Figure 2. Word Cloud Plot of Hollywood Movie Title Data (2010-2015 Year) with 1000 Minimum Frequency.

Figure 3. SNA Plots of Top 10 Popular Movie Title Keywords. (a)
Figure 3. SNA Plots of Top 10 Popular Movie Title Keywords. (b)

Figure 3. SNA Plots of Top 10 Popular Movie Title Keywords. (c)
Figure 4. SNA Plots of Top 25 Popular Movie Title Keywords. (a)

Figure 4. SNA Plots of Top 25 Popular Movie Title Keywords. (b)
Figure 4. SNA Plots of Top 25 Popular Movie Title Keywords. (c)

Figure 3 and figure 4 provide SNA (Social Network Analysis) plots of the top 10 and top 25 popular movie keywords, respectively. We can understand SNA as a plot that provides how the words relate, the importance of the words, and the structure of the words that we analyze. First, in plots (a) and (b) in figure (3) and (4), the words form a network that we can visualize their relationship, importance, structure, etc. from the nodes formed by a point and the bridges that connect points. In plot (a) of the figure (3), the ten popular movie keywords form an irregular octagon. When the number of words increases to 25 (figure (3(a))), the shape is close to an irregular circle. Notice the words one and love inside the SNA plot (3(a)). We can look at the centrality; the words located in the center of the SNA plot are more relevant and influential than peripheral words. We can also find that the words man and best are centralized in the irregular circle (figure 4(a)), which means the two words are the most important of the 25 popular movie keywords. The density of centralized words also has a higher density, which the ties are connected to one hole in a network relative to the total number possible. The distance between two words can also indicate how strong the relationship is. Plot (b) in the figures 3 and 4 indicates distance more obvious. The more nuclear the words are, the smaller the distance between two words. For example, in the figure 3(b), the distance between two central words: bad and city are significantly shorter than the length of the tie around the word: man. It indicates that centralization represents importance. In plot (c) of the figures (3) and (4), we use arrows to replace the ties in the plots (a) and (b), represent the relationships by a variety of colors, sizes, and shapes. It is easier to visualize the structure in plot (c) than in the plots (a) and (b). Explicitly speaking, words of the same color have a strong relationship. So we can predict the future evolution of the graph. For example, in the plot 3(c), we study 10 popular movie keywords and find that the words atlas and movie are both red. In plot 4(c), these two words have the same color and shape as in the figure 3(c), and two more red words join pattern. It
indicates that when the observation size becomes large, the relationships between words become firm and complicated. By investigating the relationships between the most frequently used movie title keywords and return on investment (ROI), together with movie performance data, our goal is to perform the return on investment of the group P (movie titles that contain popular movie keywords) which are higher than group N (movies that do not contain popular movie keywords). Summary statistics of 200 popular words and movie performances are presented in Table 3.

**Table 3**: Summary Statistics of Movie Performance with Top 200 Popular Keywords

| Top 200 Keywords (n=364) | Statistics | Audiences | BoxOffice | Budget | MetaScore | Theaters | Weeks | ROI     |
|-------------------------|------------|-----------|-----------|--------|-----------|----------|-------|---------|
| Mean                    | 9,158,320  | 74,214,064| 53,289,365| 50.88  | 2,436.76  | 13.03    | 226.00|
| Median                  | 4,757,457  | 38,540,831| 30,000,000| 51.00  | 2,936.00  | 13.03    | 9.44  |
| Std. Dev.               | 11,694,470 | 95,521,339| 56,679,094| 17.04  | 1346.22   | 6.56     | 1692.87|
| Min.                    | 424        | 3,450     | 6,000     | 9.00   | 1.00      | 1.00     | -99.95|
| Max.                    | 77,374,930 | 652,270,625| 250,000,000| 95.00  | 4,390.00  | 53.00    | 28,752.20|

| Non-Popular Keywords (n=359) | Statistics | Audiences | BoxOffice | Budget | MetaScore | Theaters | Weeks | ROI     |
|-----------------------------|------------|-----------|-----------|--------|-----------|----------|-------|---------|
| Mean                        | 6,014,505  | 48,243,856| 41,728,508| 52.34  | 2,066.12  | 12.49    | 174.44|
| Median                      | 3,796,003  | 29,807,260| 25,000,000| 51.00  | 2,557.00  | 12.00    | -8.53 |
| Std. Dev.                   | 7,453,036  | 60,195,129| 46,092,087| 16.48  | 1,368.18  | 6.88     | 1,413.94|
| Min.                        | 160        | 1,309     | 17,000    | 5.60   | 1.00      | 1.00     | -99.99|
| Max.                        | 49,291,270 | 400,738,009| 250,000,000| 100.00 | 4,311.00  | 36.40    | 22,664.40|

In Table 3, 364 movies used the top 200 popular movie title words and 359 movies used non-popular movie title keywords to compare movie performances.

**Table 4**: t-test for Popular Movie and Non-Popular Movie Title Keyword Groups

| Statistics | Audiences** | BoxOffice** | Budget** | MetaScore | Theaters** | Weeks |
|------------|-------------|-------------|----------|-----------|------------|-------|
| t-test     | 4.31        | 4.38        | 3.01     | -1.17     | 3.67       | 1.07  |
| p-value    | 0.00        | 0.00        | 0.00     | 0.24      | 0.00       | 0.29  |
| 95% CI     | (1,711,193,  | (14,325,449,| (4,022,049,| (-3.91,    | (172.43,   | (-0.45,|
|            | 4,576,437)  | 37,614,965) | 19,099,664)| 0.99 | 568.85 | 1.52 |

**significant at 0.01**

**Table 5**: Wilcoxon-test for Popular Movie and Non-Popular Movie Title Keyword Groups

| Statistics | Audiences** | BoxOffice** | Budget** | MetaScore | Theaters** | Weeks |
|------------|-------------|-------------|----------|-----------|------------|-------|
| Test Statistics | 73465 | 74505 | 71506.5 | 63183.5 | 76245 | 70057.5 |
| p-value    | 0.00        | 0.00        | 0.03     | 0.44      | 0.00       | 0.09  |

**significant at 0.05**
Among six single movie performance variables, Tables 4 and 5 show that Audience, BoxOffice, Budget, and Theaters have a statically significant difference between P and N at the 5% significance level. From the 95% confidence interval for the mean difference of the variables (Audiences, Box office, Budget, Theaters) in Table 4, the lower bound and upper bound are all positive. It indicates that the mean of the popular movie title keyword group for the variables is higher than the mean of the non-popular movie title keyword group for the variables. However for the variables (MetaScore and Weeks), it is not statistically significant at the 10% significance level. Also, from the 95% confidence interval for the mean difference of the variables (MetaScore and Weeks), there exists zero in between the lower bound and upper bound. It suggests that the mean of the popular movie title keyword group for the variables is not different from the mean of the non-popular movie title keyword group for the variables.

By using R package “WRS” from [10] that compares user-defined quantiles of both distributions using a Harrell–Davis estimator in conjunction with a percentile bootstrap. We have five experiment units, and we conducted Wilcoxon Rank Sum Test to test the equality of means in two independent groups (group P and group N) related to each experiment unit, respectively. To be specific, our goal of this test is to compare the medians between the two populations use paired data. We can find that $t$-test statistic increases based on the increase in the number of Top Popular Movie Title Keywords (from 24149 to 71500). That means the differences in the median between two groups increase as our popular movie title keywords increase from 10 to 200. The $P$-value is less than a significant level $\alpha = 0.05$ when $n=200$, which means we rejected $H_0: \text{Median}_1 = \text{Median}_2$, and the test is substantial when $n=200$. Based on our alternate hypothesis $H_a: \text{Median}_1 \neq \text{Median}_2$, we can conclude that for a statistical significance level $\alpha = 0.05$, the distributions of Group P (200) and Group N (not 200) are not normally distributed by the normality test of Shapiro-Wilk. For skewed distributions, we used Wilcoxon Rank Sum Test to compare the medians of two independent groups. Based on comparing median of two independent groups by controlling Type I errors $\alpha$ [10] we are examining two independent groups (top 200 Popular Group and Non-popular Group) via the upper and lower quantiles in table 6. This test can obviate the situation that tied values (tied values occur when two or two more observations are equal) occur. It is “getting a reasonably accurate estimate of the standard error” [10]. Our goal is to test : $H_0: \theta_{q_1} - \theta_{q_2} = 0$ ($\theta_{q}$ is the $q^{th}$ quantile corresponding to the $j^{th}$ group). Using $\theta_{q}$ to estimate the population when we take some random samples. The estimator 1 ($\hat{\theta}_1$) is for the quantile of the top 200 popular movie title keywords, and the estimator 2 ($\hat{\theta}_2$) is for the quantile of not top 200 popular movie title keywords. We are considering the study comparing the value of ROI by using a .95 confidence band. The sample sizes are $n_1 = 364$ and $n_2 = 359$. The value of the difference of two estimators ($\hat{\theta}_1 - \hat{\theta}_2$) is getting more statistically significant along with $q^{th}$ quantile increasing. Comparing the .20, .25, .45, and .50 quantiles, the corresponding values of the difference of the lower confidence and the upper confidence are bigger than 0. For example, when $q^{th} = 20$ (20th percentile below our all observations $n_1$ and $n_2$), the corresponding value of lower confidence is 0.267, and upper confidence is 25.217.
Table 6: Comparing two independent groups via the upper and lower quantiles

| j | Quantile (q) | n1 | n2 | Est.1 | Est.2 | Est.1-Est.2 | ci.low | ci.up |
|---|--------------|----|----|-------|-------|-------------|--------|-------|
| 1 | 0.05         | 364| 359| -98.987 | -99.402 | 0.415 | -0.347 | 2.667 |
| 2 | 0.10         | 364| 359| -91.755 | -97.654 | 5.898 | -0.287 | 13.170 |
| 3 | 0.15         | 364| 359| -80.251 | -90.598 | 10.346 | -0.884 | 19.563 |
| 4 | 0.20         | 364| 359| -69.897| -80.471 | 10.574 | 0.267 | 25.217 |
| 5 | 0.25         | 364| 359| -55.211| -72.696 | 17.484 | 1.668 | 30.259 |
| 6 | 0.30         | 364| 359| -42.904| -58.565 | 15.660 | -2.411 | 33.192 |
| 7 | 0.35         | 364| 359| -29.116| -43.185 | 14.068 | -4.601 | 32.931 |
| 8 | 0.40         | 364| 359| -14.924| -30.202 | 15.278 | -1.913 | 33.376 |
| 9 | 0.45         | 364| 359| -2.031 | -19.791 | 17.759 | 2.019 | 33.138 |
| 10| 0.50         | 364| 359| 10.301 | -9.145  | 19.447 | 1.649 | 36.039 |
| 11| 0.55         | 364| 359| 22.791 | 3.938   | 18.853 | -0.056 | 37.450 |
| 12| 0.60         | 364| 359| 36.667 | 18.137  | 18.530 | -0.581 | 41.933 |
| 13| 0.65         | 364| 359| 55.232 | 33.046  | 22.186 | -6.914 | 52.089 |
| 14| 0.70         | 364| 359| 86.476 | 55.217  | 31.259 | -7.896 | 66.566 |
| 15| 0.75         | 364| 359| 121.988| 83.984  | 38.003 | -13.674| 84.444 |
| 16| 0.80         | 364| 359| 169.541| 135.110 | 34.430 | -24.754| 94.836 |
| 17| 0.85         | 364| 359| 228.729| 193.467 | 35.262 | -37.603| 135.824|
| 18| 0.90         | 364| 359| 392.445| 297.228 | 95.216 | -41.046| 255.489|
| 19| 0.95         | 364| 359| 733.460| 495.237 | 238.223| -79.218| 653.925|

Note. n1= the number of top 200 popular movie title keywords
n2= the number of other than top 200 popular movie title keywords

Figure 5. 95 % Confidence Interval for quantiles with (top 200 popular group – non-popular group)
We use figure 5 to visualize the information from table 12. The horizontal axis indicates the number of our total observations ($n_1 + n_2=723$); the vertical axis shows the value of $(\hat{\theta}_1 - \hat{\theta}_2)$. The nonlinear line represents the difference between two estimators $(\hat{\theta}_1 - \hat{\theta}_2)$. The points below the nonlinear line represent the values of the lower confidence, and the points above the nonlinear line represent the values of the upper confidence. As we mentioned in table 6, the overall trend of the upper confidence interval is upward, whereas the overall direction of the lower confidence interval is downward. Regarding the values of $\hat{\theta}_1 - \hat{\theta}_2$, the overall trend is upward. We can conclude that the values of ROIs from these quantiles (percentile of the total observations) are always positive.

6. CONCLUSION AND LIMITATION

The findings of this research show the importance of movie title selection and related financial performances by using text mining and social network analysis. Clearly, specific movie titles work better than others. The results of this research can be applied to other entertainment businesses such as the titles of music, plays, and even novels. While this research provides a clear comparison of movie title choice and the corresponding financial performances, it also suggests a direction and guidance for future research. First, this study did not consider the impact of multiple keywords in a movie. Thus, the impact of only one distinctive word in a movie toward financial performance was measured and compared. If one movie title has multiple words, it is unknown how the results may differ. In such a case, we would need to identify what portion of the primary words influences movie performance. This research reveals that there are future research opportunities for understanding movie marketing in different ways, but it is limited in that it cannot provide a complete generalization of all movie title keywords and financial performances. Future studies may draw upon these research opportunities to resolve this limitation.

REFERENCES

[1] Legoux, R., Larocque, D., Laporte, S., Belmati, S. and Boquet, T. (2016). The effect of critical reviews on exhibitors' decisions: Do reviews affect the survival of a movie on screen?, International Journal of Research in Marketing, 33 (2), 357-374.

[2] Ho, J. Y. C., Chang J., and Krider, R. E. (2016) Mere newness: Decline of movie preference over time, Canadian Journal of Administrative Science, 34, 33-46

[3] Du, J., Xu, H., and Huang, X. (2014). Box office prediction based on microblog. Expert Systems with Applications, 41(4), 1680–1689.

[4] Lee, K., Park, J., Kim, I., and Choi, Y. (2016). Predicting movie success with machine learning techniques: ways to improve accuracy. Information Systems Frontiers, 1-12.

[5] Chiang, I. (2014). Using Text Mining Techniques To Analyze How Movie Forums Affect The Box Office. International Journal of Electronic Commerce Studies,5(1), 91-96.

[6] Qiu, Z. and Gao, Q. (2016). Movie Success Predictor and Two Brand-new Bagging Algorithms. Project Report, 1-10.
[7] Lash, M. T. and Zhao, K. (2016). Early Predictions of Movie Success: the Who, What, and When of Profitability. Journal of Management Information Systems, 30(3) 874-903.

[8] Xiao, J., Li, X., Chen, S., Zhao, X., & Xu, M. (2017). An inside look into the complexity of box-office revenue prediction in China. International Journal of Distributed Sensor Networks, 13(1), 1-14.

[9] Rui, H., Liu, Y., & Whinston, A. B. (2013). Whose and What Chatter Matters? The Impact of Tweets on Movie Sales. Decision Support Systems, 55(4), 863-870

[10] Wilcox, R. R., Erceg-Hurn, D. M., Clark, F. and Carlson, M. (2014). Comparing two independent groups via the lower and upper quantiles. Journal of Statistical Computation and Simulation, 84 (7), 1543-1551.
# APPENDIX

List of Top 200 Popular Movie Keywords

| Rank | Word   | n   | Rank | Word   | n   | Rank | Word   | n   | Rank | Word   | n   |
|------|--------|-----|------|--------|-----|------|--------|-----|------|--------|-----|
| 1    | man    | 15466 | 26   | dragon | 5624 | 51   | hotel  | 2826 | 76    | Angeles | 2826 |
| 2    | movie  | 15466 | 27   | fast   | 5624 | 52   | jack   | 2826 | 77    | annie   | 2826 |
| 3    | part   | 15466 | 28   | furious | 5624 | 53   | joe    | 2826 | 78    | apes    | 2826 |
| 4    | day    | 11248 | 29   | game   | 5624 | 54   | kid     | 2826 | 79    | back    | 2826 |
| 5    | life   | 9842  | 30   | girl   | 5624 | 55   | left   | 2826 | 80    | beyond  | 2826 |
| 6    | love   | 9842  | 31   | green  | 5624 | 56   | legend | 2826 | 81    | black   | 2826 |
| 7    | one    | 9842  | 32   | kill   | 5624 | 57   | lincoln | 2826 | 82    | blood   | 2826 |
| 8    | last   | 8436  | 33   | shrugged | 5624 | 58   | machete | 2826 | 83    | book    | 2826 |
| 9    | night  | 8436  | 34   | time   | 5624 | 59   | mars   | 2826 | 84    | bosses  | 2826 |
| 10   | atlas  | 7030  | 35   | activity | 4218 | 60   | men    | 2826 | 85    | breaking | 2826 |
| 11   | best   | 7030  | 36   | age    | 4218 | 61   | Paranormal | 2826 | 86    | call    | 2826 |
| 12   | big    | 7030  | 37   | america | 4218 | 62   | planet | 2826 | 87    | captain | 2826 |
| 13   | dawn   | 7030  | 38   | american | 4218 | 63   | rise   | 2826 | 88    | christmas | 2826 |
| 14   | days   | 7030  | 39   | blue   | 4218 | 64   | run    | 2826 | 89    | cop    | 2826 |
| 15   | evil   | 7030  | 40   | boy    | 4218 | 65   | runner | 2826 | 90    | crazy   | 2826 |
| 16   | good   | 7030  | 41   | chapter | 4218 | 66   | safe   | 2826 | 91    | dead    | 2826 |
| 17   | house  | 7030  | 42   | diary  | 4218 | 67   | secret | 2826 | 92    | die    | 2826 |
| 18   | like   | 7030  | 43   | earth  | 4218 | 68   | son    | 2826 | 93    | doctor  | 2826 |
| 19   | red    | 7030  | 44   | feet   | 4218 | 69   | spy    | 2826 | 94    | dollar  | 2826 |
| 20   | war    | 7030  | 45   | gods   | 4218 | 70   | street | 2826 | 95    | dolphin | 2826 |
| 21   | world  | 7030  | 46   | guardians | 4218 | 71   | tale   | 2826 | 96    | door    | 2826 |
| 22   | bad    | 5624  | 47   | happy  | 4218 | 72   | wanted | 2826 | 97    | drive   | 2826 |
| 23   | battle | 5624  | 48   | home   | 4218 | 73   | Way    | 2826 | 98    | dust    | 2826 |
| 24   | city   | 5624  | 49   | hood   | 4218 | 74   | Woman  | 2826 | 99    | end    | 2826 |
| 25   | dark   | 5624  | 50   | horrible | 4218 | 75   | words  | 2826 | 100   | escape  | 2826 |
| Rank | Word | n   | Rank | Word     | n   | Rank | Word     | n   | Rank | Word     | n   | Rank | Word     | n   |
|------|------|-----|------|----------|-----|------|----------|-----|------|----------|-----|------|----------|-----|
| 101  | ever | 2826| 126  | impossible | 2826| 151  | mirror   | 2826| 176  | real     | 2826|
| 102  | Exotic | 2826| 127  | inside    | 2826| 152  | mission  | 2826| 177  | redemption| 2826|
| 103  | family | 2826| 128  | insidious | 2826| 153  | monsters | 2826| 178  | resident | 2826|
| 104  | first | 2826| 129  | iron      | 2826| 154  | muppets  | 2826| 179  | rio      | 2826|
| 105  | five  | 2826| 130  | jackass   | 2826| 155  | need     | 2826| 180  | road     | 2826|
| 106  | four  | 2826| 131  | john      | 2826| 156  | never    | 2826| 181  | rush     | 2826|
| 107  | fury  | 2826| 132  | journey   | 2826| 157  | new      | 2826| 182  | saga     | 2826|
| 108  | games | 2826| 133  | jump      | 2826| 158  | next     | 2826| 183  | san      | 2826|
| 109  | ghost | 2826| 134  | justin    | 2826| 159  | now      | 2826| 184  | scary    | 2826|
| 110  | glory | 2826| 135  | kickass   | 2826| 160  | number   | 2826| 185  | sex      | 2826|
| 111  | god   | 2826| 136  | killer    | 2826| 161  | paris    | 2826| 186  | shrek    | 2826|
| 112  | great | 2826| 137  | know      | 2826| 162  | parker   | 2826| 187  | skin     | 2826|
| 113  | greatest | 2826| 138  | ledge     | 2826| 163  | paul     | 2826| 188  | sold     | 2826|
| 114  | grey  | 2826| 139  | lone      | 2826| 164  | penguins | 2826| 189  | soldier  | 2826|
| 115  | guys  | 2826| 140  | long      | 2826| 165  | perfect  | 2826| 190  | stand    | 2826|
| 116  | hangover | 2826| 141  | los       | 2826| 166  | perrys   | 2826| 191  | star     | 2826|
| 117  | hard  | 2826| 142  | lost      | 2826| 167  | piranha  | 2826| 192  | stardom  | 2826|
| 118  | haunted | 2826| 143  | machine   | 2826| 168  | pitch    | 2826| 193  | stars    | 2826|
| 119  | hercules | 2826| 144  | madagascar| 2826| 169  | place    | 2826| 194  | steel    | 2826|
| 120  | hit   | 2826| 145  | madeas    | 2826| 170  | presents | 2826| 195  | story    | 2826|
| 121  | hobbit | 2826| 146  | magic     | 2826| 171  | prince   | 2826| 196  | taken    | 2826|
| 122  | hot   | 2826| 147  | marigold  | 2826| 172  | project  | 2826| 197  | think    | 2826|
| 123  | hunger | 2826| 148  | max       | 2826| 173  | punch    | 2826| 198  | thor     | 2826|
| 124  | hunter | 2826| 149  | mike      | 2826| 174  | purge    | 2826| 199  | train    | 2826|
| 125  | iii   | 2826| 150  | million   | 2826| 175  | raid     | 2826| 200  | transformers | 2826|

(cont.)
