SRAS: A Social-respected Analysis System for Modern Movie Investment

ABSTRACT

The return on investment (ROI) analysis about movies is a challenging but important task for movie investors in their decision making process. For investors, they are expected to get high box-office revenue with appropriate investments. Apart from movie special effects, most of investments are used to remuneration for movie actors(actress). Due to complicated factors such as audience reactions, screenplay, film types etc, it's not easy for investors to estimate upfront ROI, which makes film investment a gamble. Although most current research works aim to predict box-office revenue more specific, they can not provide investors more direct profits information about the success of the movie they would like to invest.

In this paper, we design and implement an integreated system, called Social Network Investment Platform (a.k.a SRAS), that aims to do some analysis of movies for investors. SRAS provides various modules for users to predict box-office, capture public opinion about a film, assess the value of the actor, monitor the changes of the ratio of box-office about actors, assess actor replacement and so on.

KEYWORDS

Data Analysis, Targeted Sentiment Analysis, Box-office Predicting

1 INTRODUCTION

In modern society, watching movies is becoming a central part of consumer culture which makes the great earning of investments about movies. However, how to evaluate the reasonability of investments becomes an important problem for investors. Owing to complex movie markets, the success of a movie is closed related to audience reactions, screenplay, movie types and famous cast and directors. Before a movie releases, we cannot guarantee it can be popular or not.

In recent years, with the heat of the movie market, large amounts of money pour into movie industry. Investments about movies consist of cast's and directors' salary, movie itself production. advertising fees etc. Actors and directors salary occupies a large proportion in cost of movies. The immaturity of the Chinese film market itself leads to a high-risk, high-reward investment environment.Many investors determine the investment prospects of the film is based on the recent case of the same type of movie, but they do not have in the analysis of the case from the aspect of artistic creation (such as the type of film plot, actor and director of skill etc.) the ability to summarize experience of selling products, most will eventually reduced to the accumulation of the casting and some business element. This is a lot of investment some important reasons can be started the project and ultimately a huge loss.

Recently, the group acts of "fans" do not only create a great social effect, but also produce a huge economic benefit. Without bankable stars the film script aroused no interest. In order to pursue high profit, the usual way is to use a large number of "celebrities" to promote the box office through the fan effect but only a few were success. But it has caused the soaring star worth, the consequent increase in production costs. The reason is that in spite of massive "fans", the degree of match between fans and movies is the most important factor affecting the box office. Another is that box office is decided by actors, scripts, publicity, release schedule and other aspects, actors' contribution to the movie cannot be equal to box office. The lack of an effective quantitative method to assess the creative record of the box office contribution prevent investors from assess the investment star's rationality.

Besides the cost of production about movies, investors are concerned about box-office revenue. The research on movies maily focuses on movie box-office prediction[1, 2, 21]. They use linear regression, neural networks[25] and other models to predict box office more accurately by constructing multiple features, for instance, Google [23] uses search, ad click data, and line rows to predict box office and gets 94% accuracy. Sentiment analysis in movie reviews [17, 20, 24] is another aspect to do some analysis in movies and many experiments[10, 13] prove that reviews data can improve box-office predicting accuracy.

In recent years, the research on movie mainly focuses on the construction of user-oriented movie recommendation system[7, 30]. However, those work are too simple to do some analysis on the components of revenue further. Generally speaking, investors are anxious to have a tool or a platform to do a more intuitive analysis of a movie's return on investment. In general, an investor-oriented movie analysis platform is for the Chinese movie market, providing an investor-oriented integrated emotional analysis, box office prediction, actor evaluation, box office contribution analysis of the analysis of the box office analysis of the movie, focusing on the protagonist, director, audience impact on the box office.

1.1 Solutions and Challenges

As methods developments in machine learning especially text mining increasing rapidly, much data mining techniques are proved to be effective method in both natural language process (NLP) and deep learning (DL). Social-respected analysis system(SRAS) is an end-to-end platform which aim to integrate social data of movies when analysing their investment value and potential future. SRAS makes great contribution in mining movie industry data from three points of view. We summarized it the actress played by actor, box-office pay back for investors, and sentiment reflect by audience. We proposed target-depend sentiment analysis for sentiment side, dynamic predition for box-office and finally the heteogenous network for investor. The key problem SRAS focused on are listed below.

DIFFICULTY 1.1. *Given a comment, how do you analyze the names and emotions mentioned in this comment?*

The existing sentiment analysis of movie reviews rarely seldom further analyzes the sentiment of the commentators. In social media, audiences may show different affective tendencies toward different protagonists of the movie, where traditional short text sentiment analysis can not capture these sentiments. To solve this problem, SRAS provides an object-oriented commentary sentiment analysis that performs named entity recognition and sentiment classification for each comment to more accurately capture the viewer's preference for a particular director (actor) of a particular movie.

DIFFICULTY 1.2. *How to quantify the director's contribution to the box office?*

Director (starring) contribution to the box office refers to the audience for the expectations of one or a few people to generate consumer spending movie tickets, analyze the contribution of a box office contribution to help get the actual business director or director value. In the analysis of the box office data, it is difficult to quantify the actual box office revenue of every star / director in a movie. There is also a lack of analysis of related modules in the existing commercial movie analysis system. SRAS provides a dynamic analysis module that accounts for the share of the box office's contribution, not only to quantify the box office's contribution but also to analyze the changes in the commercial value of a master before and after the release of the movie through changes in contributions over time.

DIFFICULTY 1.3. *Joint value evaluation and replacability among actors.*

Since effect of an actor suffer multi uncertainties and how to make a joint-evaluation of actor impact is a big question. A lot of method focus on giving linear weight for each factor. But this weight might change and impractical when goal of analysis changing. In SRAS , we mining joint value based on relationship among actors using heterogeneous network. Qualifying similarity and measuring increase of influence by co-star of two specific actor for target movie show great power on practicability of the system which enhance availability and effectiveness of SRAS a lot.

DIFFICULTY 1.4. *In which way the unstructured data cooperate with structured data?*

It is a problem how much effect by just using reviews(the unstructured data) or combine many other factor in analysis. We see that integrate sentiment help to prediction and investment, considering textual data an intensifying factor for attribute of movie itself. Whether the effect existing in each other if we analyse the synergy effects is unknown. To solve this problem, constrained network and iterative algorithm have been proposed to capture effect between the two kinds of data.

1.2 Roadmap

The rest of this paper is organized as follows, Section 2 presents an overview of SRAS . In section 3 we further study name entities recognition based on disambiguation and then describe how SRAS extract target-dependent sentiments and analyse changeable trend on sentiment. Section 4 presents dynamic impact of leader creators based on both sentiments and self attributes. In section 4.2 present the core value of system by phased prediction of movies which helps to maximum box-office based on reasonable investments now and future. Section 6 presents system evaluation as well as the case study. In section 7, we introduce the related work. Finally, Section 8 concludes the paper.

2 PLATFORM OVERVIEW

SRAS is a web-based integrated system for movie analysis. The screenshots of SRAS are shown in Figure 1 and the architecture overview is displayed in Figure 2. Using SRAS , users can carry out a series of analysis based on the movie centric information.

In general, SRAS consists of three layers of components, including Basic Processing Task Layer, Higher Algorithm Library Layer and Visualization Layer.

Basic Processing Task Layers provides data cleaning, data standardization and data conversion. Due to the different data sources, the data obtained can not be applied directly to the analysis. In fact, the movie data need to consider the following questions, name ambiguity elimination, used to evaluate the quality of the film at the box office, the film's reviews of object identification, actor knowledge base construction. These methods are provided in the basic processing task layer.

Higer Algorithm Library Layer contains the core technology of SRAS , including Target-Dependent Sentiment Analysis Module and Dynamic Impact of Actors and Directors Module. The layer considers the importance of film reviews to a film and combine with Machine Learning.

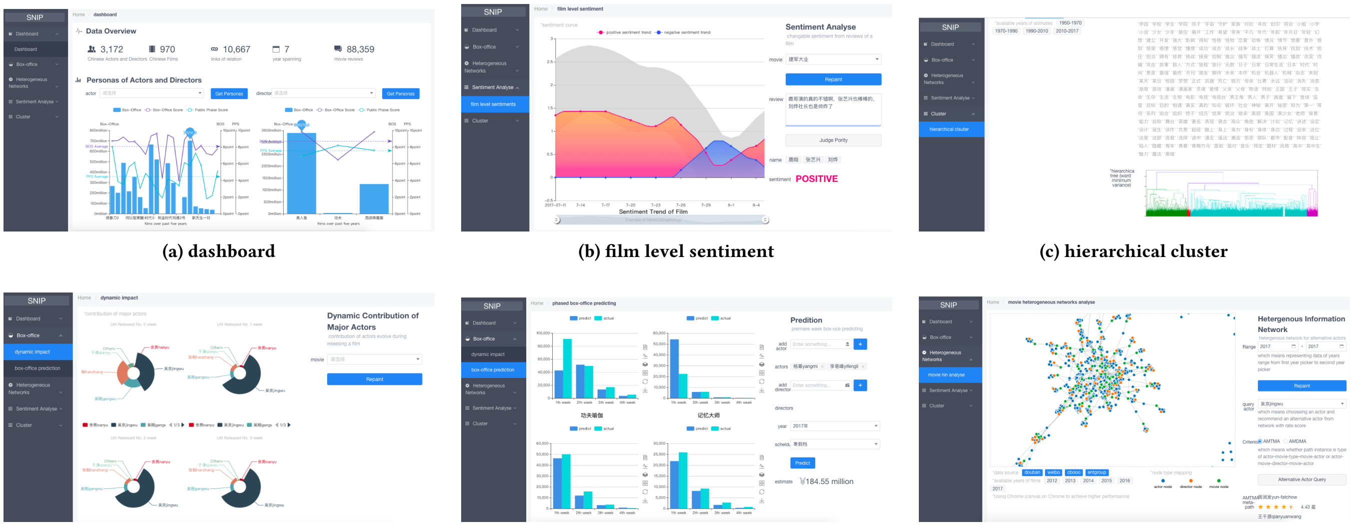
Visualization Layer is the interface between the system and the users. To intuitively present the analysis results, this layer is designed to be user-friendly and easy to operate. Specifically, Sentiment Tendency is to capture the changes of sentiment among audience during movie released. Box-office prediction is to predict the box office of a movie by actors, directors and audience factors. Alternative actors is to get the replaceable list of actors.

3 TARGET-DEPENDENT SENTIMENT ANALYSIS

3.1 Named Entity Disambiguation

Named entity disambiguation is one of the most important problems in natural language processing. In named entity category, person name has strong ambiguity so that person name disambiguation is the most difficult category. However, because the project inquiry is a well-known director and famous actor, and Chinese name and English name, makes the probability of Fijian phenomenon is greatly reduced, in this collection of 2371 directors and 801 actors, one by one, through the Baidu encyclopedia content, movie actor information, can be a good solution the same problem.

The main purpose of sentiment analyse is to discover views and opinions diversity [15] of different targets in films which reflect what attract audience watching the movie. The sample dataset in this area are review corpus $R = \{r_1, r_2, \ldots, r_n\}$ and each r_i has film name f_i it belonging to, time point $time_i$, scores towards this film s_i and user u_i who scores. These attributes of reviews are discrete (has domain specific value) and many sentences in it contribute to the whole sentiment of this single review. Since reviews are spread in social media platform such as twitter or micro-blogs, the sentences are limited and review of movie are usually enough short. Thus, many phenomenon occurred on short texts or natural language add difficulties when analysing the sentiments of targets (e.g. Rhetoric,



(a) dashboard (b) film level sentiment (c) hierarchical cluster

(d) dynamic impact (e) phased box-office prediction (f) heterogeneous network

Figure 1: Screenshots of SRAS

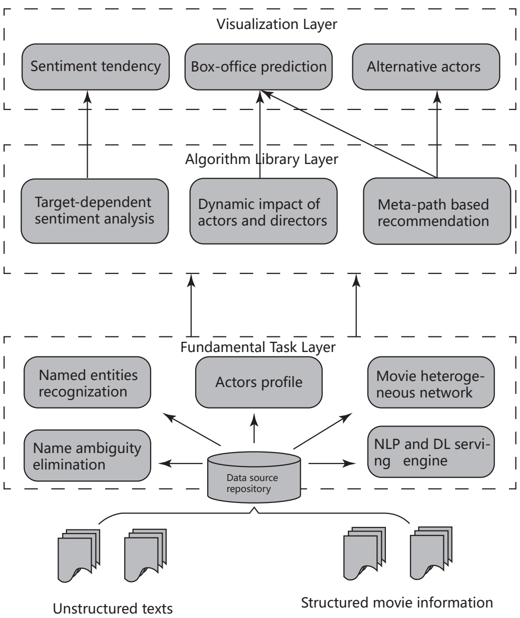


Figure 2: System architecture

metaphor, proverb and nicknames)[18]. Unlike many works focus on proceed in sentence-level document-level and time-period level step by step[4], we try to figure out the sentiment to a specific actors or director(leading creator of this film)[8]. A useful prior knowledge in reviews together with reviewers' rating score or stars made by user is evaluation score of specific film. We make the best of this knowledge to enlarge our sentiment words database because user usually behave consistent sentiment to same target.

3.2 Named Entities Recognition

If we want to know sentiments behind human expression, targets that people are taking about should be identified firstly. Here we concentrate on comment target about leader creators in films. The main challenging is the nicknames for actors and directors. Especially when one comment a film just referring his used acted film actress, it is necessary to seek other solution due to weakness of traditional solutions.

Traditional solutions weak in training data and much more training time while we want to adjust algorithm frequently. Give a brief introduction about sequence to sequence approach. Given a word sequence $X = \{x_1, x_2, \ldots, x_n\}$ we observed and the 4-tag label set $Tags = \{O, B, E, M\}$, the objective task is to find correct corresponding labels $Y = \{y_1, y_2, \ldots, y_n | y_i \in Tags\}$. By optimizing the loss function with parameter θ

$$J(\theta) = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} loss(f(x_i; \theta), y_i) \tag{1}$$

recurrent neutral networks such as *LSTM*, *BidirectionalLSTM*, *BidirectionalLSTM+CRF* [4] have been proved to be state-of-art structures. They consider both long time and short terms information and find tags with maximum possibility for sequence labeling. However, previous machine learning approaches need huge domain training data, while it is hardly to label these social media corpus accurately which are full of informality and flexibility. Especially actors get constantly evolving new roles in various films as time goes by, we hope find a simple and efficient way that recognize most of the comment targets in films reviews. There is a more fast and adaptive approach we adapted in this paper. In order to mining entities faster, we apply key word matching for leader creators in films. Online knowledge database which store film meta information can be strong support providing comprehensive and dependable analysis. Focused crawling technology based on web linkage contribute to construction of prior knowledge of **actors**, **roles** and **directors**.

Take a short example, Benedict Timothy Carlton Cumberbatch has nick name 'Curly Fu' and 'Peanut' which arisen by Sherlock he used to star in. We capture the Jaccard distance between 'Curly Fu' and 'Cur Fu' in word-level to qualify similarity for misspelling while Levenshtein distance (word edit distance) between 'Curly Fu' and 'Curly Ben' which represent word transform tolerance when doing word matching. Give thresholds control how similar and how different we can accept.

Algorithm 1 Framework of nickname mining for our system

Input:

The name set of **actors** and **directors** in each film
The name set of **roles** for each actors used to played
Threshold $\boldsymbol{\alpha}$ for minim similarity of misspell and variation
Threshold $\boldsymbol{\beta}$ for max tolerance of misspell and variation

Output:

Potential comment targets mapping dictionary A, P

1: Aggregating reviews group by corresponding film name along with preparing actors list A_i, directors list P_i in ith film according given **actors** and **directors**;
2: Constructing actors mapping A for every actor in a_i, so dose P for every director in p_i;
3: Quality similarity between every noun morphemes w and t in **roles**, **actors** and **directors** by *Jaccard* char distance in all comments for one film;
4: Recognizing potential nickname by the similarity and threshold $\boldsymbol{\alpha}$ we set, if $Jaccard(w,t) > \alpha$ add it to mapping dictionary A or P;
5: Checking potential nickname in mapping dictionary, if $Levenshtein(w,t) > \beta$, delete it from mapping dictionary A or P;
6: **return** A, P

Assume average noun morphemes length is m for all n reviews and t targets for all F films. Algorithm 1 need compare $c \times nmt$ times. Since word comparing consume limited space and time and $m \ll n, t \ll n$, we extract potential comment targets in $O(n)$ time complexity which greatly less than deep learning approach. We can easily adjust designed parameter and improve efficiency with parallel framework when processing TB data.

In order to decrease the storage space of data, we replace film name and user name by mapping them using film id and user id. $f_i \in F = \{f_1, f_2, \ldots, f_m\}$ and $u_i \in U = \{u_1, u_2, \ldots, u_U\}$. Let score for each review be $s_i \in [0, 10]$ and $time_i$ be the timestamps. The features we finally retrieved can be summarized in table **??**.

3.3 Binary Sentiment Classification

Sentences tend to be more complex when targets we want to analyse increase and sentiments detection being a hard work when analysing more objects. We take leader creator of film as major targets in this paper. Examples containing multi targets are shown as follows.

(1) Live here now! The point of life is looking for the point. —*A dog's purpose*
(2) It has been a year since I were familiar with Alexander Sandro Gonzalez Inarritu whose films have a cruel irony and full bitterness.—*Bird man*
(3) Mia had dreamed of becoming an actress known by more audience, while Sebastian want to won a place in his loving jazz. —*LaLa land*

Sentence(1) is a non-target review, while Sentence(2) shows one-target sample and Sentence(3) stands for multi-target sentences. Considering targets after NER process in section 4.1 consist of actors map list $A = \{a_1, a_2, \ldots, a_A\}$ and directors map list $P = \{p_1, p_2, \ldots, p_P\}$, we cluster sentences into comments set with two polarity(positive +1 and negative −1). Binary sentiment classifier play an important role in sentiment classification. Sentiment polarity $SP = \{+1, -1\}$ labeled by classifier for each sentences in corpus st_i^j of ith film with corresponding target j, $j \in A \cup P$ and targets capacity $T = ||A|| + ||P||$. Let T_i be number of targets in film i. We try to find out following segments.

$$Neg = \bigcup_{1}^{F} ng_i \mid polarity\ of\ ng_i = -1,\ ng_i \in \bigcup_{j=1}^{T_i} st_i^j \tag{2}$$

$$Pos = \bigcup_{1}^{F} ps_i \mid polarity\ of\ ps_i = +1,\ ps_i \in \bigcup_{j=1}^{T_i} st_i^j \tag{3}$$

Since we have to retrieve *Neg* and *Pos* shown in Equation (2) and (3), we take training data R_p with scores larger than 9 for positive sentiment and R_n which scores lower than 3 for negative one. We exclude reviews R_{ambi} which scores range between 4 and 8 that might be ambiguous over sentiment in training phrase. Instead, R_p and R_n show more concentrated sentiments in words. We train a classifier on R_p and R_n, then we split R with context window according referred target in each review. Reviews segments cover target's contextual information which construct sentiment corpus Neg and *Pos* towards specific targets are objectives for sentiment classifier. We mainly apply it on targets segments split from R_{ambi}. Finally, we get comprehensive sentiment polarity on targets segments in R.

Lexicon based sentiment analysis is constrained by sentence structure, latent word meaning and confused word features. We introduce bidirectional long short term memory (*Bi-LSTM*) neural network [4, 29] for end-to-end sentiment classification and automated feature learning as figure 3 shows.

Specially, word sequence are represented as a low dimensional , continuous and valued vector. It is called word embedding and we use pre-trained vectors so as to make better use of semantic and grammatical associations. Assume each d-dimension vector $W_i = \mathbb{R}^{d \times |V|}$ for ith word in whole vocabulary V. Targets words with L_r length cover $tw = \{W_r, W_{r+1}, \ldots, W_{r+L_r-1}\}$ and context words $cw = \{W_{r-L_w}, \ldots, W_{r-1}, W_{r+L_r}, \ldots, W_{r+L_r+L_w-1}\}$ with window length L_w surround corresponding targets words tw. *BiLSTM* maps word vectors to fix-length sentence vector by recursively transformation above vectors of previous time step h_{t-1}. Cells for *BiLSTM* in Figure 3 contains neural gates: input gates, forget gates, and output gates which adaptively remember input vector, forget history, and generate output vector. We add softmax layer to output sentiment of sentence vector which cover hidden

Review No.	Contents	Referred	Film No.	time point	Scores
1	Yang Yang's hard temperament is enough to hold up the role	Yang Yang	Ten great II of peach blossom	2016-07-25T13:36:12.000+0800	10
2	Yang Yang's acting is really embarrassing, Crystal Liu is OK	Yang Yang, Crystal Liu	Ten great II of peach blossom	2017-08-07T20:18:55.000+0800	5
3	The story of embarrassment, feeling Yehua did not love Baiqian	Yang Yang, Crystal Liu	Ten great II of peach blossom	2017-08-07T21:09:07.000+0800	2
...	...	...	...	...	...

Table 1: review targets after named entities recognition

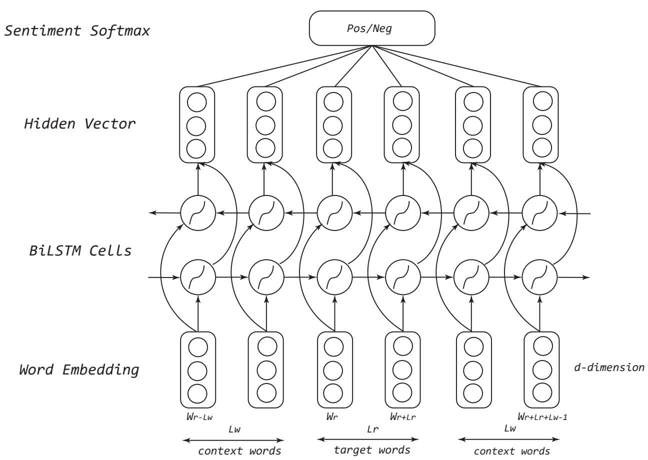


Figure 3: Structure of Bi-LSTM Sentiment Classifier, where L_r for target words length and L_w for context length we extract from original sentence. First layer in Bi-LSTM stands for forward hidden layer and the second layer are the backward one.

representation of vectors for specific target segment as Figure 1 shows.

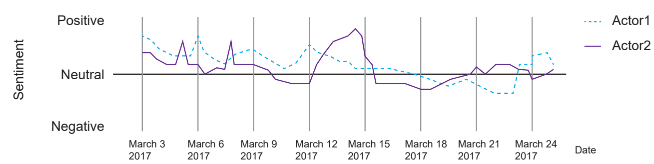
3.4 Actor-Level and Film-level Sentiment Trend

It counts that cinema arrange movie schedule when movie on scene. We hold sentiment tendency to help understand and measure movie quality in some ways. Predicting future movie box-office gain which based on actor-Level and film-level sentiment trend is described as follows. In this section, we mainly care about how to predict latent sentiment transform. Obviously, we quantify explicit sentiment towards special actors or directors. Although all information came from overall corpus, we can not deny that different sentiment exist in different time period during movie life. Audience's sentiments have trends and always behave dynamically. Previous work concentrate on sentiment polarity and we extend it by capture dynamic time-period sentiment. Denote a time period range from $time_r = [time_i, time_j]$, we capture review statistical sum R_{f_i} of each movie f_i at time period $time_r$ from Pos and Neg. Then we get different target j's positive or negative sentiment transformation at different $time_r$.

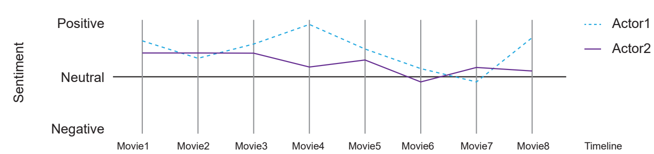
$$SentimentPos_i^j(time_r) = 10 \cdot \frac{|pos_i^j \ in \ time_r|}{R_{f_i}} \tag{4}$$

$$SentimentNeg_i^j(time_r) = 10 \cdot \frac{|neg_i^j \ in \ time_r|}{R_{f_i}} \tag{5}$$

Actor-level sentiment are shown figure 4 below. a) shows the change of sentiment in one movie life period which indicates what attract people. b) shows the average sentiment of a specific actor which indicates highest moments and lowest moment of an actor career



(a) Period Sentiment Trend: Change of sentiment in a specific film



(b) Dynamic Sentiment on Leader Creator : General sentiment transform of an actor in past five years

Figure 4: Film-level sentiment trend analyse of different time period. a) shows the change of sentiment in one film life period which indicates what attract people. b) shows the average sentiment of a specific actor which indicates highest moments and lowest moment of an actor career

Dynamic sentiment on leader creator shows how people transform their attention from different targets in movies. We pick out most interesting part (actor) from them and label it the main factor which mainly contribute to the box-office of specific film. The trend analyse is shown in Figure 4.

To measure the smoothness of the sentiment time series, we use slice window to capture the changes in the emotional inclination of the reviews. We Denote upper bound and lower bound of a day r,$UBP_i^j(r)$, $LBP_i^j(r)$,$UBL_i^j(r)$ and $LBN_i^j(r)$.

$$UBP_i^j(r) = mean_i^j(time_r) + \lambda \cdot std_i^j(time_r) \tag{6}$$

$$LBP_i^j(r) = mean_i^j(time_r) - \lambda \cdot std_i^j(time_r) \tag{7}$$

where

$$mean_i^j(r) = \frac{1}{2d} \sum_{k=r-d}^{r+d} SentimentPos_i^j(time_k) \tag{8}$$

$$std_i^j(r) = \frac{1}{2d} \sum_{k=r-d}^{r+d} (SentimentPos_i^j(time_k) - mean_i^j(time_r))^2 \tag{9}$$

Because the time series follows Gaussian Distribution, and we set $\lambda = 2$, means that confidence intervals is 0.0456,

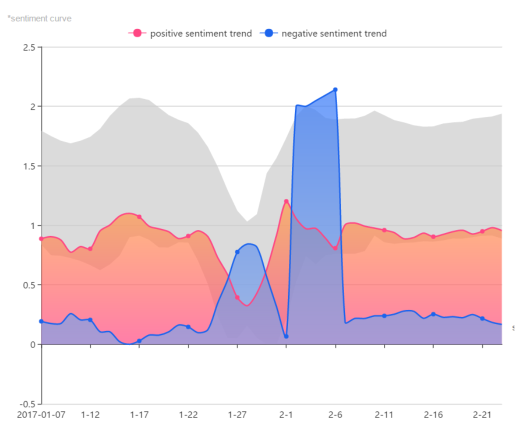


Figure 5: Film trend of Kung Fu Yoga

4 BOX-OFFICE INFLUENCE ESTIMATION AND PREDICTION

An obvious phenomenon is that the box office of movies has significant difference with the year and type. Similarly, due to the difference of movie types, the original movie ratings can not accurately describe the word of mouth of the movie because of different types. In order to eliminate these effects, we use $BoxScore$ to measure box-office quality of a movie. If there is a movie called m which box office is $boxoffice(m)$ with movie type set $T(m) = T_1, T_2, \ldots, T_k$ and release year $y(m)$, the $BoxScore$ of m is

$$BoxScore(m) = \frac{1}{|T|} \sum_{t \in T} \frac{BoxOffice(m)}{\max_{t \in T(n), y(n)=y(m)} BoxOffice(n)} \tag{10}$$

and the word of mouth score $WOMScore$ of m is

$$WOMScore(m) = \frac{1}{|T|} \sum_{t \in T} \frac{WOM(m) - \min_{t \in T(n)}(WOM(n))}{\max_{t \in T(n)}(WOM(n)) - \min_{t \in T(n)}(WOM(n))} \tag{11}$$

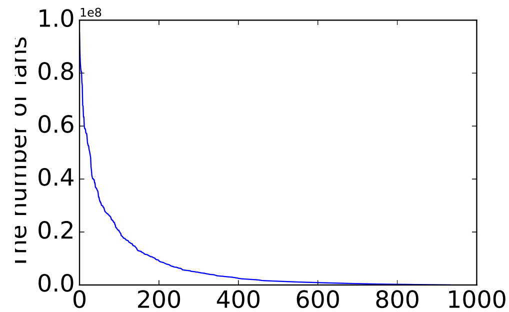


Figure 6: Distribution of Fans

We provide profile analysis of famous actors and directors which is shown in Figure 1. The box office value of a famous director or actor is one of the most important parts of the box office. Famous Cast (directors) are the main performers(directors) over 10 movies or won the film excellent actor(director) award or have more than ten million followers (shown in Figure6) in social networks. Portraying the box office of famous actors or directors in historical movies can help us analyze their value and Volatility.

We use bos, bov, wms, wmv to reflect the average box office performance, the box office performance fluctuation, the average word-of-mouth performance and the volatility of word-of-mouth performance so that we can depict the performance of the historical movie to provide guidance for the selection of the corner.

The goal of this module is to find who is the most bankable person in a movie and to capture the dynamic trends about actors and directors during the period of a movie released. As figure 7 shows, the box office influences can be made up of three parts: the actors'/directs' own box office appeal, media exposure and feedback from audience. For simplify, we regard actors and directors as creators.

4.1 Contribution of Cast and Crew

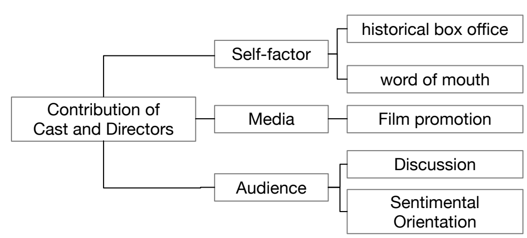


Figure 7: Composition of creative influence

The creator's own box office appeal. Actor's own box office appeal includes historical box office results and historical word of mouth performance, in which the film's historical movie box office performance can objectively reflect the creative ability of gold absorption, while the historical reputation score can objectively reflect the creative performance by Audience recognition, so this article at the same time using the history of the box office and the history of reputation to measure the creative box office call ability.

We define the box office quality boq to measure the box office generated by the creator itself based on the historical film data. As for directors, we define

$$boq(c) = wms(c) \cdot bos(c) \tag{12}$$

Considering different role importance, as for cast, we define

$$boq(c) = f(c) \cdot wms(c) \cdot bos(c) \tag{13}$$

where $f(c)$ is the influence factor and it depends on the position k that the actor c is. The larger k is , the smaller $f(c)$ gets.

Media exposure. The movie's promotional period is always accompanied by a wide range of media coverage, the media's title content reflects the current public focus, if an actor or director appear in the title a lot, you can explain the public familiarity and attention more. We define $meq_j(c)$ as the number of times the creator c referred to by the media in the jth week.

Feedback from the audience. Liu [19] mentioned the impact of word-of-mouth on the box office is very large and Craig [5]added to the equation online buzz variables expressing awareness and purchase intention and examined factors that contributed to higher levels of e-WOM. With the rapid development of the Internet, people can express their emotions on social media at any time. Thanks to the development of social media networks, the promotion of the movie has also gradually increased the proportion of online publicity. At the same time, the content of the discussion of the movie has also formed a considerable amount of data. The movie review often includes the concept of " If you can get the audience's idea of fittfittan actor or master from the commentary, you can see whether the source of the audience's emotional inclination toward the movie comes from the influence of the movie's main actor (including the protagonist).We define $aco_j(c)$ as the number of times the creator c referred to by the active comments in the jth week.

In the short life of the movie, the contribution of the genre to the box office is obviously not constant. With the development of the Internet media, real-time feedback of the movie users on the movie will affect the watching desire of the non-movie users, and the positive contribution Refers to the potential box office or the desire to watch the increase; the negative is the box office to bring the negative image, to dispel the wishes of watching. Therefore, in the film's life, the star effect on the box office's contribution is dynamic, the project will be divided into the life of the movie before the release, the first week of release, the second week of release, the third week of release, released the fourth Friday A period, respectively, to explore the five periods of major box office contributions.

We define $DIP_j(c)$ to measure the influence of a famous cast c in jth week

$$DIP_j(c) = w_1 \frac{boq(c)}{\sum_c boq(c)} + w_2 \frac{\sum_{i=0}^{j} mep_j(c)}{\sum_{i=0}^{j} \sum_c mep_j(c)} + w_3 \frac{\sum_{i=0}^{j} aco_i(c)}{\sum_{i=0}^{j} \sum_c aco_i(c)} \tag{14}$$

Then we can get the DIR (Dynamic Impact Ratio) as $DIR_j(c) = \frac{DIP_j(c)}{\sum_{k \in C} DIP_j(k)}$ to measure the ratio of the box office contribute from creator c to all box box office contribute in jth week. Where w_1,w_2,w_3 is three coefficients to embody the importance of creator's own box-office appleal, media exposure and feedback from the audience. The choice of proportion involves expert experience and in practice, we set $w_1 = 0.1, w_2 = 0.2, w_3 = 0.7$.

4.2 Box-office Predicting

Movie box office is influenced by many factors, such as investment budget, script, director, actor, post production, producer and producer, media publicity and movie reputation. On the whole, we quantify the factors that affect movie box office from the film itself, the director, the actor and the audience. The opening week due to the lack of effective comments, so just for the release of the film at the box office during the period of life prediction problem, we consider the factors associated with the film at the box office and did not join the historic influence of sentiment analysis, in the opening week after we join in the model of the value of emotional factors.

4.3 Premiere week box-office predicting

Symbol	Description
sc	The mean bos of famous creators
gsc	The mean wms of famous creators
std	The variance bos of famous creators
$gstd$	The variance wms of famous creators
$starcounts$	The number of famous creators
$schedule$	The schedule that the movie came out
$year$	The year that the movie came out

Table 2: Features

Table 2 shows the features we used in the task of predicting premiere week box office. As is metioned above, the box office of a movie depends on not only movie factors itself including actors, directors, type, movie propaganda and so on but also influenced by audience reations. In addition, the time the movie released also has a certain impact on the movie box office. In general, predicting premiere week box-office can capture the influences of famous creators(actors and directors) to the movie. We provide Premiere week box-office predicting by using traditional box-office predicting model Support Vector Regression. This Module can help people to asses the influence of stars in the movie.

4.4 Phased Box-office Predicting

In the past prediction of box office, we usually only forecast the overall box office, but often ignored the movie's influence on movie box office due to the audience's attitude. Such as "wolf 2", the movie box office to be among the world's top 100, because in the release process, because the audience warmly, attracted the audience is not the original film (film series the fans, fans, like the kind of audience) to watch a movie, which leads to the box office continued to rise however, the traditional model is unable to capture this phenomenon. Therefore, as shown in Figure 8, from the pre launch to the 1 months after the release, we set up the box office prediction model with the weekly variation of the weekly release to predict

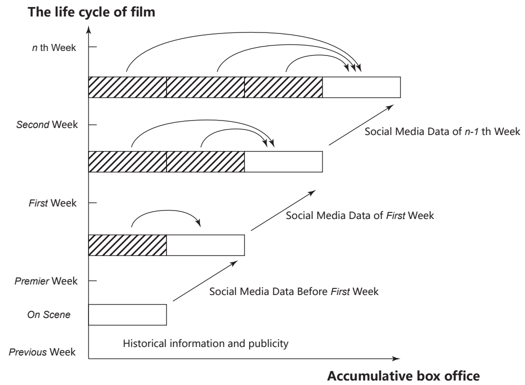


Figure 8: Multi-stage box office prediction model

the box office for the first week, the box office for second weeks, the third week box office and the box office at the fourth week.

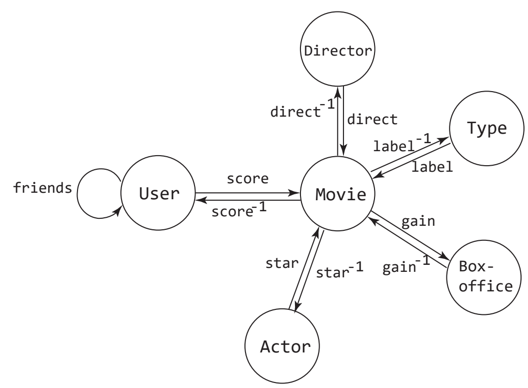
In this section, we add two type features: $weibo_i$, The ratio of creative comments and all comments until the ith week and $week_i$ The box office in the ith week. With these additional messages obtained from audience and current box office, the results we predict are more similiar to real box office, shown in Table 3.

Algorithm[31]	Metric	First	Second	Third	Forth
SVR	MSE	**0.0939**	0.2716	0.3949	0.7445
	R2score	**0.6349**	0.0693	0.6605	0.5734
	EVC	**0.6082**	0.1977	0.7136	0.6332
RF	MSE	0.2683	0.3192	1.0082	0.6839
	R2score	-1.8019	-1.0037	0.0296	0.1461
	EVC	0fiff4194	-0.1191	0.0253	0.4586
GBDT	MSE	0.6354	0.5121	0.7647	2.0396
	R2score	-2.1466	-0.7831	0.2974	0.1664
	EVC	0.2024	-0.0841	0.6118	0.0505
LR	MSE	1.0951	0.5105	0.7335	1.3732
	R2score	-3.5674	-0.7488	0.3694	0.2142
	EVC	-3.0608	-0.6947	0.3728	0.3254
LASSO	MSE	0.7427	**0.0528**	**0.0896**	**0.5025**
	R2score	-2.0977	**0.8189**	**0.9229**	**0.7124**
	EVC	-0.9281	**0.9164**	**0.9758**	**0.7183**

Table 3: Evaluation of Predicting Model

5 ACTORS RECOMMENDING

In order to discover important relation and promotion between two actors, we focus on mining movie heterogeneous information network[27]. Discovering interesting topology of network provide great insight about the system innate character. A movie heterogeneous information network (*MHIN*) is a domain information network with multiple types of entities and relations [28]. We construct our *MHIN* with multiple objects such as movies (M), directors (D), users (U), Actors (A), Types (T) and box-offices (B). Figure 9 shows typical *MHIN* schema defined by us. Links exist between users and movies denoting score and score-by relations, between movies and actors(directors) denoting star(direct) and star-by(direct-by) relations, between movies and types denoting label and (labeled) relations, between movies and box-offices denoting gain and gain by relations. Also, we can add other attributes into the movie heterogeneous network (e.g years (Y), producers (P)). Meta-path is a connect relation between two types of objects in *MHIN*. Denote network schema for our *MHIN* is $\mathcal{S} = (\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ represent object types and $\mathcal{R}$ indicate different types of relationships. A meta-path $\mathcal{P}$ is defined in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_k} A_{k+1}$. Length of $\mathcal{P}$ is defined by number of relations it contains. Semantic composite relations are implied in meta-paths[**?**]. For example, we can evaluate the similarity of movies or actors by length-2 meta-path MDM (movie-director-movie), MAM (movie-actor-movie) or AMA (actor-movie-actor), AMTMA(actor-movie-type-movie-actor).



(a) Structure of movie heterogeneous network

Figure 9: Social network schema of MHIN.

5.1 Rank Based Cluster For Actors

In this section, we introduce a way [**?**] to make distinguish between A-list and B-list actors. We prefer applying cluster based classifier on actors popularity recognition. Strong available mutually reinforcing relations between cluster and rank are helpful to deal with different levels of actors. We defined recursion formula by following empirical rules.

RULE 5.1. *high rank actors star more high rank movies*

RULE 5.2. *high rank movies attract more high rank actors*

RULE 5.3. *actors get high rank with high rank co-starers*

Suppose we want K clusters of actors. According the star relationship between actors and movies, we define matrix $M_{MA}(i,j) = c_{ij}$ representing actor j's contribution of movie i (see Section 3.1) and similarly, $M_{AA}(i,j) = m_{ij}$ stands for number of movies that actor i and actor j co-stared. Note that $W_{AM} = W_{MA}^T$, and i =

$\{1, 2, \ldots, m\}, j = \{1, 2, \ldots, n\}$. According the 3 rules we proposed, we have

$$r_A(i) = \alpha \sum_{i=1}^{m} W_{AM}(j,i) r_M(i) + (1-\alpha) \sum_{j=1}^{n} W_{AA}(i,j) r_A(j) \tag{15a}$$

$$r_A(j) \leftarrow \frac{r_A(j)}{\sum_{j'=1}^{n} r_A(j')} \tag{15b}$$

$$r_M(i) = \sum_{j=1}^{n} W_{MA}(i,j) r_A(j) \tag{15c}$$

$$r_M(i) \leftarrow \frac{r_M(i)}{\sum_{i'=1}^{m} r_M(i')} \tag{15d}$$

Note that $\alpha \in [0,1]$ is a believe factor of weighed component of rule 3, $r_A(j)$ and $r_M(i)$ are normalized rank score vector. We finally get r_A which should be primary eigenvector of $\alpha W_{AM} W_{MA} + (1-\alpha) W_{AA}$. Further, we capture posterior probability $\pi_{i,k}$ that a_i from cluster k, Once an actor acts a movie, he is more likely to star high ranked film and for movie, its success are more likely contributed by high rank actor. Thus we have K dimensional vector $s_{a_i} = \{\pi_{i,1}, \pi_{i,2}, \ldots, \pi_{i,K}\}$ where $\pi_{i,k}$ denotes a_i's coefficient for component k.

The cluster center and the distance between each actor and each cluster can be defined as

$$S_{A_k} = \frac{\sum_{a \in A} s(a)}{|A_k|} \tag{16a}$$

$$Distance(a, A_k) = 1 - \cos(s_{a_i}, S_{A_k}) \tag{16b}$$

Algorithm 2 Distinguish Popularity of Actors

Input:

Our movie information network $MHIN = (M, A; W)$
Cluster Number K

Output:

K clusters of actors A_i and rank of actor in each cluster

1: iter = 0;
2: Init partitions for A, get $PA^{iter} = \{A_i^{iter}\}_1^K$;
3: Repeat following until $PA^{iter} - PA^{iter-1} < \epsilon$ or iterations reach limitation
4: For all cluster calculate r_A followed by rank function
5: Evaluate Θ for mixture model and get component efficient estimations s_{a_i} for each actor a_i
6: Update centers $S_{A_k}^{iter}$ of each cluster A_k
7: Reassign each actor a_i according distance between a_i and each cluster center A_k^{iter}

5.2 Constrained Alternative Actors

In this section, we discuss method choosing alternative actors for casting agents. The basic idea is measure similarity between actors. Since movie information network has been built, we then introduce meta path-based similarity framework [28] for alternative actors. The example is show in table 4.

Denote actors a_1 and a_2 then $p_{a_1 \rightsquigarrow a_2}$ is a path instance between a_1 and a_2, $p_{a_1 \rightsquigarrow a_1}$ is a path instance between a_1 and a_1, $p_{a_2 \rightsquigarrow a_2}$ is a path instance between a_2 and a_2. We then get similarity between a_1 and a_2 by calculate

Type	Path instance	Meta-path
I(AMA)	Andy-M_1-Sara	Actor-Movie-Actor
II($AMTMA$)	John-M_2-Comedy-M_3-Sara Ben-M_4-War-M_3-Sara Drew-M_5-Dracula-M_2-John	Actor-Movie-Type-Movie-Actor
III($AMDMA$)	Diana-M_6-Spielberg-M_1-Andy John-M_2-Luc Besson-M_5-Drew Sara-M_3-Tom Tykwer-M_5-Drew	Actor-Movie-Director-Movie-Actor

Table 4: Different Types of meta-path in MHIN

$$s(a_1, a_2) = \frac{2 \times |\{p_{a_1 \rightsquigarrow a_2} : p_{a_1 \rightsquigarrow a_2} \in \mathcal{P}\}|}{|\{p_{a_1 \rightsquigarrow a_1} : p_{a_1 \rightsquigarrow a_1} \in \mathcal{P}\}| + |\{p_{a_2 \rightsquigarrow a_2} : p_{a_2 \rightsquigarrow a_2} \in \mathcal{P}\}|} \tag{17}$$

Pay attention that A-M-A is not a meta path. It is insignificant find co-star actor as the recommend actor. For example, Andy and Sara co-star M_1 and is recommend who is similar to Sara in the case of choose alternative actor for Andy. We use A-M-T-M-A and A-M-D-M-A to find actor similarity among movie types and directors. Another constrain is sex that female actor should recommend female and male the same. Base on this measure of actors, we can not only know similarities but also get multi metrics that reflect how actors link with others. Query a specific actor $a_i$and then return a recommend actor list r_{a_i} that ordered by similarities. Investors can determine the scope of film plan or prepare backup candidate for chief actor. In SRAS we also analyse different types of meta path instances for providing comprehensive alternative information for other actors.

6 SYSTEM EVALUATION AND CASE STUDY

In the case study, we use two concrete tasks to demonstrate how investors can apply SRAS to the analysis of movie box office.

Task 1: Movie Inclusion. It's important for investors to analyze the performance of a movie. Movie performance not only refers to box office income, word of mouth, but also includes the changes in the audience and the change of the influence of the director during the screening process.

Fig.4 shows dynamic impact of creators in "Wolf Warriors II". Before the movie released (week 0), the attentions people payed to Hans Zhang, Gang Wu and Jason Wu were equal while with the movie's release and the increase in the heat of the film, more and more people went to the movies for the sake of Jason Wu. Meanwhile, the dynamic influence process can reflect a major change in the importance of the movie. We can see that Jason Wu has greatly improved his value through the movie himself. To some extent, Jason Jing has more box office attractiveness, and investing in the actor's next movie is very likely to get high box office gains.

Task 2: Evaluation of investment quality. Investors usually need to assess the movie's expected box office and measure the

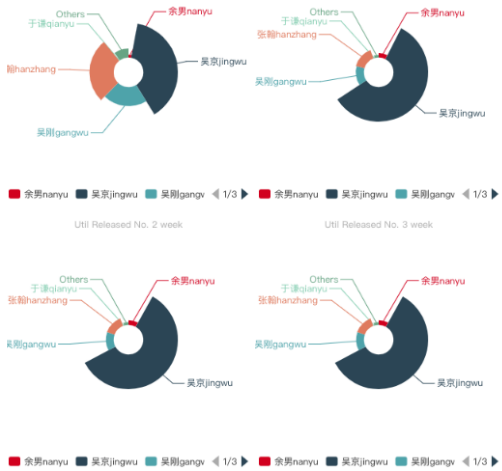


Figure 10: Dynamic Impact of Wolf Warriors II

actor’s return on investment. Because there is a strong correlation between movie box office and audience feedback, the box office of the first week is less affected by audience feedback, so the first week box office can be regarded as the baseline of box office. The first week box office prediction can measure the expected revenue of the movie, and the actor replacement module can help investors to screen the appropriate actors at the right price at the appropriate box office in the first week.

In SRAS , we give a interface to query expected box-office when combine many factors into an investment. Such as Chinese actor $YangMi$ and $YifengLi$ and year selector into 2017 winter, the expected outcome is 184.55 million. We confirm the result as table 5 shows. The data indicated that two star co-star relation really play an important role in box-office improvement compared with single star.

Actors	year	schedule	Outcome
YangMi + YifengLi	2017	Winter	184.55 m
YangMi + YifengLi	2016	May Day	143.83 m
YifengLi	2016	May Day	123.60 m

Table 5: Estimate first week box-office for future investment

Task 3: Quantify sentiment and choose suitable actor. The following part will show the film sentiment analysis. Taking a pretty hot film Kung Fu Yoga in 2017 into consideration, at the first ten days, positive sentiment control the major opinion. While as time goes on, negative sentiment burst much higher than normal positive sentiment. We can see form figure 5 that people begin to criticize Jack Chen about this film. Even though at last, the public sentiment turn steady, we still believe Kung Fu Yoga has a tough time.

At last, we want to give a rank list of alternative actors when investors choose major actor. We give rank order form both AMTMA and AMDMA meta-path based method. If we query $JingWu$ (Chinese major actor of Wolf Warriors II), we get highest score of 4.43 for Chow Yun Fat(Famous Chinese actor) based on his performers’ experience especially film types that has been played in. Table 6 give details.

Alternative Actor Name	rank score
yun-fatchow	4.43
qianyuanWang	4.38
qingYe	4.38
tomeroz	4.36
hakonFung	4.35

Table 6: Highest 5 recommend alternative actors for JingWu

7 RELATED WORK

Box-office forecasting is a challenging but import task for movie contributors in their decision making process. There are lots of works in box-office predicting [26, 32]. Empirical studies of the determinants of box office revenues have mostly focused on post-production factors - that is, ones known after the film has been completed and/or released with audience reactions. Early in [19] shows that word of mouth (WOM) can help explain box office revenue. The results of experiments from [16] show that the utilization of SNS data can improve he forecasting accuracies of machine learning-based algorithms and their combination. [11] regard the textual and content analysis of the screenplays of films as chief among the pre-production factors. [12] presents new box-office forecasting models to enhance the forecasting accuracy by utilizing review sentiments and employing non-linear machine learning algorithms.

The influence of the stars on the box office in the movie is another hot spot. [6] shows that the role of a star is not the influence of the early box office, but the extension of the film’s release period. The results of the study[3] found that stars had only an impact on the box office of the premiere, and the impact on the movie’s box office was even negative. However, Both [9, 22] find stars who participate in movies can affect movie revenue very positively. Furthermore, [14] proves that the internet talk about movie stars can bring positive effects to the box office.

Comparing with previous work, SRAS succeed in integrating unstructured reviews with structured information to enhance and extend application of system. We first extract and qualify sentiment which is core part of audience react information. Specially, we choose more state-of -art techniques (e.g LSTM[15], Heterogenous Network[27]). It is hard to overemphasize the value of NLP in system. Many study [12, 16, 29] focus on single point of film industry taking the IMDB data as the foundation. Instead, we make hard to integrate many aspects which is the distinctive features of the Chinese film market, as well as give interpretable visualized result for follow-up decision. In addition, dynamic analyse of tendency and impact also employ time series model and statistical analysis. To the best of knowledge, SRAS is a comprehensive solution to overcome tradition difficulty on film data analysis currently.

8 CONCLUSION

The return on investment (ROI) analysis about movies is a challenging but important task for movie investors in their decision making process. For investors, they are expected to get high box-office revenue with appropriate investments. Apart from movie special effects, most of investments are used to remuneration for movie actors(actress). Due to complicated factors such as audience reactions, screenplay, film types etc, it's not easy for investors to estimate upfront ROI, which makes film investment a gamble. Although most current research works aim to predict box-office revenue more specific, they can not provide investors more direct profits information about the success of the movie they would like to invest.

In this paper, we design and implement an integreated system, called Film Social Network Investment Platform (a.k.a SRAS), that aims to do some analysis of movies for investors. SRAS provides various modules for users to predict box-office, capture public opinion about a film, assess the value of the actor, monitor the changes of the ratio of box-office about actors, assess actor replacement and so on.

REFERENCES

[1] Sitaram Asur and Bernardo A Huberman. 2010. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 492–499.

[2] Suman Basuroy, Subimal Chatterjee, and S Abraham Ravid. 2003. How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of marketing* 67, 4 (2003), 103–117.

[3] Byeng-Hee Chang and Eyun-Jung Ki. 2005. Devising a practical model for predicting theatrical movie success: Focusing on the experience good property. *Journal of Media Economics* 18, 4 (2005), 247–269.

[4] Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2017. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems With Applications* 72 (2017), 221–230.

[5] C Samuel Craig, William H Greene, and Anthony Versaci. 2015. E-word of mouth: early predictor of audience engagement: how pre-release fiE-WOMfi drives box-office outcomes of movies. *Journal of Advertising Research* 55, 1 (2015), 62–72.

[6] Arthur De Vany and W David Walls. 1999. Uncertainty in the movie industry: Does star power reduce the terror of the box office? *Journal of cultural economics* 23, 4 (1999), 285–318.

[7] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J Smola, Jing Jiang, and Chong Wang. 2014. Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 193–202.

[8] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In *Meeting of the Association for Computational Linguistics*. 49–54.

[9] Anita Elberse. 2007. The power of stars: Do star actors drive the success of movies? *Journal of Marketing* 71, 4 (2007), 102–120.

[10] Gerda Gemser, Martine Van Oostrum, and Mark AAM Leenders. 2007. The impact of film reviews on the box office performance of art house versus mainstream motion pictures. *Journal of Cultural Economics* 31, 1 (2007), 43–63.

[11] III Hunter, Starling David, Susan Smith, and Saba Singh. 2016. Predicting box office from the screenplay: A text analytical approach. *Journal of Screenwriting* 7, 2 (2016), 135–154.

[12] Minhoe Hur, Pilsung Kang, and Sungzoon Cho. 2016. Box-office forecasting based on sentiments of movie reviews and Independent subspace method. *Information Sciences* 372 (2016), 608–624.

[13] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A Smith. 2010. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 293–296.

[14] Ekaterina V Karniouchina. 2011. Impact of star and movie buzz on motion picture distribution and box office revenue. *International Journal of Research in Marketing* 28, 1 (2011), 62–74.

[15] Arzoo Katiyar and Claire Cardie. 2016. Investigating LSTMs for Joint Extraction of Opinion Entities and Relations. In *Meeting of the Association for Computational Linguistics*. 919–929.

[16] Taegu Kim, Jungsik Hong, and Pilsung Kang. 2015. Box office forecasting using machine learning algorithms based on SNS data. *International Journal of Forecasting* 31, 2 (2015), 364–390.

[17] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* 50 (2014), 723–762.

[18] Jiwei Li and Eduard Hovy. 2014. Sentiment Analysis on the People's Daily. In *Conference on Empirical Methods in Natural Language Processing*. 467–476.

[19] Yong Liu. 2006. Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing* 70, 3 (2006), 74–89.

[20] Asha S Manek, P Deepa Shenoy, M Chandra Mohan, and KR Venugopal. 2017. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. *World wide web* 20, 2 (2017), 135–154.

[21] Pablo Marshall, Monika Dockendorff, and Soledad Ibáñez. 2013. A forecasting system for movie attendance. *Journal of Business Research* 66, 10 (2013), 1800–1806.

[22] Randy A Nelson and Robert Glotfelty. 2012. Movie stars and box office revenues: an empirical analysis. *Journal of Cultural Economics* 36, 2 (2012), 141–166.

[23] Reggie Panaligan and Andrea Chen. 2013. Quantifying movie magic with google search. *Google WhitepaperfiIndustry Perspectives+ User Insights* (2013).

[24] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.

[25] Ramesh Sharda and Dursun Delen. 2006. Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications* 30, 2 (2006), 243–254.

[26] Scott Sochay. 1994. Predicting the performance of motion pictures. *Journal of Media Economics* 7, 4 (1994), 1–20.

[27] Yizhou Sun, Jiawei Han, Charu C Aggarwal, and Nitesh V Chawla. 2012. When will it happen?: relationship prediction in heterogeneous information networks. (2012), 663–672.

[28] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of The VLDB Endowment* 4 (2011), 992–1003.

[29] Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Target-Dependent Sentiment Classification with Long Short Term Memory. *arXiv: Computation and Language* (2015).

[30] Duyu Tang, Bing Qin, Ting Liu, and Yuekui Yang. 2015. User Modeling with Neural Network for Review Rating Prediction.. In *IJCAI*. 1340–1346.

[31] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[32] Li Zhang, Jianhua Luo, and Suying Yang. 2009. Forecasting box office revenue of movies with BP neural network. *Expert Systems with Applications* 36, 3 (2009), 6580–6587.