Improving Movie Gross Prediction Through News Analysis

Wenbin Zhang and Steven Skiena
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400 USA
Email: {wbzhang, skiena}@cs.sunysb.edu

***Abstract*—Traditional movie gross predictions are based on numerical and categorical movie data. But since the 1990s, text sources such as news have been proven to carry extra and meaningful information beyond traditional quantitative finance data, and thus can be used as predictive indicators in finance. In this paper, we use the quantitative news data generated by *Lydia*, our system for large-scale news analysis, to help us to predict movie grosses. By analyzing two different models (regression and k-nearest neighbor models), we find models using only news data can achieve similar performance to those use numerical and categorical data from The Internet Movie Database (IMDB). Moreover, we can achieve better performance by using the combination of IMDB data and news data. Further, the improvement is statistically significant.**

I. Introduction

The movie industry is of intense interest to both economists and the public because of its high profits and entertainment nature. In 2007, the total revenue of U.S. movie market was $8.74 billion and it continues to grow. An interesting question is to forecast pre-release movie grosses, because investors in the movie market want to make wise decisions. The investors could be widely either earlier stage investors like movie studios or movie distributors, or later stage ones like movie retailers, exhibitors, home video makers, or even book or CD-ROM publishers.

Traditionally, people predict gross based on historical IMDB data analysis regarding specific characteristics, e.g., the movie's genre, MPAA rating, budget, director, number of first-week theaters, etc., but with somewhat limited success. Nevertheless, recent publications ([1], [2], [3], et al.), have shown the media's power on forecasting financial market like stock prices, volatilities, or earnings. Considering the encouraging results, it is reasonable to infer that news has predictive power for movie grosses as well. We are unaware of any previous attempt to apply linguistic analysis to movie gross prediction. Therefore, here we focus on improving movie gross prediction through news analysis.

Our primary goal is to prove that we can give better pre-release prediction of movie grosses if we use news data, because commercially successful movies, actors, or directors are always accompanied by media exposure. Our experiments use *Lydia* ([4], *http://www.textmap.com*), a high-speed text processing system, to analyze news publicity and output movie news data, and then to help our movie gross prediction.

In this paper, our particular contributions are: 1) We provide a comprehensive way to evaluate news data and linguistic sentiment indexes as well as give a detailed analysis for movie news data; 2) We build k-nearest neighbor models for movie gross predictions, which have not been studied in previous movie prediction literatures; 3) Through large scale analysis, we prove that news data is capable of helping people to build models with better performance. We do not use any post-release data in the following experiments, and all the predictions are out-of-sample predictions. In practice, our approach provides a feasible and more accurate estimation regarding the investment worthiness for some pre-release investors and almost all the post-release investors.

The contents of this paper are organized as follows. First, we will review related work briefly. Second, we will describe the movie data sources, both traditional movie data and news data, and give a correlation analysis. We then set up different models with traditional movie data, movie news data, and their combination respectively as well as evaluate their performance. Finally, we conclude that we can improve traditional movie gross prediction through news analysis.

II. Related Work

Different people work on movie gross prediction from different perspectives. Most previous work ([5], [6], [7], [8], [9], et al.), 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. For example, Simonoff and Sparrow [6] gave three predictions for movie *The Horse Whisperer*, which had an actual gross of $74.37 million. Its predicted grosses for the pre-release, first weekend and Oscar models are $1.405 million, $63.932 million and $59.391 million respectively. However, both the first weekend model and Oscar model are post-release models. Sawhney and Eliashberg [7] also claimed that their model works pretty well by taking the first three weeks of gross data as input, but admitted that it is much more difficult to give shape estimation for either model parameters or gross if we 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.

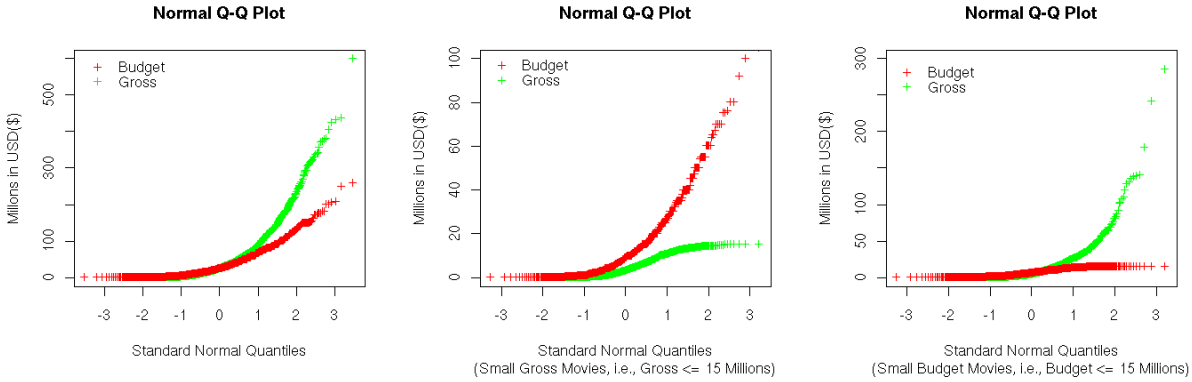
Moreover, there has been substantial interest in the NLP community on using movie reviews as a domain to test sentiment analysis methods, e.g., [10], [11], et al. Basically speaking, they apply information retrieval or machine learning techniques to classify movie reviews into some categories and hope 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". Pang and Lee [12] gives a detail review in this domain. However, to the best of our knowledge, news and sentiment analysis has not been previously studied as a predictor of movie grosses. In addition, Mishne and Glance [13] show that movie sales have some correlation with movie sentiment references, but they neither build prediction models or show the value of the correlation because they think the result is not good enough for accurate modeling.

III. Movie Data and Correlation Analysis

There are two kinds of movie data used in this paper, movie specific variables and movie news data. The movie specific variables are collected from traditional movie websites like IMDB, but the movie news data is obtained from *Lydia*. We need to analyze the correlation between movie grosses and traditional movie variables or news variables, and then let it guide us to set up reasonable models for movie gross prediction. The correlation between variables is measured

Movie Variables	Categories	Movies	Mean	Median	Min	Max	Corr	p-Value
Budget	All	1500	45.97	25.48	0.0008	600.79	0.672	**<0.001**
Opening Screens	All	1500	45.97	25.48	0.0008	600.79	0.647	**<0.001**
First Week Gross	All	1500	45.97	25.48	0.0008	600.79	0.841	**<0.001**
World Gross	All	1500	45.97	25.48	0.0008	600.79	0.936	**<0.001**
Release Date	Holiday	640	55.13	32.22	0.008	600.79	0.132	**<0.001**
	Non-holiday	860	39.15	21.44	0.0008	436.72	-0.132	**<0.001**
MPAA Rating	G	41	82.72	58.40	0.669	339.71	0.103	0.261
	PG	201	65.60	42.27	0.119	436.72	0.128	**0.035**
	PG-13	500	59.04	35.28	0.011	436.72	0.154	**<0.001**
	R	646	30.68	16.98	0.0008	216.33	-0.221	**<0.001**
	NC-17	17	18.18	7.4	0.030	70.10	-0.049	0.426
Source	Sequel	127	90.24	64.96	0.146	436.72	0.224	**0.006**
	Not Sequel	1373	41.88	22.73	0.0008	600.79	-0.224	**<0.001**
Origin Country	USA	1191	50.28	30.31	0.0008	600.79	0.141	**<0.001**
	Not USA	309	29.38	11.55	0.009	317.56	-0.141	**0.006**

Table I: Correlation Coefficient of Movie Variables versus Movie Grosses. The given value of Mean, Median, Min, and Max grosses are in terms of million dollars. The bold numbers show the corresponding correlations are statistically significant at a 0.05 significance level.



(a) All Movies (b) Low-grossing Movies (c) Low-budgeting Movies

Figure 1: The normal quantile-quantile plots used for the normality test of the movie budgets or grosses distributions. Figure (a), (b), and (c) show the results for all movies, low-grossing movies, low-budgeting movies in our test data set respectively, which indicate movie budgets or grosses are not following strictly Gaussian distribution.

Entities	Duration	Gross (Pre-rel)	Gross (Post-rel)	Budget (Pre-rel)	Budget (Post-rel)
Movie	1 week	0.707	0.781	0.497	0.480
	1 month	0.672	0.779	0.463	0.474
	4 months	0.629	0.749	0.437	0.455
Director	1 week	0.494	0.602	0.311	0.389
	1 month	0.371	0.495	0.218	0.389
	4 months	0.192	0.317	0.117	0.078
Top 3 Actors	1 week	0.640	0.726	0.476	0.528
	1 month	0.569	0.683	0.448	0.477
	4 months	0.493	0.618	0.413	0.424
Top 15 Actors	1 week	0.646	0.725	0.533	0.595
	1 month	0.575	0.686	0.477	0.530
	4 months	0.511	0.618	0.415	0.433

Table II: Correlation Coefficient of Logged Pre-release News Article Counts versus Logged Grosses under various scenarios. The rows indicate what kind of entities are examined in terms of what kind of duration, i.e., 1 week, 1 month, or 4 months. The columns indicate the correlation is for gross or budget in terms of pre-release(or post-release) article counts.

by both strength and significance. The strength is futher evaluated by the correlation coefficient r, or sometimes by the coefficient of determination r^2, while the significance can be verified by t-test $t = r\sqrt{\frac{N-2}{1-r^2}}$, in which r is the correlation coefficient, N is the sample size, and $N-2$ is the degree of freedom of the t-test. In this paper, we use significance level 0.05 to test the statistical significance because 0.05 is conventionally a standard threshold to evaluate significance although in our experiments most of the results are statistically significant to an even stricter level.

A. Traditional Movie Variables and Correlation Analysis

Traditional movie data is available at *http://www.imdb.com* and *http://www.the-numbers.com*. We wrote a spider program and downloaded data for all movies released from 1960 to 2008. Table I summarizes the relationship between some important movie variables and grosses by providing the correlation coefficients and some other statistical data. The most important movie variables include numerical variables like budget or opening screens, and categorical variables like source or MPAA rating. Another important variable is genre. IMDB defines 19 genres, and in our experiments, we find some genres like "*Action*" and "*Adventure*" are positively correlated with grosses, while others genres like "*Biography*" and "*Documentary*" are negatively correlated with grosses. We notice the correlation coefficient between movie gross and the first week gross is as high as 0.841, which explains why some decent models could be built with post-release data in some previous literatures.

A particularly interesting question is movie budget vs. gross distribution. Figure 1 shows the normal quantile-quantile plots of movie budget and gross. Figure 1(a) shows both budget and gross are not strictly Gaussian distribution, because there are more low-grossing movies than high-grossing movies. However, if we particularly study the low-grossing movies or low-budgeting movies, Figure 1(b) and 1(c) show that high budget may result in low gross, and low budget may result in high gross as well. In this paper, we will pay more attention on high-grossing movies because they can generate more revenue and have more media exposure.

B. News Data and Correlation Analysis

Movie news data is generated form the *Lydia* system, which does high-speed analysis of online daily newspapers. The input of *Lydia* includes the coverage of around 1000 nationwide and local newspapers. One difficulty for movie news analysis is title matching, which causes lots of false positives or false negatives during entity identification phase. For example, *Lydia* may fail to identify certain movies' name like "15 Minutes", "Pride", "Next", "Interview", etc. correctly. Our solution is to filter out these "bad" data before our analysis based on three rules: 1) Common-word-named movies should be removed; 2) Movies should have a reasonable news coverage; 3) News approaching to a movie's opening date should have more references of this movie than news far away from its opening date. Eventually, we get a data set size of 498 movies, and we divided these movies into two parts - 60% as the training set and the rest 40% as the predicting set.

Lydia generates an entity database. For each entity, the *Lydia* data includes the daily article counts, daily frequency counts, as well as daily sentiment (both positive and negative) counts in seven categories: *General*, *Business*, *Crime*, *Health*, *Politics*, *Sports*, and *Media*.

Based on above raw counts, we evaluated the accumulated news references for the first week (*1-week* data), the second week through the 4th week (*1-month* data), and the 5th week through the 16th week (*4-month* data) period before the release of movies respectively. Our correlation analysis includes the evaluation of the media coverage in terms of four different entities - movie titles, directors, top 3 actors, and top 15 actors. Table II shows the correlation analysis of logged pre-release news reference counts versus logged grosses or budget under different scenarios. Table III shows the correlations between movie grosses and sentiment counts in seven categories.

1) *Movie Grosses versus News Reference Counts*: Some significant observations from our experiments are below.
   - Article counts vs. Frequencies: Grosses have higher correlation with article counts than with total entity references.
   - Time Value of Money: Higher correlations could be achieved if we take inflation into consideration based on the year-by-year interests rate before the correlation is evaluated. Therefore, time value of money is always used in this paper hereafter.
   - Raw correlation vs. Logged correlation: The logarithm operation generates higher correlations for news reference counts and grosses (or budget).
   - Grosses vs. Budget: News references are more highly correlated with grosses than budgets.
   - Pre-release and Post-release References: Table II shows that the post-release data correlates with grosses better than pre-release data.
   - Time Periods: The 1-week data has the strongest correlation, and the correlations of 1-month data and 4-month data decrease accordingly.
   - News Entities: Director references have the least correlation with grosses; movie titles and top actors have better correlations with grosses (or budget).
   - Seven Sentiment Categories: "*General*" and "*Media*" sentiment counts have the highest correlation with grosses among all seven sentiment categories.
   - Negative References vs. Positive References: From Table III, we can see that positive references are better correlated with grosses than negative ones for all sentiment categories except "*Crime*" and "*Health*".
   - Low-grossing Movies vs. High-grossing Movies: For low-grossing movies, the news references for top 3 actors are better gross predictors than those of top 15 actors. For high-grossing movies, we have the opposite conclusion.
2) *Movie Grosses versus Derived Sentiment Indexes*: Based on raw sentiment references, we derive several sentiment measures, including *polarity, subjectivity, positive references per reference, negative references per reference, and positive-negative differences per reference*. They are defined as the follows.
   - $polarity = \frac{pos\_senti\_refs}{total\_senti\_refs}$
   - $subjectivity = \frac{total\_senti\_refs}{total\_refs}$
   - $pos\_refs\_per\_ref = \frac{pos\_senti\_refs}{total\_refs}$
   - $neg\_refs\_per\_ref = \frac{neg\_senti\_refs}{total\_refs}$
   - $senti\_diffs\_per\_ref = \frac{pos\_senti\_refs - neg\_senti\_refs}{total\_refs}$

   Figure 2 shows the correlations between grosses and all these five statistics are not strong. However, correlation coefficients for several of them, such as *polarity, negative references per reference, and positive-negative differences per reference* are still statistically significant at a 0.05 significance level.
3) *Pairwise Correlation of Various News Statistical Measures*: Figure 2 shows that the pairwise correlation details. We notice article count, frequency, positive frequency, and negative frequency are highly correlated each other. To avoid multi-collinearity, our prediction model preferably use only one of them. We can also use some derived sentiment indexes because they are not strongly correlated with raw references.

IV. Prediction Models and Comparison

Two basic modeling methodologies used in this paper are regression and *k*-nearest neighbor classifiers. Regression models forecast

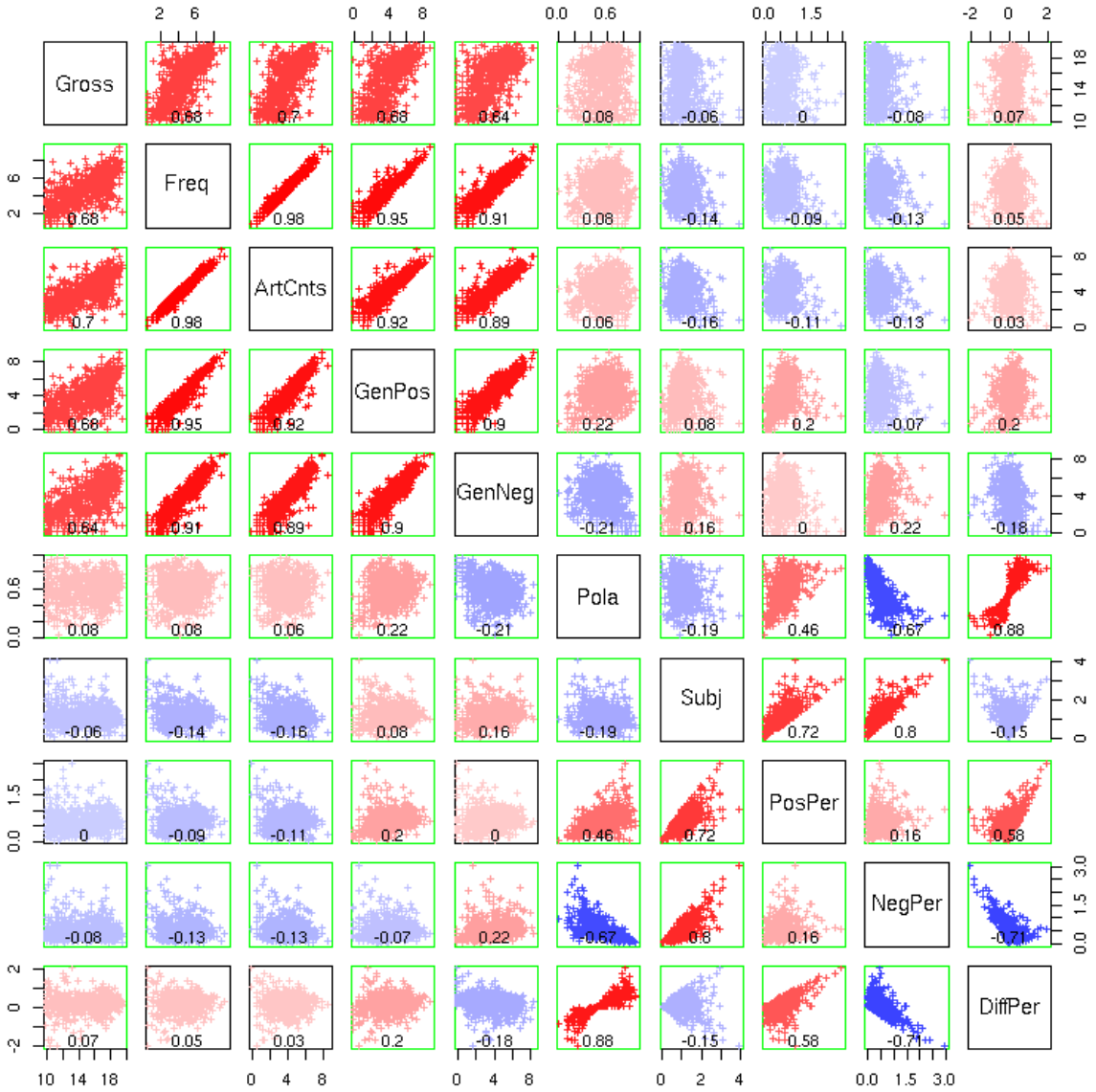


Figure 2: The pairwise plot (using 1-week, pre-release data) of move gross, news references, sentiment references, and derived sentiment indexes for the time period from January 1990 to December 2007. **Notes:** 1) The ten variables respectively are: movie gross (Gross), news frequencies (Freq), news article counts (ArtCnts), general positive counts (GenPos), general negative counts (GenNeg), polarity (Pola), subjectivity (Subj), positive references per reference (PosPer), negative references per reference (NegPer), positive-negative differences per reference (DiffPer). 2) The correlation coefficient between two variables shows on the bottom of each box. 3) Red color indicates two variables are positively correlated, while blue color indicates two variables are negatively correlated. 4) Saturation of colors indicates the strength of the correlation. 5) A green box indicates the corresponding correlation coefficient is statistically significant, while a black box means the opposite side. **Some significant observations:** 1) News references are highly correlated with movie grosses. 2) Positive references have higher correlation with grosses than negative references. 3) Frequencies, article counts, positive references, and negative references are highly correlated each other. 4) Polarity is positively correlated with movie grosses and the correlation is not strong but yet statistically significant, and so does the positive-negative differences per reference. 5) Negative references per reference is negatively correlated with movie grosses and the correlation is not strong but still statistically significant. 6) Subjectivity is negatively correlated with movie grosses, but the correlation is not statistically significant. 7) Derived sentiment indexes are not highly correlated with news references, which will give us some new information other than the raw counts.

Scenarios		General	Business	Crime	Health	Politics	Sports	Media
1 week	Positive	0.692	0.666	**0.418**	**0.520**	0.615	0.684	0.695
	Negative	0.665	0.564	**0.594**	**0.624**	0.565	0.444	0.513
1 month	Positive	0.665	0.651	**0.401**	**0.520**	0.603	0.669	0.675
	Negative	0.650	0.579	**0.580**	**0.616**	0.564	0.466	0.507
4 months	Positive	0.625	0.626	**0.370**	**0.497**	0.561	0.635	0.643
	Negative	0.608	0.544	**0.541**	**0.557**	0.531	0.438	0.490

Table III: Logged Movie Grosses versus Logged Pre-release Positive or Negative Sentiment Counts in Seven Sentiment Categories, in terms of movie title coverage. The bold numbers show that positive references are better correlated with grosses than negative ones except for "*Crime*" and "*Health*" categories. One reason is that a movie may be more attractive due to excess violence.

grosses by a regression equation. By contrast, k-NN models identify the most "similar" movie of the target movie from the training set by examining their similarities, because we think that "similar" movies should have similar grosses.

To evaluate performance (or accuracy) of models, many measures are proposed. Hyndman and Koehler [14] gave a detail description about them. Here we make some adjustment for our purpose. We suppose G is the actual gross and P is the predicted gross, and then we have below evaluation methods:

1) *AMAPE (Adjusted Mean Absolute Percentage/Relative Error)*: $AMAPE = \frac{\sum_{i=1}^{n} |APE_i|}{n}$, while $APE = \max_{abs}(\frac{G-P}{G}, \frac{G-P}{P})$ is adjusted percentage error. The operator "$\max_{abs}$" chooses the element that has the biggest absolute value. For example, for a movie whose actual gross is \$50 million, predictions \$75 million and \$33.3 million are equally good for this movie because they have the same $|APE|$ value of 0.5.
2) *Score of Models*: $Score = \frac{\sum_{i=1}^{n} (100 - \min(100, |APE_i|))}{n}$
3) $\alpha\%$ *percentage coverage*: $PC_{\alpha\%} = \frac{Number\ of\ movies\ whose\ |APE| \leq \alpha\%}{Total\ number\ of\ movies\ (n)}$

A. Prediction from Traditional Movie Variables

Traditional movie models are our base models. We build separate models according to budget information availabilities, i.e., "*budget*" and "*nobudget*" cases.

1) *Regression Models (Reg_{budget} and $Reg_{nobudget}$)*: Model Reg_{budget} use variables budget, holiday flag, MPAA rating, sequel flag, foreign flag, opening screens, and genres. Model $\text{Reg}_{nobudget}$ is the same, but with removing budget indicator. Therefore, the regression model is: $ln(G) = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + ... + \beta_k P_k + \mathfrak{e}$, where G is gross, P_is are predictors, β_is are coefficients of predictors, and $\mathfrak{e}$ is random noise.
2) *k-Nearest Neighbor Models (kNN_{budget} and $kNN_{nobudget}$)*: The similarity of movies could be measured by "distance", which is further evaluated in a multi-dimensional space. Firstly, we define the distance for each dimension. For example, the distance of two budget value B_1, B_2 are defined as: $dis(B_1, B_2) = \frac{max(B_1,B_2) - min(B_1,B_2)}{min(B_1,B_2)}$. The distance for other variables are defined accordingly. We then get the completed distance formula by Euclidean measure $Dis = \sqrt{\sum_{i=1}^{n} {dis_i}^2}$, Manhattan measure $Dis = \sum_{i=1}^{n} |dis_i|$, or regressing the training data set to determine the coefficients for all dimensions. Our experiments show that regressor works the best among all above three approaches. The basic reason is that different variables have different scales in terms of their influence on movie grosses but only the regressing method considers the difference. After this, we find the k movies from the training set which are the k nearest neighbors. In addition, our results show that the k-NN models work poorly when $k = 1$, but they work well enough while $k = 7$. Table IV shows some nearest neighbor pairs.

The performance data shows Reg_{budget} is better than $\text{Reg}_{nobudget}$, which means that budget is capable of improving performance substantially in regression models. The performance of K-NN models strongly depended on the training set size. With additional training data, and the increasing of k (but yet still a small number), the performance of K-NN models will be further improved. For all models, the high-grossing movies are predicted significantly better than low-grossing movies. The overall performance of K-NN models is similar to, but the high-grossing performance is better than that of regression models. If we use regression models for low-grossing movies and k-NN models for high-grossing movies, the best prediction will be expected.

B. Prediction from News Variables

In this section, we will predict movie grosses using news data only. Several models are built as follows.

1) *Regression Models Using News References Only ($nReg_{1w}$ and $nReg_{mov+act15}$)*: Model nReg_{1w} takes three indicators, the pre-release 1-week news article counts in terms of movie titles, top 3 actors, and top 15 actors. By contrast, $\text{nReg}_{mov+act15}$ takes six indicators, the pre-release 1-week, 1-month and 4-month news article counts in terms of movie titles and top 15 actors. The simulation result shows that models nReg_{1w} and $\text{nReg}_{mov+act15}$ have similar performance, and both of them perform better than other news-reference-based models, which means our predictors are chosen properly.
2) *Regression Models Using News References plus Sentiment Data ($nReg_{1w+senti1}$ and $nReg_{1w+senti2}$)*: Based on nReg_{1w}, $\text{nReg}_{1w+senti1}$ adds raw sentiment counts , while $\text{nReg}_{1w+senti2}$ adds derived sentiment statistics like polarity or subjectivity. However, both their overall performance and high-grossing performance have no significant improvements compared to the base model nReg_{1w}, because sentiment counts are highly correlated with the news article counts and thus carry little extra information while regressing.
3) *k-Nearest Neighbor Models ($nkNN_{1w}$, $nkNN_{mov+act15}$, and $nkNN_{1w+senti1}$)*: The three k-NN models use the same indicators as corresponding regression models. The distance of two movies can be easily computed by normalizing the reference or sentiment counts. Surprisingly, the sentiment data in the k-NN models shows some predictive power and the improvement is statistically significant. The basic reason is that the sentiment data will be helpful in identifying more similar movie pairs. Moreover, k-NN models have worse overall performance but better high-grossing performance than corresponding regression models. In addition, k-NN models using news data can achieve similar performance with IMDB models, especially for high-grossing movies.

C. Prediction from Combined Variables and Performance Comparison

We have shown that decent models can be built using either traditional IMDB data or news data. Now we build models with the combination of IMDB data and news data, and indeed yield even better prediction results. For example, in the *"nobudget"* case ("budget" is not an input variable), $\text{Reg}_{nobudget}$ is the regression model with IMDB data and it yields only a coefficient of determination R^2 of

No.	*MPAA*	*Genre*	*Source*	*Cout.*	*Scrns*	*Bgt($M)*	Gro($M)	*Date*	*Name*
1	R	Comedy	Original Screenplay	USA	7	3.500	0.221	09/14/07	Ira and Abby
	R	Comedy	Original Screenplay	USA	7	3.500	0.107	02/17/06	Winter Passing
2	PG-13	Adventure	Based on Book or Short Story	UK	4285	150.000	292.005	07/11/07	Harry Potter and the Order of the Phoenix
	PG-13	Adventure	Based on Book or Short Story	UK	3858	150.000	290.013	11/18/05	Harry Potter and the Goblet of Fire
3	R	Action	Original Screenplay	USA	2	1.000	0.000884	04/21/06	In Her Line of Fire
	R	Adventure	Original Screenplay	USA	2	1.000	9.015	12/02/05	Transamerica

Table IV: Example of Nearest Neighbor Pairs Identified with Numerical and Categorical Indicators (Model kNN$_{budget}$). Pair 2 shows that the algorithm identifies one "Harry Potter" movie with another "Harry Potter" movie as its nearest neighbor, which indicates a very good comparison. Pair 3 is a strange pair, which is an almost perfectly matched pair but their grosses differ substantially. But generally speaking, the prediction based on nearest neighbors achieves similar performance with regression models.

Regression Models					
Predictor	Model	Perf$_{Overall}$		Perf$_{High}$	
		AMAPE	Score	AMAPE	Score
IMDB	Reg$_{nobudget}$	**7.83**	**92.8**	**8.97**	**92.41**
	Reg$_{budget}$	3.53	96.47	2.03	97.97
News	nReg$_{1w}$	**8.72**	**92.1**	**4.02**	**96.2**
	nReg$_{mov+act15}$	10.46	92.07	2.87	97.13
Combined	Reg$_{nobudget}$+nReg$_{1w}$	**3.82**	**96.81**	**2.48**	**97.52**
	Reg$_{nobudget}$+nReg$_{mov+act15}$	3.79	96.21	2.4	97.6
	Reg$_{budget}$+nReg$_{1w}$	2.76	97.24	1.57	98.43
	Reg$_{budget}$+nReg$_{mov+act15}$	2.63	97.37	1.54	98.46
k-Nearest Neighbor Models					
Source	Model	Perf$_{Overall}$		Perf$_{High}$	
		AMAPE	Score	AMAPE	Score
IMDB	kNN$_{nobudget}$	**18.66**	**89.9**	**2.44**	**97.56**
	kNN$_{budget}$	11.68	92.11	1.16	98.84
News	nkNN$_{1w}$	**24.22**	**87.25**	**1.79**	**98.21**
	nkNN$_{mov+act15}$	21.6	87.57	2.2	97.8
Combined	kNN$_{nobudget}$+nkNN$_{1w}$	**11.13**	**92.03**	**1.14**	**98.87**
	kNN$_{nobudget}$+nkNN$_{mov+act15}$	10.89	92.17	1.16	98.84
	kNN$_{budget}$+nkNN$_{1w}$	3.37	96.88	1.06	98.91
	kNN$_{budget}$+nkNN$_{mov+act15}$	5.82	95.13	1.01	98.99

Table V: Performance Comparison for IMDB, News, and Combined Models. The bold numbers show the comparison of a group of experiments. The data proves: 1) For regression methods, the news models have similar overall accuracy to, but better accuracy of high-grossing movies than IMDB models. 2) For k-NN methods, the news models have worse overall accuracy, but yet still better accuracy of high-grossing movies than IMDB models. 3) For both regression and k-NN methods, the combined models prove superior to both IMDB and news models for either overall accuracy or accuracy of high-grossing movies. Other groups of experiments indicate the same results.

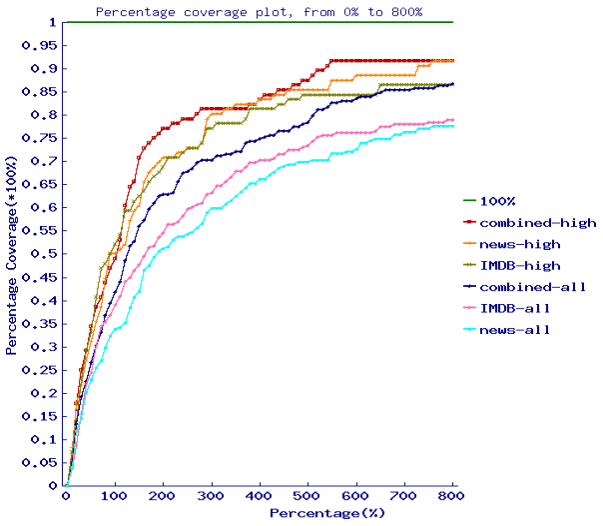


Figure 3: Comparison of Regression Models (*"nobudget"* case). These models use IMDB data, news data and their combination respectively. The combined model works best among all three models, both for overall performance and high-grossing performance.

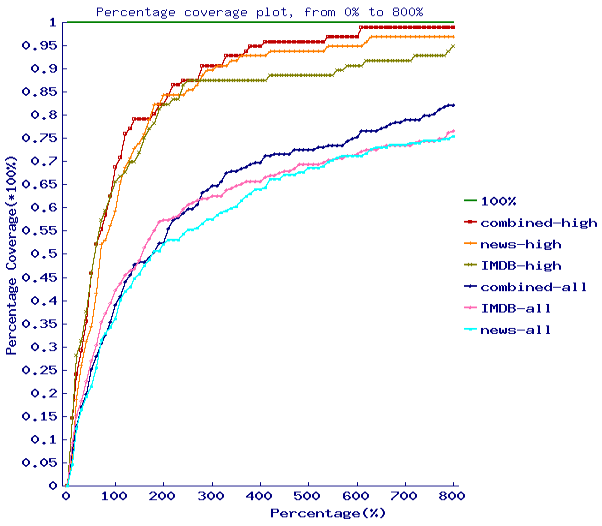


Figure 4: Comparison of k-NN Models (*"nobudget"* case). These models use IMDB data, news data and their combination respectively. The combined model works best among all three models, both for overall performance and high-grossing performance.

0.448, which is almost the same with the result of pre-release model from Simonoff and Sparrow [6]. By contrast, $\text{Reg}_{nobudget}$+nReg_{1w} is the corresponding regression model with IMDB data plus news data, and it achieves a R^2 of 0.788, which indicates a big improvement.

We studied the adjusted percentage error (or residual) plots for IMDB models, news models, and combined models. The results show some movies' grosses are highly overestimated and some others are highly underestimated if we use only IMDB data, and the error plots are not symmetric or with zero mean. However, the high deviations are smoothed by news indicators, i.e., highly underestimated or overestimated grosses are eliminated in news models or combined models. Furthermore, the combined models make the the adjusted percentage error plots symmetric, which is a sign of another benefit of using news data.

The completed performance data of IMDB models, news models, and combined models are listed in Table V. Compared to pure IMDB models or pure news models, the combined models yield nice performance improvement, either for regression models or k-NN models, which can be indicated by smaller AMAPE and higher scores. Our t-test proves the improvement is statistically significant.

Figures 3 and 4 show the "Percentage vs. Percentage Coverage" comparison (*"nobudget"* case) of IMDB, news, and the combined models. The X-axis shows the $\alpha\%$ percentage, while the Y-axis shows the corresponding $\alpha\%$ percentage coverage. These plots show both overall performance and high-grossing performance of combined models are higher than those of IMDB or news models. We have exactly the same conclusion for *"budget"* case. Furthermore, the comparison of Figures 3 and 4 also shows that regression models work better for overall performance, while k-NN models perform better for high-grossing performance. That is, regression models are more suitable for low-grossing movies, but k-NN models are more suitable for high-grossing movies.

V. Conclusions

We have discussed the correlation of movie grosses with both traditional IMDB data and movie news data, and built models with IMDB data, news data, and their combination respectively. Our experiments proved media's predictive power in movie gross prediction.

Detailed conclusions are as the follows. Firstly, movie news references are highly correlated with movie grosses, and sentiment measures including derived sentiment indexes are also correlated with movie grosses. Secondly, movie gross prediction can be done by either IMDB data, news data, or their combination. Prediction models using merely news data can achieve similar performance with models using IMDB data, especially for high-grossing movies, while the combined models using both IMDB and news data yield the best result. Therefore, news data is proven to be capable of improving movie gross prediction in our analysis. Thirdly, both regression and k-nearest neighbor classifiers can be used for movie gross prediction. With the same indicators, regression models have better low-grossing performance, but k-NN models have better high-grossing performance. Finally, article counts for movie entities are good movie gross predictors. News sentiment data are good predictors for k-NN models, but not good predictors for regression models.

For future work, we plan to compare our results with large-scale analysis of blog data, web reviews, as well as news data, to determine what kind of sources has greater predictive power over what time scale.

References

[1] G. Fung, J. Yu, and W. Lam, "Stock prediction: Integrating text mining approach using real-time news," in *Proceedings of IEEE Int. Conference on Computational Intelligence for Financial Engineering*, 2003, pp. 395–402.

[2] W. S. Chan, "Stock price reaction to news and no-news: Drift and reversal after headlines," *Journal of Financial Economics*, vol. 70, pp. 223–260, 2003.

[3] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy, "More than words: Quantifying language to measure firms' fundamentals," in *Proceedings of 9th Annual Texas Finance Festival*, May 2007.

[4] L. Lloyd, D. Kechagias, and S. Skiena, "Lydia: A system for large-scale news analysis," in *Proceedings of 12th String Processing and Information Retrieval (SPIRE 2005)*, vol. LNCS 3772, Buenos Aires, Argentina, 2005, pp. 161–166.

[5] A. Chen, "Forecasting gross revenues at the movie box office," *Working paper, University of Washington, Seattle, WA*, June 2002.

[6] J. S. Simonoff and I. R. Sparrow, "Predicting movie grosses: Winners and losers, blockbusters and sleepers," *Chance*, vol. 13(3), pp. 15–24, 2000.

[7] M. S. Sawhney and J. Eliashberg, "A parsimonious model for forecasting gross box-office revenues of motion pictures," *Marketing Science*, vol. Vol. 15, No. 2, pp. 113–131, 1996.

[8] R. Sharda and E. Meany, "Forecasting gate receipts using neural network and rough sets," in *Proceedings of the International DSI Conference*, 2000, pp. 1–5.

[9] R. Sharda and D. Delen, "Forecasting box-office success of motion pictures with neural networks," *Expert Systems with Applications*, vol. 30, pp. 243–254, 2006.

[10] B. Pang and L. Lee, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, July 2002, pp. 79–86.

[11] P. Chaovalit and L. Zhou, "Movie review mining: a comparison between supervised and unsupervised classification approaches," in *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, 2005.

[12] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. Vol. 2, No 1-2, pp. 1–135, 2008.

[13] G. Mishne and N. Glance, "Predicting movie sales from blogger sentiment," *In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 155–158, 2006.

[14] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Internationl Journal of Forecasting*, vol. 22, pp. 679–688, 2006.