Literature Review of Trailer Research

Mike Ma[1] *

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

We have read more than 20 papers and select 10 of related ones to do analysis in our Literature Review. The first part is the related works on movie and trailer prediction. The second part is a brief analysis of related works. The third part is an analysis of worthy papers, focusing on inspiration to our research. The final part is potential research points for our research and updated proposed plan.

Related Works.

Movie and trailer prediction researches could be classified as three categories:
1.Predicting the Gross income(including stock price) of movie.(90%)
2.Predicting the rating(popularity) of movie.(15%)
3.Predicting the popularity of trailers(5%).

The directions could also be classified as three categories:
1.Using direct data(related pre-release and post- release data of trailers and movies) to build machine learning models and do correlation analysis.
2.Sentiment Analysis of comments of movies(not comments of trailers) using shallow NLP and machine learning.
3.Video shots analysis.

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 $R\hat{2}$=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.
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%).
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.
Marton and Taha[8] have showed that the popularity 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] describes 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.

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".
In the video shots analysis part, researches could be classified as two categories: 1. Relationship between Media and Audiences' Affecting State: Related research focused on 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 on 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. Predict 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. Therefore, their experiment process is worth to study, although in deep learning we don't need correlations analysis like them. And although in their paper there is nothing about prediction as they said, all are correlations.
4. The real power of artificial markets[6]: It indicates that the prices of movie stocks accurately predict real box office results.
5. Predicting movie sales from blogger sentiment[2]: They analysis both pre-release data and post-release data. Their sentiment analysis methods are worthy. In addition, they focuse on the opening weekend data rather than total sales since this 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. I believe it is the third most second most related paper among all to us.They give me inspiration of 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. I strongly recommend to use the same way to differentiate our trailer, and give each trailer a sequel number. I would do this in data-cleaning process. Besides, their NLP methods toward Youtube comments are also worth 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. I believe it is also the second most related paper among all to us.
8. Predicting the Future With Social Media[9]: To be honest, this is my favourite paper among all those. 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. I even want to copy all his methods, because their data could be replaced by my Youtube comments-rate and sentiments of Youtube trailers comments. I could even add more pre-release data from trailers to further the prediction. I am highly expect the potential results.

Potential Points.

1. Given meta-data about a movie and trailer, predict the popularity of the trailer after a fixed period of time.
2. Given meta-data about a movie and trailer, predict the popularity(ratings) of movie after a fixed period of time.
3. Given meta-data about a movie and trailer, predict the gross income of movies and the stock price of movie after a fixed period of time.
4. Test whether the sentiment of user-generated comments correlates with the movie's box office information better than simple counts of like/dislike from Youtube. Make predictions on movie popularity and stock price in terms of sentiments of user-generated comments. Dataset would be processed using a set of positive and negative words and then classified as positive, negative or neutral. Use sentiments data as new predictors.

Proposed Plan.

1. Collect Related Data(July 23 - Sep 1)

- Description, duration, viewCount, likeCount, dislikeCount, favoriteCount, commentCount, Stars, Directors, ReleaseTime, Synopsis from about 10K trailers from 10 channels in youtube, including different language trailers.
- About 100K Comments of all 8 categories trailers from the biggest channel in youtube.
- About 3K all 8 categories trailers videos from the biggest channel in youtube, 720K, 40GB.
- About 1000 trailers Sharing info from the trailer channel
- Title, Link, Certificate, Content, rating, runtime, Director, Stars, Writers, Genre, PublishYear, Metascore, Gross revenue, Headline, Plot Keywords, certificate, Official Sites, Country, Language, Filming Location, Gross USA, Cumulative World Gross, Production Co, Sound Mix, Color, Aspect Ratio, Discription, TrailerRuntime, UK Release Time, Budget, Storyline of about 1K Movies from 2011-2020 movie lists in IMDb

2. Data Cleaning and save data to Cloud(Sep 1-Sep 8 if possible)
3. Read Papers and finish literature review.(Aug 19-Sep 1)
4. Learn deep learning, NLP and build models(Sep 1-Sep 10)
5. Begin Experiment.(After data cleaning)
6. Write a formal paper.(Maybe after I go back, I'm not ambiguous about this)

Bibliography.

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.