Technical Report

Mike Ma[1] *

[1]SIRC. Ontario Tech University, ON, CA

We collected raw data of all movie information(81270 movies) from IMDb during 2013-2019; all trailer information(8616 trailers) from 9 channels of YouTube during 2013-2019; trailer reviews(864551 reviews) from 10 movie categories of the biggest YouTube channel "Movieclips Trailers"; as well as 3000 trailer videos from 10 movie categories of the biggest YouTube channel "Movieclips Trailers". For the research purpose of predicting movie gross income and popularity by trailer information, we matched the above all trailer information and all corresponding movies together, in order to match meta-data of trailer and movie. We used sentiment analysis tool TextBlob to calculate the sentiment and subjectivity of each trailer comment. By multiplying 'sentiment' and 'commentlike' count, we got 'preference' of each trailer. We also put forward a data normalization method to solve the time inconsistency problem. By co-efficiency analysis, We finally found that the trailer comments information has a much better performance and reliability than simple intuitive data 'like/dislike' to predict the movie gross income.

Motivation

The success or failure of a movie is often determined in its first weekend of play. In order to make this opening successful, movie producers must employ a number of promotional strategies including movie trailers, to publicize the movie for a significant period of time prior to its release. A movie trailer, as described in Wikipedia, is an advertisement or a commercial for a feature film that will be exhibited in the future. Of some ten billion videos watched online annually, film trailers rank the third, after news and user-created video. It is clear that businesses have a strong interest in tapping into this huge data source to extract information that might improve their decision making process. For example, predictive models derived from movie and its trailers may facilitate filmmakers making more profitable decisions.

In our research, by analyzing the correlation between the meta-data of trailer and movie, we aim to derive a predictive model from movie and its trailer meta-data, and focus on the prediction of movie gross income and movie popularity using trailer information as an effective pre-released meta-data.

RELATED WORK

For details, see 'Literature Review of Trailer Research'

Predicting the Gross income(including the stock price) of movie.

Research has been done to generate models for predicting revenues of movies. Most of them derived results from single data sources.

Specifically, Joshi and others[1] use linear regression that joined meta-data with text features from pre-release critique to predict earnings for movies with a coefficient of determination R^2=0.671.

Mishne and Glance[2] correlate sentiments in blog posts with movie box-office scores. The correlations they observed for positive sentiments are fairly low and not sufficient to use for predictive purposes. They neither build prediction models or show the value of the correlation because they think the result is not good enough for accurate modeling.

Zhang and Skiena[4] have used a news aggregation model along with IMDb data to predict movie box-office numbers.

In a very interesting approach Asur and Huberman set up a prediction system for the revenue of movies based on the volume of Twitter mentions[5]. They build a linear regression model based on the chatter of Twitter and achieve an adjusted coefficient of determination of 0.97 on the night before the movie release for the first weekend revenue of a sample of 24 movies. In addition, they even tried to predict the Hollywood Stock Price given that social media can accurately predict box office results and the Hollyhood Stock Exchange adjusts the price for a movie stock to reflect the actual box office gross. They tested social media data efficacy compared with historical HSX prices at forecasting the stock prices of the HSX index and their tweet-rate proves to be significantly better at predicting the actual stock value than the historical HSX prices. It's a good inspiration of considering stock price because according to [6], Prices of movie stocks accurately predict real box office results, which demonstrates the strong correlation between the movie stock price and real box office gross income.

In a later work, however, Wong et al. show that Tweets do not necessarily represent the financial success of movies[7]. They consider a sample of 34 movies and compare the Tweets about the movies to evaluations written by users of the movie review websites. They argue that predictions based on social media could have high precision but low recall.

Predicting the rating(popularity) of movie.

Sharda and Delen[3] have treated the prediction problem as a classification problem and used neural networks to process pre-release data, such as quality and popularity variables, and classify movies into nine categories ranging from 'flop' to 'blockbuster'. Apart from the fact that they are predicting ranges over actual numbers, the best accuracy that their model can achieve is fairly low(36.9%).

Marton and Taha[8] have showed that the populartity of a movie can be predicted much before its release by measuring and analyzing the activity level of editors and viewers of the corresponding entry to the movie in Wikipedia. It's novel because it is the only research using data from wikipedia, but the data features are too simple and low to support their conclusion.

In a rather novel approach, Oghina et al. have made use of Twitter and YouTube activity streams to predict the ratings in the Internet Movie Database(IMDb), which is among the most popular online movie database[9].

Reference[10] described a movie rating approach based on data mining of 240 movies from IMDb where Weka and J48 were used to create the prototype model.

Reference[11] also applied Weka and J48 to generate three classes of movies: Hit, Neutral or Flop, in order to predict the movie box office performance. It's amazing that they also use trailer information and their data is quite similar to us and give us inspirations of more possible data features. They generate a model consisting of genre of successful movies ranked by user ratings of the IMDb, popularity of director, leading actor and leading actress represented by the number of comments and views of official movie trailers accessible by YouTube, and sentiment toward a movie derived from YouTube viewers' comments. However, their conclusion is not reliable considering their only 35 movies data.

In [12], methods to predict the popularity of movies were discussed to evolve as a guiding strategy for Content Distribution/ Delivery Network(CDD). Actor and director popularity were considered as base criteria for predicting the popularity of a movie.

Predicting the popularity of trailer.

The only research focusing on trailers is [14], they analyzed whether subjective multimedia features be developed to predict the viewer's preference presented by like or dislike during and after watching movie trailers. The results showed that the single low-level multimedia feature of shot length variance is highly predictive of a viewer's "like/dislike" for a large portion of movie trailers. However, their features are too narrow and data is from only 1375 trailers. There are still lots of features to demonstrate the popularity of trailers. Moreover, there has been substantial interest in the NLP community on using movie reviews as a domain to test sentiment analysis methods. e.g.,[13], et al. Basically speaking, they apply information retrieval or machine learning techniques to classify movie reviews into some categories and hope to produce better classification accuracy than human being. The classification categories are like "thumbs up" vs."thumbs down", "positive"vs."negative", or "like" vs."dislike".

Video shots analysis.

In the video shots analysis part, researches could be classified as two categories:
1. Relationship between Mediaand Audiences' Affecting State:
Related research focusedon the emotion recognition of videos or movies. In these articles, some electronic signals, such as electroencephalogram(EGG), facial images, color features, the combination of audio and visual features and so on, are employed as the fundamental data for registering the viewer's affecting state.
2. Visual Data Feature Extraction:
Shot segmentation and key-frame were extracted and the lighting key cues, motion,shot density, color energy cues, and other miscellaneous cues, including some audio data, were used as the features to predict the affecting potential of a film.[14]

Brief analysis of related works.

1. Most of the works are focusing on movie gross prediction. Different people work on movie gross prediction from different perspectives. Most previous work forecast movie grosses based on IMDb data with regression or stochastic models. However, their models either work poorly or need post-release data in order to make reasonable prediction, which are not acceptable in practice, because it is difficult to give shape estimation for either model parameters or gross if they don't have any early stage movie gross data. Although the post-release models are also useful in some situations, pre-release models are of more practical importance. Luckily, trailer is a good pre-release data source but seldom considered in related works.

2. While there has been research on predicting movie sales, almost all of them have used meta-data information on the movies themselves to perform the forecasting, such as the movies genre, MPAA rating, running time, release data, the number of screens on which the movie debuted, and the presence of particular actors or actresses in the cast. Trailer is seldom considered in related works.

3. Since predictions based on classic quality factors fail to reach a level of accuracy high enough for practical application usage of user-generated data to predict the success of a movie becomes a very temping approach. It indicates that sentiments analysis from reviewers' comments are worth to do.

4. Most predictions work on using movie data to predict movie performance. Seldom predictions work on using trailer data to predict movie performance. Only one prediction works on using trailer data to predict trailer performance, while their features are too simple(only like/dislike) and data size are too small.

5. Most predictions work on movies focuses on forecasting revenues, not ratings. Seldom predictions work on movies focuses on ratings. Only two predictions work on trailers.

However, trailers are good pre-release data but ignored by many.

6. Predicting movie stock price is worth to try because the strong correlation of movie performance and movie stock price. Most papers ignored this point.

7. Video shots analysis is not practical and not necessary in current stage considering its complexity and lots of simpler problems need to focus on. What also well worth mentioning is that almost all papers about video shots analysis get no more than 30 citations.

Analysis and Inspiration.

1. Improving Movie Gross Prediction Through News Analysis[4]: Their sentiment statistics are good. They derive several sentiment measures, including polarity, subjectivity, positive references per reference, and positive-negative differences per reference and give their definitions.

2. Early Prediction of Movie Box office Success Based on Wikipedia Activity Big Data[8]: They use features both individually and combined to repeat experiments, such as T,V,S,V,T,S,V,T,V,S. And they indicate that applicability of prediction model on movies with medium and low popularity levels remains an open question.

3. Predicting IMDB movie ratings using social media[9]: Their data is similar to us, including surface features from platform Youtube and Tweeter, as well as textual features from tweets as well as Youtube comments.

4. The real power of artificial markets[6]: It indicates that the prices of movie stocks could accurately predict real box office results.

5. Predicting movie sales from blogger sentiment[2]: They analyze both pre-release data and post-release data with good sentiment analysis methods. In addition, they focus on the opening weekend data rather than total sales since it normalizes the figure across movies that were released on different dates, preventing earlier movies from having a higher income just because they have been "out there" longer.

6. Prediction of Movies Box Office Performance Using Social Media[11]: Their data are almost the same to us. It is the only paper use both movie data and trailer data to predict the performance of movie. Their provides inspiration in new data, including genre frequency, the popularity(followers) of actor and director. They also differentiate the movie by if it is sequel movie or the first-version movie like us. In addition, their NLP methods toward Youtube comments are also worth to copy.

7. Predicting Movie Trailer Viewer's Like/Dislike via Learned Shot Editing Patterns[14]: It is the only paper use trailer data to predict trailer data, their experiment process and charts are worth to copy.

8. Predicting the Future With Social Media[9]: The whole paper is good and citations are above 8000. They analysis the relationship between tweets-rate, sentiments of tweets and movie outcomes as well as movie stock price. Their data is from Tweet. Their method could be repeated with our Youtube comments-rate and sentiments of Youtube trailers comments, and we could expect a good outcome.

DATASET CHARACTERISTICS

Raw Data.

We collected raw data of all movie information(81270 movies) from IMDb during 2013-2019; all trailer information(8616 trailers) from 9 channels of YouTube during 2013-2019; trailer reviews(864551 reviews) from 10 movie categories of the biggest youtube channel "Movieclips Trailers"; as well as 3000 trailer videos from 10 movie categories of the biggest youtube channel "Movieclips Trailers".

Matched Data.

For the research purpose of predicting movie gross income and popularity by trailer information, we matched the above all trailer information and all corresponding movies together, in order to use trailer information as pre-released meta-data to predict the movie gross income, rating, and popularity. We got merged data table consists of both trailer info and movie info.

Sentiment Preference.

We used sentiment analysis tool TextBlob to calculate the sentiment and subjectivity of each trailer comment. By multiplying 'sentiment' and 'commentlike' count, we got 'preference' of each trailer. By normalizing the comment time and movie publish time, we saw the reliability of 'preference' compared to non-normalized simple 'like/dislike'.

Coefficient analysis.

By coefficient analysis, we tested the correlation between 'movie rating' and meta-data of trailer, correlation between 'movie USA gross income' and meta-data of trailer, as well as 'movie Worldwide gross income' and meta-data of trailer.

For dataset details, see 'Data introduction report'.

TITLE MATCHING of TRAILERS MOVIES

After we collected the raw data of trailer info from YouTube and movie info from IMDb, it's necessary to solve the title matching problem, and merge the trailer info and movie info into one table.
The title matching rate is about 70%, which means about 70% trailer could match the correspondent movie, based on the dataset 8616 trailers 81270 movies.

Method of title matching.

1. Export the trailer data table and movie data table from

MySQL into Json.
2. Create list for each field.
3. Use re.match() to match.
4. Check the format of unmatched part, if it included subtitle or series number, the matched part would not be considered.
5. store all the fields into one database.

TITLE REVIEWS SENTIMENT ANALYSIS

Textblob Introduction.

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

Sentiment Analysis Experiment.

TRAILER REVIEWS DATA NORMALIZATION

Major problem.

The only paper focusing on trailer prediction is “Predicting movie trailer viewer’s “like/dislike” via learned shot editing patterns”, However, they directly use “like/dislike” in YouTube to do the prediction, which has a problem. Think about this:

Movie 1 is released 3 days ago, whose trailer has 400 likes.
Movie 2 is released 3 years ago, whose trailer has 4000 likes.

However, we couldn’t say movie 2 is better than movie 1, while in their paper they ignored this problem.
To solve the time problem(described in technical report) and normalization problem, we use sentiment analysis to calculate the preference from each trailer review to compare the performance and reliability of simple ’like dislike’ for prediction in previous papers.

Method Introduction - Reviews Data Normalization.

1. Get “movie release time” and “trailer comments time” from the title-matched table and comments table.

2. Transfer “trailer comments time” (’1 days ago’ transfered to ’2 days until the data collecting time,2018-08-01’).
0-24 hours ago — 2019-07-30 (1 days)
1 days ago — 2019-07-29 (2 days)
1 weeks ago - 2019-07-18(13 days)
1 months ago - 2019-06-02(59 days)
1 years ago - 2017-08-02(729 days)

3. Compare “movie release time” and “trailer comments time” to current data-collecting time(2019-08-01) individually. Calculate the differences:
Difference 1 = 2019-08-01 - “movie release time”
Difference 2 = 2019-08-01 - “trailer comment time”

4. Prerequisites for comments selection
(1)Guarantee the review of trailer before the movie released(guarantee the trailer reviews are pre-released data).
Difference 2 - Difference 1>0
(2)Set the comments selection period as 6 months(180 days) before the movie released time.
0 <Difference 2 - Difference 1 <180
(3)Delete trailer info which is released less than 6 months before data collecting time(2019-8-1).
Difference 2 >180

5. Comments selection
Using the value of “difference 2 - difference 1”, we could normalize each Trailer review in a fixed period of time(6 months)

CORRELATION ANALYSIS

In this experiment, we calculated the co-efficiency between the movie rating and meta-data of trailer, co-efficiency between the movie USA gross income and meta-data of trailer, co-efficiency between the movie worldwide gross income and meta-data of trailer.

The meta-data of trailer includes ’trailer series’, ’view count’, ’like count’, ’dislike count’, ’comment count’.

Correlation Results.

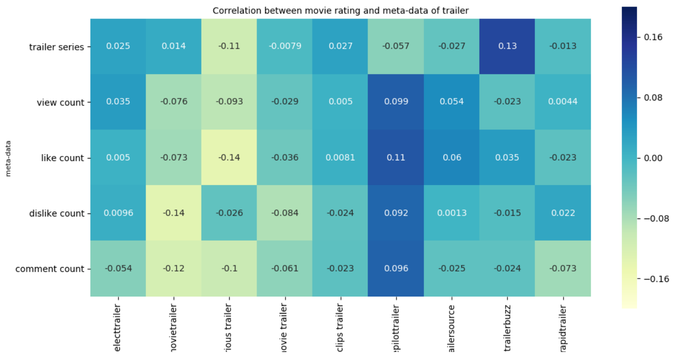

Fig. 1. Correlation between movie rating and meta-data of trailer

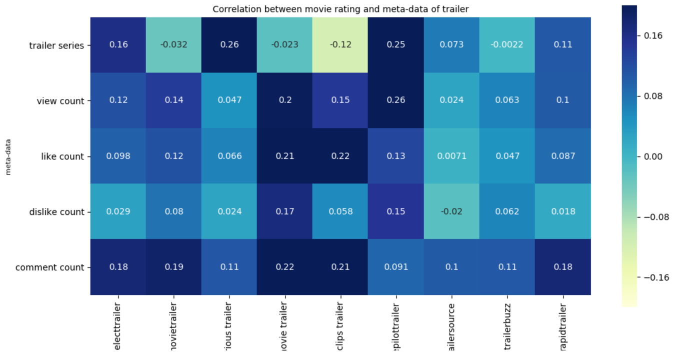

Fig. 2. Correlation between movie USA gross income and meta-data of trailer

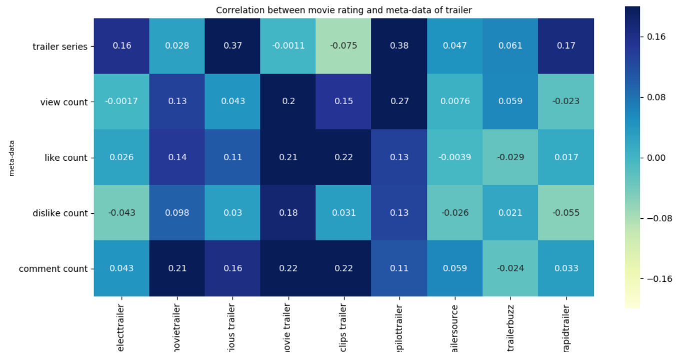

Fig. 3. Correlation between movie USA gross income and meta-data of trailer

Correlation Analysis.

Be aware that each channel has different trailer amount and subscribers, which means we should not compare the co-efficiency by columns. Rather, the below analysis based on the comparison between the rows.

According to the visualization above, we could not see any Characteristics from Fig.1. In Fig.2. and Fig.3., the color in the last row is much darker than the above 4 rows, which means "trailer comments count" has a much higher co-efficiency (more than 10 times) with "movie gross income" than simple "view count", "like count", and "dislike count".

From this result we could draw the conclusion that The comments preference value from trailer comments may be much better than simple "view count", "like count", "dislike count" to predict the movie gross income. And with the method to normalize the reviews time period suggested above, the comments preference value is much more reliable than simple meta-data.

By way of coefficient analysis, we preliminary found the results as above. From the coefficient analysis result above, we could draw the conclusion that 'comments count' has a much higher correlation with USA gross income and world-wide gross income than simple meta-data like 'view count', 'like count', and 'dislike count'.(about 10 times higher). Therefore we believe that the preference from trailer reviews have much higher performance than simple meta-data of trailer in YouTube. We would prove the assumption in further co-efficiency analysis.

CURRENT CONCLUTION

By coefficient analysis, "trailer comments count" has much higher coefficiency(more than 10 times) with "movie gross income" than simple "view count","like count", and "dislike count". according to our method to calculate the trailer comments 'preference' with Textblob, the comments preference value of trailer may have much better performance than simple "view count", "like count" "dislike count" to predict the movie gross income, and possibly the movie stock price. And with the method to normalize the reviews time period, comments preference value is also much more reliable than simple meta-data. We make up for the time problem of data in the past literature.

In our further research, we aim to derive a predictive model from movie and its trailer meta-data, and focus on the prediction of movie gross income using trailer comments information. Test the conclusion from our correlation analysis.

References

1. M. Joshi, D. Das, K. Grimpel, and N. A.Smith, "Movie revies and revenues: An experiment in text regression." *NAACL-HLT*, 2010.
2. G.Mishne and N.Glance, "Predicting movie sales from blogger sentiment." in *In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.
3. R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks." in *Expert Systems with Applications*, vol. 30, 2006, pp. 243–254.
4. W.Zhang and S.Skiena, "Improving movie gross prediction through news analysis." in *In Web Intelligence*, 2009, p. 301304.
5. H. B. Asur S, "Predicting the future with social media." in *In Processings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010, pp. 492–499.
6. D. M. Pennock and S. Lawrence, "The real power of artificial markets." in *Science: American Association for the Advancement of Science.*
7. W. FMF, S. S, and C. M, "Why watching movie tweets won't tell the whole story?" in *In Processings of the 2012 ACM workshop on Workshop on online social networks. New York, NY, USA: ACM, WOSN'12*, 2012, pp. 61–66.
8. M. Mestyan and T. Yasseri, "Early prediction of movie box office success based on wikipedia activity big data," in *PLoS ONE 8(8): e71226. doi:10.1371/journal.pone.0071226*, 2013.
9. O. A, B. M, T. E, and de Rijke M, "Predicting imdb movie ratings using social media." in *In: ECIR 2012: 34th European Conference on Information Retrieval. Springer-Verlag, Barcelona, Spain: Springer-Verlag*, 2012, pp. 503–507.
10. S. Kabinsingha, S. Chindasorn, and C. Chantrapornchai, "Movie rating approach and application based on data mining," in *International Journal of Engineering and Innovative Technology(IJEIT)*, 2012.
11. K. R. Apala and M. Jose, "Prediction of movies box office performance using social media," in *2013 IEEE/ACM International Conference on Adances in Social Networks Analysis and Mining*, 20123,.
12. L. Zhu, M. Zhu, and S. Yao, "The popularity of movies predict system based on data mining technology for cdn." in *IEEE International Conference on the 3rd Computer Science and Information Technology*, 2010.
13. P. Chaovalit and L. Zhou, "Movie review mining: a comparison between supervised and unsupervised classification approaches." in *In Proceedings of the Hawaii International Conference on System Sciences(HICSS)*, 2005.
14. Y. Hou, T. Xiao, and S. Zhang, "Predicting movie trailer viewer's like/dislike via learned shot editing patterns," in *IEEE Transactions on Affective Computing*, no. 1, 2016.

Table 1. Correlation between movie rating and meta-data of trailer

Rating correlation	trailer series	view count	like count	dislike count	comment count	sentiment	preference
filmselecttrailer	0.02533283	0.03526385	0.0050197	0.00958571	-0.053747		
freshmovietrailer	-0.01440181	-0.07580638	-0.07292159	-0.13561284	-0.11801474		
furious trailer	-0.10815836	-0.093043	-0.14001656	-0.02585848	-0.10327457		
joblomovietrailer	-0.00786402	-0.0292394	-0.03607939	-0.08427883	-0.06071927		
movieclips trailer	0.02700788	0.00504539	0.00806546	-0.02381409	-0.02317481		
moviepilottrailer	-0.05652897	0.09932129	0.1084559	0.09204856	0.09605606		
movietrailersource	-0.0271737	0.05370674	0.06031079	0.00127742	-0.02493521		
newtrailerbuzz	0.12772609	-0.022946	0.03457465	-0.01498928	-0.02350908		
rapidtrailer	-0.01301163	0.00444308	-0.0230677	0.02161095	-0.07269746		

Table 2. Correlation between movie USA gross income and meta-data of trailer

USA Gross Income	trailer series	view count	like count	dislike count	comment count	sentiment	preference
filmselecttrailer	0.16238215	0.11563359	0.09771129	0.02928777	0.17571305		
freshmovietrailer	-0.03247056	0.14004506	0.11536722	0.07975358	0.1853188		
furious trailer	0.2574477	0.04731082	0.06555064	0.0238418	0.10809361		
joblomovietrailer	-0.02304071	0.19566287	0.20924704	0.17024959	0.22012332		
movieclips trailer	-0.11687523	0.14675415	0.21936088	0.05769043	0.2123239		
moviepilottrailer	0.24708663	0.2620017	0.12796143	0.14920888	0.09145186		
movietrailersource	0.07302531	0.0243072	0.00707041	-0.02030733	0.10357889		
newtrailerbuzz	-0.00220898	0.06282479	0.0468655	0.0620744	0.10764305		
rapidtrailer	0.10672939	0.10335676	0.08667873	0.01767974	0.17640523		

Table 3. Correlation between Worldwide USA gross income and meta-data of trailer

Worldwide Gross	trailer series	view count	like count	dislike count	comment count	sentiment	preference
filmselecttrailer	0.1573731	-0.00174282	0.02645849	-0.04269658	0.0427057		
freshmovietrailer	0.02815839	0.13289882	0.14492246	0.0978303	0.21468858		
furious trailer	0.37442343	0.04289278	0.10565868	0.0302329	0.16334648		
joblomovietrailer	-0.0011454	0.20321018	0.21359551	0.18066378	0.22128298		
movieclips trailer	-0.07532938	0.14924679	0.22110975	0.03148821	0.22409317		
moviepilottrailer	0.37532633	0.26685664	0.13342402	0.13186306	0.11057562		
movietrailersource	0.04660466	0.00761057	-0.00388638	-0.02634109	0.05929984		
newtrailerbuzz	0.06052892	0.0593351	-0.02875914	0.0214751	-0.02366493		
rapidtrailer	0.17006521	-0.02297426	0.01676616	-0.05546407	0.03273731		

Table 4. Subscriber and data amount of Rating experiment channels in YouTube

Rating Analysis	Data amount	Subscriber
filmselecttrailer	225	3.08M
freshmovietrailer	233	4.16M
furious trailer	183	3.23M
joblomovietrailer	394	2.08M
movieclips trailer	1092	14M
moviepilottrailer	266	818K
movietrailersource	232	1.3M
newtrailerbuzz	168	1.14M
rapidtrailer	253	794K

Table 5. Subscriber and data amount of USA gross income experiment channels in YouTube

USA gross income Analysis	Data amount	Subscriber
filmselecttrailer	66	3.08M
freshmovietrailer	139	4.16M
furious trailer	87	3.23M
joblomovietrailer	233	2.08M
movieclips trailer	952	14M
moviepilottrailer	116	818K
movietrailersource	97	1.3M
newtrailerbuzz	85	1.14M
rapidtrailer	57	794K

Table 6. Subscriber and data amount of Worldwide gross income experiment channels in YouTube

Worldwide gross income Analysis	Data amount	Subscriber
filmselecttrailer	225	3.08M
freshmovietrailer	233	4.16M
furious trailer	183	3.23M
joblomovietrailer	394	2.08M
movieclips trailer	1092	14M
moviepilottrailer	266	818K
movietrailersource	232	1.3M
newtrailerbuzz	168	1.14M
rapidtrailer	253	794K