Moviepedia: A Movie Analytic

Xinyi Liu
Computer Science, NYU
New York City, US
xl2700@nyu.edu

Xinsen Lu
Computer Science, NYU
New York City, US
xl2783@nyu.edu

Yiming Li
Computer Science, NYU
New York City, US
yl6183@nyu.edu

***Abstract* —** In this paper, an analytic which provides insights in movies for script writers and movie investors. Compared to common movie success predictions using information like film crews and plot keywords, this movie analytic focus on provide evidences of decision in an early stage of movie production. By analyzing gross box office incomes, user ratings, directors, actors and actress, and commonly mentioned keywords from reviews and their sentiments regarding to genres, writers are able to avoid clichés and get inspired from past great works, while movie producers may discover the potential success of a rough idea or a script at the very beginning.

***Keywords* —** *Movie, Big Data, Analytic, Sentiment*

I. Introduction

Hundreds of new movies are shown in movie theaters every year. There are many factors can affect the quality and outcome of each movie, such as casts, directors, genres, and scripts. While it's easy to get feedback of released films through movie databases like IMDb (Internet Movie Database), people may feel difficult to predict a success from a script or determine whether an idea worth investing.

This analytic provides insights in movies by genres through analyzing gross box office incomes, user ratings, numbers of Facebook likes, storylines, commonly mentioned keywords from reviews and their sentiments. While general statistics including the most popular movies and the trend of the development of movie industry are stated, other facts including directors and actors and actresses mastering certain genres are also discovered. The feedbacks of users from both IMDb and Twitter are investigate and compared, to see if there are any outliers. Common evaluations related to specific genres are displayed in the form of word clouds, with which writers may avoid clichés and get inspired from feedback of similar works.

II. Motivation

A good movie contains many factors, such as great acting, beautiful scenery, famous directors, as well as smooth movie scripts. Movie ratings, box office, and reviews from audience are representatives of success of movies. Most movie success prediction models available now take budgets, plot keywords, film crews as inputs and output estimated user rating or box office. However, script writers and movie producers may still need an analytic based on statistics to help them make decisions at an early stage.

The audience's tastes can be discovered from the average rating or box office for each genre. The budget needed, and the potential profits shall follow the trend. With summarized user reviews regarding to genres in the format of word cloud, it's clear and straightforward to get what the audience like and dislike. A producer can determine whether a script or an outline is another cliché from historical statistics and get some tips on how to get desired cast and crews.

III. Related Work

A lot of works has been done in the field of big data, movie analytics, prediction, sentiment analysis from plain texts, etc.

A movie analytic is provided by T. Ashwitha, A. Rodrigues and N. Chiplunkar [1]. The authors focus on getting ratings of action movies on user gender, ratings of adventure movies on user gender, highest rated movie for each year, and maximum Facebook likes in a year from IMDb Datasets using Hive. While this research proves that big data tools accelerate the analysis process, more insights can be gained other than simple calculations on historical data.

G. Mishne and N. Glance dive a little deeper, predicting movie sales from blogger sentiment [2]. The authors use Natural Language Processing and other sentiment analytical tools to analyze whether the words of blogger indeed have a big influence on the success of a movie. The result is that "there is good correlation between references to movies in weblog posts — both before and after their release — and the movies' financial success". Moreover, the paper also shows that it will have a good pre-prediction result when it is conjunction with other factors such as genre and season.

M. Mestyán, T. Yasseri and J. Kertész also tend to predict movie sales in their paper "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data" [3]. The author builds a multivariate linear regression model for predicting the box office revenues, which determine whether movies are good or not. The Wikipedia model is compared with a Twitter-based model. The Twitter-based model has a slightly success over the Wiki one, but the later can predict one month before. On the contrast, the Twitter-based one does a good job only one night before the result comes.

Tweets can be important as a rich source of big data to get reviews from the public. R. Paul proposes an analytic on how tweets are related to Indian premier league is conducted in his paper "Big Data Analysis of Indian Premier League using Hadoop and MapReduce" [4]. Tweet streams are used as input and the selected tweets are transformed into JSON format. The analysis is done by using MapReduce programs. The author focuses on the time interval in which cricket fans tweets most,

the most popular player, and the dominant team of the game. This paper provides some inspiration on how to deal with tweets from different aspects while taking all factors into consideration.

M. Kumar and A. Bala conduct a sentiment analysis on tweets using Hadoop for intelligent analysis and storage of big data in their paper "Analyzing Twitter Sentiments through Big Data" [5]. They use Naïve Bayes to realize the study of classification. Mahout commands are used to test the correctness of classifiers. The classifiers are then used on tweets related to Bharti Airtel. It will be useful in classifying the opinion from people in different fields.

B. Liu, E. Blasch, Y. Chen, D. Shen and G. Chen also provide a sentiment classification implementation with MapReduce framework in "Scalable Sentiment Classification for Big Data Analysis Using Naïve Bayes Classifier" [6]. The system requires four extra modules, which are the work flow controller, the data parser, the user terminal and the result collector based on Hadoop infrastructure. After cleaning up the datasets, including the removal of punctuations, three MapReduce jobs are launched: training, combing, and classifying. Instead of a standard library like Mahout, this model has an accuracy of about 82%.

IV. Design and Implementation

A. Description of Datasets

1. IMDb 5000 Dataset [7]

IMDb 5000 Dataset provides metadata (*movie_id*, *movie_title*, *release_year*, *color*, *duration*, *genres*, plot_*keywords*, *language*, *country*, *content_rating*, *storyline*, *movie_facebook_likes*, *imdb_score*, *cast_info*, *director_info*, *num_user_for_reviews*, *num_critic_for_reviews*, *num_voted_users*, *production_budget*, *worldwide_gross*, *domestic_gross*) of about 5,000 movies released before mid 2016 in IMDb. After cleaning and removing duplicates in the dataset, there are over 4,600 records. It has the size of about 20 MB and is collected just once.

2. IMDb Reviews Dataset [8]

This dataset contains IMDb movie IDs, user reviews and ratings of movies listed in the IMDb 5000 Dataset. It is downloaded by web crawler from the IMDb website and is collected just once. There are about 220,000 reviews on over 4,000 movies with size of approximately 200 MB.

3. Twitter Reviews Dataset [9]

This dataset has tweets sharing reviews of movies listed in the IMDb 5000 Dataset. The tweets are downloaded from Twitter API with the movie titles and is collected just once. There are about 180,000 reviews on over 4,000 movies with size of approximately 60MB.

B. Design Details

Figure 1 shows the design of this movie analytic. It can be divided to four main tasks: data ingest, cleaning and profiling, analysis and visualization.

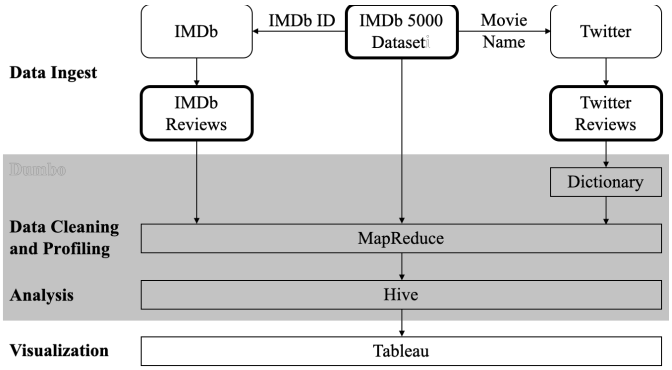

Figure 1. Design Diagram of Moviepedia

1. Data Ingest

We start with the IMDb 5000 Dataset [7]. Since the box office information may be outdated, so we run the web crawler provided by the author of this dataset to get updated *production_budget*, *worldwide_gross* and *domestic_gross* from The Numbers [10], as well as *num_user_for_reviews*, *num_critic_for_reviews*, *num_voted_user* and *imdb_score* from IMDb [8]. With the unique IMDb movie IDs listed, we use web crawler to get reviews from users for every movie, which forms the IMDb Reviews Dataset. By calling Twitter Search API with movie titles in the IMDb 5000 Dataset, we extract tweets related to the movie, which belongs to the Twitter Reviews Dataset. The three datasets are related by the unique IMDb movie ID.

The detailed introduction of each dataset is provided in the previous section. These datasets are moved to NYU Dumbo for further profiling and cleaning.

2. Data Cleaning and Profiling

For the IMDb 5000 Dataset, we apply six MapReduce jobs:

- join *movie_id*, *production_budget*, *worldwide_gross* and *domestic_gross* by same IMDb search links, and drop if multiple match with each other
- join output of the first MapReduce job with updated *num_user_for_reviews*, *num_critic_for_reviews*, *num_voted_user* and *imdb_score* from IMDb by same *movie_id*, and drop if multiple match with each other
- join output of the second MapReduce job with the original IMDb 5000 Dataset by *movie_id*, remove any duplicates, parse fields and fill empty entries with 0 or *NULL* values
- reconstruct tables containing information of movies, genres and actors/actresses, respectively, from output of the third MapReduce job, and deal with non-English characters

For the IMDb Reviews Dataset, we tokenize every review while removing stop words [11]. The following shows a detailed process of the MapReduce job:

- pass stop words file as cache files
- remove stop words from each entry

- convert the reviews into lowercases and remove other non-English characters
- generate word count for each review
- conduct an add-up onto all user ratings

For the Twitter Dataset, we remove any non-English characters, emojis and *@username* using a MapReduce job. Then each cleaned tweet is assigned with a special key for each movie. Another MapReduce job is used to remove stop words [11], calculate overall sentiment levels by checking the positive or negative attribute of each word from a tagged dictionary [12], and do word counts. The following shows the procedure of the second MapReduce job for sentiment analysis:

- pass stop words file and sentiment words dictionary as cache files
- remove stop words from each tweet
- get sentiment score for each tweet by calculating the average score for all useful words in the tweet
- generate word count for each movie with format (*movie_id*, *word*, *count*, *sentiment_score*) in mapper
- get count and add sentiment score for the same word in the same movie and output it by reducer

We use python and Java to write codes for these MapReduce jobs to clean up these datasets. The outputs are stored in HDFS on NYU Dumbo for further analysis.

3. Analysis

We implement Hive queries to study

- the trend of the development in movie industry
- the key to the success of a movie by the correlation between each pair of numerical factors
- previous great works
- previous great works by genre
- characteristics of each genre
- top directors and actors/actresses within each genre to consider when forming a film crew
- top commonly mentioned words from reviews on IMDb and Twitter by genre

The outcomes we get are stated in the RESULTS section.

4. Visualization

We use Tableau to visualize the charts and word clouds. Please refer to the RESULTS section.

V. RESULTS

We focus on the trend of development in movie industry, the key factors to the success of a movie, survey of previous popular movies, characteristics of each genre, and great actors/actresses and directors.

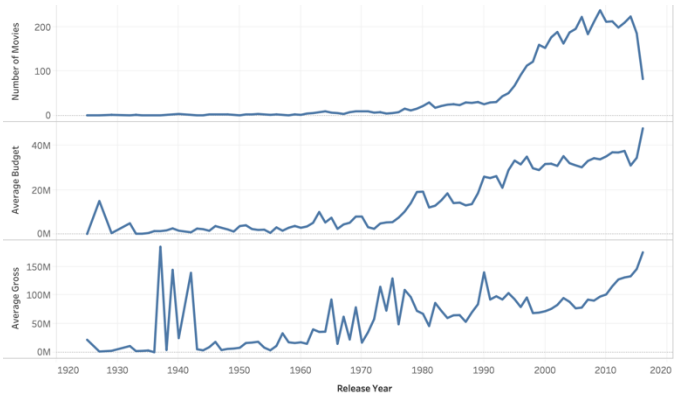

Figure 2. Development in Movie Industry

Figure 2 shows that number of movies produced, average budget and gross box office each year are gradually increasing, which means that film practitioner are shooting more, investing more, and earning more on movies. However, there is a severe drop near the tail of the curve showing number of movies released each year. This is because the IMDb 5000 Dataset was collected in the middle of 2016 [7]. We also find that the average gross box office incomes in 1937, 1939 and 1942, respectively, are extremely high. We figure out all movies released in these years in Table 1. It turns out that *Snow White and the Seven Dwarfs* (1937), *Gone with the Wind* (1939) and *Bambi* (1942) are the outliers that lead to unusual high values in the three years. We further learn that the three movies were re-released afterward [8], and the IMDb 5000 Dataset collects cumulative box office incomes [7], which contribute to the unexpected result.

Table 1. Movies Released in 1937, 1939 and 1942

Movie	Worldwide Gross (USD)
Snow White and the Seven Dwarfs (1937)	184,925,486
The Wizard of Oz (1939)	33,953,637
Mr. Smith Goes to Washington (1939)	9,000,000
Gone with the Wind (1939)	390,525,192
Casablanca (1942)	10,462,500
Bambi (1942)	268,000,000

In order to see what factors are the key to the success of a movie, we calculate a correlation matrix for all numeric fields in the IMDb 5000 Dataset (see Figure 3). From the matrix, we are able to conclude the following that movie producers can learn from:

- a movie with a higher budget tends to get more box office income (0.741)
- audience's rating on a movie is not necessarily related to its production budget (0.055)
- if a movie is popular among public (*num_voted_users*), it may have relatively higher rating to some extent (0.428)
- the popularity of cast is not necessarily related to a high rating (0.096)

These patterns can also be reflected in past great works ever until mid 2016 shown in Table 2, Table 3, Table 4 and Table 5.

	actor_1..	actor_2..	actor_3..	cast_to..	director..	duration	imdb_s..	movie_..	num_cr..	num_us..	num_vo..	product..	release..	worldw..
actor_1_fa..	1.000	0.433	0.339	0.887	0.096	0.089	0.063	0.138	0.192	0.148	0.184	0.186	0.080	0.144
actor_2_fa..	0.433	1.000	0.717	0.724	0.177	0.179	0.110	0.304	0.394	0.294	0.344	0.381	0.161	0.327
actor_3_fa..	0.339	0.717	1.000	0.692	0.215	0.204	0.112	0.366	0.463	0.359	0.412	0.475	0.204	0.402
cast_total_..	0.887	0.724	0.692	1.000	0.166	0.166	0.096	0.292	0.375	0.287	0.341	0.391	0.162	0.326
director_fa..	0.096	0.177	0.215	0.166	1.000	0.238	0.228	0.229	0.266	0.333	0.386	0.149	-0.084	0.181
duration	0.089	0.179	0.204	0.166	0.238	1.000	0.286	0.214	0.263	0.325	0.317	0.265	-0.150	0.245
imdb_score	0.063	0.110	0.112	0.096	0.228	0.286	1.000	0.288	0.325	0.316	0.428	0.055	-0.218	0.195
movie_face..	0.138	0.304	0.366	0.292	0.229	0.214	0.288	1.000	0.719	0.458	0.580	0.353	0.213	0.431
num_critic..	0.192	0.394	0.463	0.375	0.266	0.263	0.325	0.719	1.000	0.608	0.630	0.507	0.270	0.527
num_user_..	0.148	0.294	0.359	0.287	0.333	0.325	0.316	0.458	0.608	1.000	0.816	0.452	-0.003	0.566
num_voted..	0.184	0.344	0.412	0.341	0.386	0.317	0.428	0.580	0.630	0.816	1.000	0.446	0.010	0.624
production..	0.186	0.381	0.475	0.391	0.149	0.265	0.055	0.353	0.507	0.452	0.446	1.000	0.176	0.741
release_year	0.080	0.161	0.204	0.162	-0.084	-0.150	-0.218	0.213	0.270	-0.003	0.010	0.176	1.000	0.090
worldwide_..	0.144	0.327	0.402	0.326	0.181	0.245	0.195	0.431	0.527	0.566	0.624	0.741	0.090	1.000

Figure 3. Correlation Matrix of Factors

Table 2. Top 10 Money-Making Movies

Rank	Movie	Worldwide Gross (USD)
1	*Avatar* (2009)	2,776,345,279
2	*Titanic* (1997)	2,208,208,395
3	*Jurassic World* (2015)	1,648,893,208
4	*Furious 7* (2015)	1,518,722,794
5	*Avengers: Age of Ultron* (2015)	1,405,057,855
6	*Harry Potter and the Deathly Hallows: Part II* (2011)	1,341,511,219
7	*Frozen* (2013)	1,272,469,910
8	*Iron Man 3* (2013)	1,215,392,272
9	*Minions* (2015)	1,162,781,621
10	*Captain America: Civil War* (2016)	1,143,219,772

Table 3. Top 10 IMDb Rating Movies

Rank	Movie	IMDb Score
1	*The Shawshank Redemption* (1994)	9.30
2	*The Godfather* (1972)	9.20
3	*Fargo*	9.00
4	*The Dark Knight* (2008)	9.00
5	*The Godfather: Part II* (1974)	9.00
6	*The Good, the Bad and the Ugly* (1966)	8.90
7	*Pulp Fiction* (1994)	8.90
8	*12 Angry Men* (1957)	8.90
9	*Schindler's List* (1993)	8.90
10	*The Lord of the Rings: The Return of the King* (2003)	8.90

Table 4. Top 10 Popular Movies

Rank	Movie	Voted Users
1	*The Shawshank Redemption* (1994)	2,012,309
2	*The Dark Knight* (2008)	1,980,974
3	*Inception* (2010)	1,760,982
4	*Fight Club* (1999)	1,610,146
5	*Pulp Fiction* (1994)	1,570,853
6	*Forrest Gump* (1994)	1,532,783
7	*The Lord of the Rings: The Fellowship of the Ring* (2001)	1,449,146
8	*The Matrix* (1999)	1,443,721
9	*The Lord of the Rings: The Return of the King* (2003)	1,432,474
10	*The Godfather* (1972)	1,378,863

Table 5. Top 10 Popular Movies Among Critics

Rank	Movie	Reviews by Critics
1	*The Dark Knight Rises* (2012)	828
2	*Django Unchained* (2012)	810
3	*Prometheus* (2012)	805
4	*Mad Max: Fury Road* (2015)	796
5	*Batman v Superman: Dawn of Justice* (2016)	771
6	*Gravity* (2013)	757
7	*Man of Steel* (2013)	741
8	*Avatar* (2009)	723
9	*Interstellar* (2014)	716
10	*Skyfall* (2012)	712

We find that top money-making movies are shot in recent years (see Table 2), while top rating movies share a wide range of release time (see Table 3). We also discover that top popular movies among critics are recently released movies (see Table 5). This may be because the platform where critics share their views is transferring from traditional media like newspapers and magazines to internet so that IMDb is including more and more reviews from critics.

As movies may differ a lot among various genres, of which IMDb provides 26 kinds, we also state top movies by their genres in Table 6, Table 7, Table 8 and Table 9 for reference.

Table 6. Top Money-Making Movie within Each Genre

Genre	Movie	Worldwide Gross (USD)
Action	*Avatar* (2009)	2,776,345,279
Adventure	*Avatar* (2009)	2,776,345,279
Animation	*Frozen* (2013)	1,272,469,910
Biography	*American Sniper* (2014)	547,326,372
Comedy	*Frozen* (2013)	1,272,469,910
Crime	*Furious 7* (2015)	1,518,722,794
Documentary	*This Is It* (2009)	252,091,016
Drama	*Titanic* (1997)	2,208,208,395
Family	*Frozen* (2013)	1,272,469,910
Fantasy	*Avatar* (2009)	2,776,345,279
Film-noir	*The Lost Weekend* (1945)	11,000,000
Game-show	*The Bachelor*	36,882,378
History	*American Sniper* (2014)	547,326,372
Horror	*I Am Legend* (2007)	585,532,684

Music	*Alvin and the Chipmunks: The Squeakquel* (2009)	443,483,213
Musical	*Frozen* (2013)	1,272,469,910
Mystery	*Harry Potter and the Order of the Phoenix* (2007)	942,943,935
News	*Capitalism: A Love Story* (2009)	19,121,531
Reality-tv	*The Bachelor*	36,882,378
Romance	*Titanic* (1997)	2,208,208,395
Sci-fi	*Avatar* (2009)	2,776,345,279
Short	*Dolphins and Whales 3D: Tribes of the Ocean* (2008)	17,252,287
Sport	*Cars 2* (2011)	560,155,383
Thriller	*Jurassic World* (2015)	1,648,893,208
War	*American Sniper* (2014)	547,326,372
Western	*The Revenant* (2015)	532,950,503

Table 7. Top IMDb Rating Movie within Each Genre

Genre	Movie	IMDb Score
Action	*The Dark Knight* (2008)	9.00
Adventure	*The Lord of the Rings: The Return of the King* (2003)	8.90
Animation	*Spirited Away* (2001)	8.60
Biography	*Schindler's List* (1993)	8.90
Comedy	*It's Always Sunny in Philadelphia*	8.80
Crime	*The Shawshank Redemption* (1994)	9.30
Documentary	*Butterfly Girl* (2014)	8.60
Drama	*The Shawshank Redemption* (1994)	9.30
Family	*The Honeymooners*	8.70
Fantasy	*The Lord of the Rings: The Return of the King* (2003)	8.90
Film-noir	*Rebecca* (2009)	8.10
Game-show	*The Bachelor*	3.10
History	*Schindler's List* (1993)	8.90
Horror	*The Silence of the Lambs* (1991)	8.60
Music	*Samsara* (2011)	8.50
Musical	*Singin' in the Rain* (1952)	8.30
Mystery	*The Usual Suspects* (1995)	8.60
News	*Capitalism: A Love Story* (2009)	7.40
Reality-tv	*The Bachelor*	3.10
Romance	*It's a Wonderful Life* (1946)	8.60
Sci-fi	*Star Wars: Episode V - The Empire Strikes Back* (1980)	8.80
Short	*Marilyn Hotchkiss' Ballroom Dancing and Charm School* (1990)	6.70
Sport	*Friday Night Lights*	8.70
Thriller	*Fargo*	9.00
War	*Saving Private Ryan* (1998)	8.60
Western	*The Good, the Bad and the Ugly* (1996)	8.90

Table 8. Top Popular Movies within Each Genre

Genre	Movie	Voted Users
Action	*The Dark Knight* (2008)	1,980,974
Adventure	*Inception* (2010)	1,760,982
Animation	*WALL·E* (2008)	869,513
Biography	*Schindler's List* (1993)	1,038,698
Comedy	*Forrest Gump* (1994)	1,532,783
Crime	*The Shawshank Redemption* (1994)	2,012,309
Documentary	*Bowling for Columbine* (2002)	132,743
Drama	*The Shawshank Redemption* (1994)	2,012,309
Family	*WALL·E* (2008)	869,513
Fantasy	*The Lord of the Rings: The Fellowship of the Ring* (2001)	1,449,146
Film-noir	*Rebecca* (1940)	106,991
Game-show	*The Bachelor*	5,255
History	*Schindler's List* (1993)	1,038,698
Horror	*The Silence of the Lambs* (1991)	1,079,553
Music	*Whiplash* (2014)	577,293
Musical	*Frozen* (2013)	496,222
Mystery	*Se7en* (1995)	1,229,950
News	*Capitalism: A Love Story* (2009)	39,047
Reality-tv	*The Bachelor*	5,255
Romance	*Gladiator* (2000)	1,163,052
Sci-fi	*Inception* (2010)	1,760,982
Short	*Vessel* (2012)	330
Sport	*Million Dollar Baby* (2004)	568,625
Thriller	*The Dark Knight* (2008)	1,980,974
War	*Inglourious Basterds* (2009)	1,072,499
Western	*Django Unchained* (2012)	1,160,157

Table 9. Top Popular Movie Among Critics Within Each Genre

Genre	Movie	Reviews by Critics
Action	*The Dark Knight Rises* (2012)	828
Adventure	*Prometheus* (2012)	805
Animation	*Inside Out* (2015)	571
Biography	*Argo* (2012)	653
Comedy	*Suicide Squad* (2016)	665
Crime	*Drive* (2008)	694
Documentary	*Fahrenheit 9/11* (2004)	281
Drama	*Django Unchained* (2012)	810
Family	*Hugo* (2011)	690
Fantasy	*Man of Steel* (2013)	741
Film-noir	*Rebecca* (1940)	175
Game-show	*The Bachelor*	5
History	*Argo* (2012)	653
Horror	*World War Z* (2013)	658
Music	*Whiplash* (2014)	561
Musical	*Les Misérables* (2012)	493
Mystery	*Prometheus* (2012)	805
News	*Capitalism: A Love Story* (2009)	216
Reality-tv	*The Bachelor*	5
Romance	*Deadpool* (2016)	641
Sci-fi	*Prometheus* (2012)	805
Short	*Dolphins and Whales 3D: Tribes of the Ocean* (2008)	9
Sport	*Creed* (2015)	469
Thriller	*The Dark Knight Rises* (2012)	828
War	*Lincoln* (2012)	547
Western	*Django Unchained* (2012)	810

From Table 6, Table 7, Table 8 and Table 9, we may conclude that the popularity and rating vary a lot among genres, even though the movies listed are the best within their own genre. We would like to dive a little deeper to see the average performance of movies by genre. In Figure 4, it's clear that action, adventure, animation, family, fantasy and sci-fi movies generally require more investment, but may earn more. However, there is no guarantee to have an outstanding rating. Films about biography, film-noir, history, news or war may have a slightly higher rating, but earn a lot less than others.

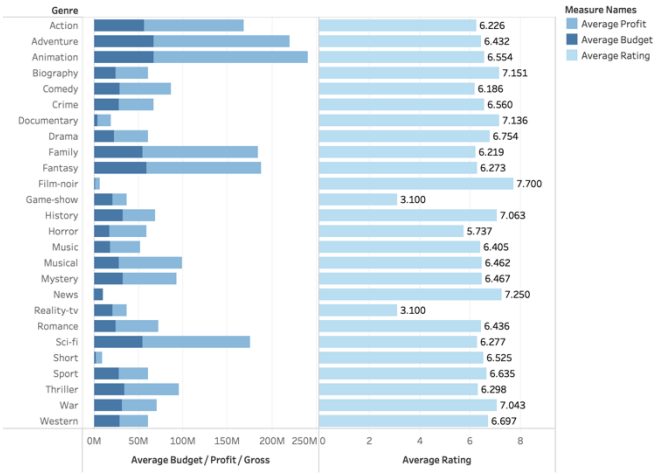

Figure 4. Average Budget, Profit and Rating by Genre

While the visualization in Figure 4 helps to select a genre, this analytic also provides some suggestions on film crew once the genre is selected. We consider both popularity and past experience on that genre of actors, actresses and directors. We calculate an index by

$$normalized\_facebook\_likes \times 20\% + normalized\_average\_box\_office\_income \times 40\% + normalized\_average\_imdb\_score \times 40\%$$

within certain genre for every actors, actresses and directors that participated ever. We take top 5 actors/actresses and directors for each genre if exist, and the results are presented in Table 10 and Table 11.

Table 10. Top 5 Directors by Genre

Genre	Top 5 Directors
Action	Christopher Nolan, Peter Jackson, Joss Whedon, Colin Trevorrow, James Cameron
Adventure	Christopher Nolan, Joss Whedon, Peter Jackson, James Cameron, Colin Trevorrow
Animation	Lee Unkrich, Chris Buck, Pete Docter, Andrew Stanton, Kyle Balda
Biography	David Fincher, Steven Spielberg, Martin Scorsese, Denzel Washington, Clint Eastwood
Comedy	Martin Scorsese, Lee Unkrich, Chris Buck, Joseph Gordon-Levitt, Andrew Stanton
Crime	Christopher Nolan, David Fincher, Martin Scorsese, Steven Spielberg, Quentin Tarantino
Documentary	Martin Scorsese, Cary Bell, Ron Fricke, Catherine Owens, Marius A. Markevicius
Drama	Christopher Nolan, James Cameron, David Fincher, Quentin Tarantino, Peter Jackson
Family	Lee Unkrich, Chris Buck, Martin Scorsese, Steven Spielberg, Alfonso Cuarón
Fantasy	James Cameron, David Fincher, Peter Jackson, Lee Unkrich, Chris Buck
Film-noir	Alfred Hitchcock, Billy Wilder, Orson Welles, Henry Hathaway
Game-show	N/A
History	David Fincher, Clint Eastwood, Steven Spielberg, Martin Scorsese, Mel Gibson
Horror	David Fincher, Ridley Scott, Stanley Kubrick, Tim Burton, William Friedkin
Music	Martin Scorsese, Kevin Spacey, Clint Eastwood, Tom Hanks, Billy Wilder
Musical	Chris Buck, Kevin Spacey, Tim Burton, Clint Eastwood, Martin Scorsese
Mystery	Christopher Nolan, David Fincher, Martin Scorsese, Steven Spielberg, Quentin Tarantino
News	Michael Moore, Sanjay Rawal
Reality-tv	N/A
Romance	James Cameron, David Fincher, Joseph Gordon-Levitt, Clint Eastwood, Tim Miller
Sci-fi	Christopher Nolan, Joss Whedon, Colin Trevorrow, Steven Spielberg, James Cameron
Short	Randall Miller, Jason Naumann, Jean-Jacques Mantello, Clark Baker
Sport	Clint Eastwood, Martin Scorsese, Sylvester Stallone, Harold Ramis, Darren Aronofsky
Thriller	Christopher Nolan, David Fincher, Colin Trevorrow, Steven Spielberg, Martin Scorsese
War	Quentin Tarantino, Clint Eastwood, Steven Spielberg, Martin Scorsese, Tony Scott
Western	Quentin Tarantino, Clint Eastwood, Alejandro G. Iñárritu, Kevin Costner, Mel Brooks

Table 11. Top 5 Actors/Actresses by Genre

Genre	Top 5 Actors/Actresses
Action	Joel David Moore, Sean Anthony Moran, Jason Whyte, Dileep Rao, Ali Astin
Adventure	CCH Pounder, Joel David Moore, Laz Alonso, Scott Lawrence, Sean Anthony Moran
Animation	Emily Hahn, Jonathan Groff, Livvy Stubenrauch, Robert Pine, Santino Fontana
Biography	Embeth Davidtz, Caroline Goodall, Mark Ivanir, Norbert Weisser, Jonathan Sagall
Comedy	Darcy Donavan, Emily Hahn, Jonathan Groff, Livvy Stubenrauch, Robert Pine
Crime	Heath Ledger, Nathalie Emmanuel, John Brotherton, Nestor Carbonell, Chin Han
Documentary	Emily Gorell, Abigail Evans, John Evans, Stacie Evans, Moise Levy
Drama	Gloria Stuart, Jonathan Hyde, Nicholas Cascone, Lewis Abernathy, Anatoly M. Sagalevitch
Family	Emily Hahn, Jonathan Groff, Livvy Stubenrauch, Robert Pine, Santino Fontana
Fantasy	Stephen Lang, Laz Alonso, Matt Gerald, Dileep Rao, Scott Lawrence
Film-noir	Laurence Olivier, Joan Fontaine, George Sanders, Judith Anderson, Gladys Cooper
Game-show	Chris Harrison
History	Embeth Davidtz, Caroline Goodall, Mark Ivanir, Norbert Weisser, Jonathan Sagall
Horror	Mads Mikkelsen, Gillian Anderson, Scott Thompson, Hettienne Park, Aaron Abrams
Music	J.K. Simmons, Miles Teller, Chris Mulkey, Austin Stowell, Paul Reiser
Musical	Alan Tudyk, Jonathan Groff, Edie McClurg, Josh Gad, Maurice LaMarche
Mystery	Alan Rickman, Robert Pattinson, Jim Broadbent, David Tennant, Elarica Johnson
News	Arnold Schwarzenegger, Wallace Shawn, Ronald Reagan, Michael Moore, Bernie Sanders
Reality-tv	Chris Harrison
Romance	Billy Zane, Gloria Stuart, Jonathan Hyde, Suzy Amis, Nicholas Cascone
Sci-fi	Joel David Moore, Laz Alonso, Sean Anthony Moran, Jason Whyte, Dileep Rao
Short	William Hurt, Elden Henson, Michael Bower, Victoria Jackson, Isaac Florentine
Sport	Taylor Kitsch, Jay Baruchel, Hilary Swank, Kyle Chandler, Adrianne Palicki

Thriller	Chris Pratt, Jake Johnson, Katie McGrath, Lauren Lapkus, Andy Buckley
War	Tom Hanks, Vin Diesel, Dennis Farina, Paul Giamatti, Edward Burns
Western	Tom Hardy, Christoph Waltz, Jamie Foxx, Kerry Washington, Ato Essandoh

We also collect commonly mentioned words from audience's reviews published on IMDb and Twitter for each genre and visualize them as word clouds. The purpose is to help movie producers and script writers to figure out what the audience like and what not. In this paper, we select reviews for action movies in Figure 5 to illustrate some patterns we find. More can be found in the Tableau visualization.

(a) IMDb Reviews

(b) Twitter Reviews

Figure 5. Commonly Mentioned Words for Action Movies

In word clouds, size encodes times mentioned in reviews, i.e., a larger size means that the word is mentioned more frequently. Meanwhile, color encodes average sentiment level, where blue denotes positive, gray denotes neutral, and red denotes negative. As stated previously, the sentiment levels of IMDb reviews are come from users' ratings where a score of 5 is set to be neutral, while we calculate the overall sentiment values for each Twitter review by looking up each word's emotion in a tagged dictionary. Thus, the word clouds for Twitter reviews seem to be more "colorful". This can be somehow explained by the average rating shown in Figure 4. People tend to rate 6 or 7 for an average movie, which makes IMDb reviews more positive.

By investigating the content shown in the word clouds, we find that people are more likely to analyze elements of movies in IMDb reviews, such as "story", "plot", "scenes" and "characters" for action movies. However, people state more about the plot, or the content in their tweets. Thus, movie producers and script writers may figure out the elements to care about from word clouds of IMDb reviews, and common imageries or clichés that might be avoided from Twitter reviews.

VI. Future Work

In this analytic, the majority of tasks that we have done is numerical analysis. We would like to include more text mining and prediction components in the future. We may cluster similar movies by their storylines, and thus make script writers and movie producers inspired from extracted keywords for each clustered group. We would also like to keep track of Google Trends, and make prediction about the potential level of a movie's success, and this may require a long-term monitoring to form the dataset.

A similar analytic can also be applied to music. We may get popular songs, artists, albums, etc. as well as commonly used elements or imageries in lyrics regarding to different styles.

VII. Conclusion

This analytic provides insights in movies by genres through analyzing gross box office incomes, user ratings, numbers of Facebook likes, storylines, commonly mentioned keywords from reviews and their sentiments. While some unexpected results do exist, we are able to explain them with further investigation. We also provide some suggestions on how to interpret the results and utilize the data to be evidences of decisions. We hope this analytic could inspire movie producers and script writers to come out more good movies.

Acknowledgment

We would like to appreciate Prof. Suzanne McIntosh from NYU for her help and guidance on this project.

We would like to appreciate Chuan Sun for providing the IMDb 5000 Dataset.

References

1. T. Ashwitha, A. Rodrigues and N. Chiplunkar. Movie Dataset Analysis using Hadoop-Hive. In 2nd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions, 2017.
2. G. Mishne and N. Glance. Predicting Movie Sales from Blogger Sentiment. Accessed at https://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-030.pdf.
3. M. Mestyán, T. Yasseri and J. Kertész. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. Accessed at https://doi.org/10.1371/journal.pone.0071226.
4. R. Paul. Big Data Analysis of Indian Premier League using Hadoop and MapReduce. In 2017 International Conference on Computational Intelligence in Data Science (ICCIDS), 2017.
5. M. Kumar and A. Bala. Analyzing Twitter Sentiments Through Big Data. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016.
6. B. Liu, E. Blasch, Y. Chen, D. Shen and G. Chen. Scalable Sentiment Classification for Big Data Analysis Using Naïve Bayes Classifier. In 2013 IEEE International Conference on Big Data, 2013.

7. C. Sun. Predict IMDb Movie Rating. Accessed at https://github.com/sundeepblue/movie_rating_prediction.
8. IMDb. IMDb Website. Accessed at https://www.imdb.com/.
9. Twitter. Twitter Developer. Accessed at https://developer.twitter.com.
10. The Numbers. The Numbers - Where Data and the Movie Business Meet. Accessed at https://www.the-numbers.com.
11. K. Bougé. Stop Words. Accessed at https://sites.google.com/site/kevinbouge/stopwords-lists.
12. HLT - Natural Language Processing. SentiWords. https://hlt-nlp.fbk.eu/technologies/sentiwords.