Movie Reviews and Revenues: An Experiment in Text Regression[*]

Mahesh Joshi   Dipanjan Das   Kevin Gimpel   Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{maheshj,dipanjan,kgimpel,nasmith}@cs.cmu.edu

Abstract

We consider the problem of predicting a movie's opening weekend revenue. Previous work on this problem has used metadata about a movie—e.g., its genre, MPAA rating, and cast—with very limited work making use of text *about* the movie. In this paper, we use the text of film critics' reviews from several sources to predict opening weekend revenue. We describe a new dataset pairing movie reviews with metadata and revenue data, and show that review text can substitute for metadata, and even improve over it, for prediction.

1 Introduction

Predicting gross revenue for movies is a problem that has been studied in economics, marketing, statistics, and forecasting. Apart from the economic value of such predictions, we view the forecasting problem as an application of NLP. In this paper, we use the text of critics' reviews to predict opening weekend revenue. We also consider metadata for each movie that has been shown to be successful for similar prediction tasks in previous work.

There is a large body of prior work aimed at predicting gross revenue of movies (Simonoff and Sparrow, 2000; Sharda and Delen, 2006; *inter alia*). Certain information is used in nearly all prior work on these tasks, such as the movie's genre, MPAA rating, running time, release date, the number of screens on which the movie debuted, and the presence of particular actors or actresses in the cast. Most prior text-based work has used automatic text analysis tools, deriving a small number of aggregate statistics. For example, Mishne and Glance (2006) applied sentiment analysis techniques to pre-release and post-release blog posts about movies and showed higher correlation between actual revenue and sentiment-based metrics, as compared to mention counts of the movie. (They did not frame the task as a revenue prediction problem.) Zhang and Skiena (2009) used a news aggregation system to identify entities and obtain domain-specific sentiment for each entity in several domains. They used the aggregate sentiment scores and mention counts of each movie in news articles as predictors.

While there has been substantial prior work on using critics' reviews, to our knowledge all of this work has used polarity of the review or the number of stars given to it by a critic, rather than the review text directly (Terry et al., 2005).

Our task is related to sentiment analysis (Pang et al., 2002) on movie reviews. The key difference is that our goal is to predict a *future* real-valued quantity, restricting us from using any post-release text data such as user reviews. Further, the most important clues about revenue may have little to do with whether the reviewer *liked* the movie, but rather what the reviewer found worth mentioning. This paper is more in the tradition of Ghose et al. (2007) and Kogan et al. (2009), who used text regression to directly quantify review "value" and make predictions about future financial variables, respectively.

Our aim in using the full text is to identify particular words and phrases that predict the movie-going tendencies of the public. We can also perform syntactic and semantic analysis on the text to identify richer constructions that are good predictors. Furthermore, since we consider multiple reviews for each movie, we can compare these features across reviews to observe how they differ both in frequency and predictive performance across different media outlets and individual critics.

In this paper, we use linear regression from text and non-text (meta) features to directly predict gross revenue aggregated over the opening weekend, and the same averaged per screen.

[*] We appreciate reviewer feedback and technical advice from Brendan O'Connor. This work was supported by NSF IIS-0803482, NSF IIS-0844507, and DARPA NBCH-1080004.

Domain	train	dev	test	total
Austin Chronicle	306	94	62	462
Boston Globe	461	154	116	731
LA Times	610	2	13	625
Entertainment Weekly	644	208	187	1039
New York Times	878	273	224	1375
Variety	927	297	230	1454
Village Voice	953	245	198	1396
# movies	1147	317	254	1718

Table 1: Total number of reviews from each domain for the training, development and test sets.

2 Data

We gathered data for movies released in 2005–2009. For these movies, we obtained metadata and a list of hyperlinks to movie reviews by crawling MetaCritic (`www.metacritic.com`). The metadata include the name of the movie, its production house, the set of genres it belongs to, the scriptwriter(s), the director(s), the country of origin, the primary actors and actresses starring in the movie, the release date, its MPAA rating, and its running time. From The Numbers (`www.the-numbers.com`), we retrieved each movie's production budget, opening weekend gross revenue, and the number of screens on which it played during its opening weekend. Only movies found on both MetaCritic and The Numbers were included.

Next we chose seven review websites that most frequently appeared in the review lists for movies at Metacritic, and obtained the text of the reviews by scraping the raw HTML. The sites chosen were the *Austin Chronicle*, the *Boston Globe*, the *LA Times*, *Entertainment Weekly*, the *New York Times*, *Variety*, and the *Village Voice*. We only chose those reviews that appeared on or before the release date of the movie (to ensure that revenue information is not present in the review), arriving at a set of 1718 movies with at least one review. We partitioned this set of movies temporally into training (2005–2007), development (2008) and test (2009) sets. Not all movies had reviews at all sites (see Table 1).

3 Predictive Task

We consider two response variables, both in U.S. dollars: the total revenue generated by a movie during its release weekend, and the *per screen* revenue during the release weekend. We evaluate these predictions using (1) mean absolute error (MAE) in U.S. dollars and (2) Pearson's correlation between the actual and predicted revenue.

We use linear regression to directly predict the opening weekend gross earnings, denoted y, based on features $\boldsymbol{x}$ extracted from the movie metadata and/or the text of the reviews. That is, given an input feature vector $\boldsymbol{x} \in \mathbb{R}^p$, we predict an output $\hat{y} \in \mathbb{R}$ using a linear model: $\hat{y} = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}$. To learn values for the parameters $\boldsymbol{\theta} = \langle \beta_0, \boldsymbol{\beta} \rangle$, the standard approach is to minimize the sum of squared errors for a training set containing n pairs $\langle \boldsymbol{x}_i, y_i \rangle$ where $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ for $1 \leq i \leq n$:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}=(\beta_0,\boldsymbol{\beta})} \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}) \Big)^2 + \lambda P(\boldsymbol{\beta})$$

A penalty term $P(\boldsymbol{\beta})$ is included in the objective for regularization. Classical solutions use an ℓ_2 or ℓ_1 norm, known respectively as ridge and lasso regression. Introduced recently is a mixture of the two, called the elastic net (Zou and Hastie, 2005):

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \Big(\tfrac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \Big)$$

where $\alpha \in (0, 1)$ determines the trade-off between ℓ_1 and ℓ_2 regularization. For our experiments we used the elastic net and specifically the `glmnet` package which contains an implementation of an efficient coordinate ascent procedure for training (Friedman et al., 2008).

We tune the α and λ parameters on our development set and select the model with the $\langle \alpha, \lambda \rangle$ combination that yields minimum MAE on the development set.

4 Experiments

We compare predictors based on metadata, predictors based on text, and predictors that use both kinds of information. Results for two simple baselines of predicting the training set mean and median are reported in Table 2 (Pearson's correlation is undefined since the standard deviation is zero).

4.1 Metadata Features

We considered seven types of metadata features, and evaluated their performance by adding them to our pool of features in the following order: whether the film is of U.S. origin, running time (in minutes), the logarithm of its budget, # opening screens, genre (e.g., Action, Comedy) and MPAA rating (e.g., G, PG, PG-13), whether the movie opened on a holiday weekend or in summer months, total count as well as of presence of individual Oscar-winning actors and directors and high-grossing actors. For the first task of predicting the total opening weekend revenue of a movie, the best-performing feature set in terms of MAE turned out to be all the features. However, for the second task of predicting the *per screen* revenue, addition of the last feature subset consisting of information related to the actors and directors hurt performance (MAE increased). Therefore, for the second task, the best performing set contained only the first six types of metadata features.

4.2 Text Features

We extract three types of text features (described below). We only included feature instances that occurred in at least five different movies' reviews. We stem and downcase individual word components in all our features.

I. n-grams. We considered unigrams, bigrams, and trigrams. A 25-word stoplist was used; bigrams and trigrams were only filtered if all words were stopwords.

II. Part-of-speech n-grams. As with words, we added unigrams, bigrams, and trigrams. Tags were obtained from the Stanford part-of-speech tagger (Toutanova and Manning, 2000).

III. Dependency relations. We used the Stanford parser (Klein and Manning, 2003) to parse the critic reviews and extract syntactic dependencies. The dependency relation features consist of just the relation part of a dependency triple ⟨relation, head word, modifier word⟩.

We consider three ways to combine the collection of reviews for a given movie. The first ("$-$") simply concatenates all of a movie's reviews into a single document before extracting features. The second ("$+$") conjoins each feature with the source site (e.g., *New York Times*) from whose review it was extracted. A third version (denoted "B") combines both the site-agnostic and site-specific features.

			Total		**Per Screen**	
	Features	**Site**	**MAE ($M)**	r	**MAE ($K)**	r
	Predict mean		11.672	–	6.862	–
	Predict median		10.521	–	6.642	–
meta	Best		5.983	0.722	6.540	0.272
text		$-$	8.013	0.743	6.509	0.222
	I	$+$	7.722	0.781	6.071	0.466
	see Tab. 3	B	7.627	0.793	6.060	0.411
		$-$	8.060	0.743	6.542	0.233
	I ∪ II	$+$	**7.420**	0.761	6.240	0.398
		B	7.447	0.778	6.299	0.363
		$-$	8.005	0.744	6.505	0.223
	I ∪ III	$+$	7.721	0.785	6.013	**0.473**
		B	7.595	**0.796**	†**6.010**	0.421
meta ∪ text		$-$	5.921	**0.819**	6.509	0.222
	I	$+$	5.757	0.810	6.063	0.470
		B	5.750	**0.819**	6.052	0.414
		$-$	5.952	0.818	6.542	0.233
	I ∪ II	$+$	5.752	0.800	6.230	0.400
		B	5.740	**0.819**	6.276	0.358
		$-$	5.921	**0.819**	6.505	0.223
	I ∪ III	$+$	**5.738**	0.812	6.003	**0.477**
		B	5.750	**0.819**	†**5.998**	0.423

Table 2: Test-set performance for various models, measured using mean absolute error (MAE) and Pearson's correlation (r), for two prediction tasks. Within a column, **boldface** shows the best result among "text" and "meta ∪ text" settings. †Significantly better than the meta baseline with $p < 0.01$, using the Wilcoxon signed rank test.

4.3 Results

Table 2 shows our results for both prediction tasks. For the total first-weekend revenue prediction task, metadata features baseline result ($r^2 = 0.521$) is comparable to that reported by Simonoff and Sparrow (2000) on a similar task of movie gross prediction ($r^2 = 0.446$). Features from critics' reviews by themselves improve correlation on both prediction tasks, however improvement in MAE is only observed for the *per screen* revenue prediction task.

A combination of the meta and text features achieves the best performance both in terms of MAE and r. While the text-only models have some high negative weight features, the combined models do not have any negatively weighted features and only a very few metadata features. That is, the text is able to substitute for the other metadata features.

Among the different types of text-based features that we tried, lexical n-grams proved to be a strong baseline to beat. None of the "I ∪ $*$" feature sets are significantly better than n-grams alone, but adding the dependency relation features (set III) to the n-grams does improve the performance enough to make it significantly better than the metadata-only baseline for per screen revenue prediction.

Salient Text Features: Table 3 lists some of the highly weighted features, which we have categorized manually. The features are from the text-only model annotated in Table 2 (total, not per screen). The feature weights can be directly interpreted as U.S. dollars contributed to the predicted value $\hat{y}$ by each occurrence of the feature. Sentiment-related features are not as prominent as might be expected, and their overall proportion in the set of features with non-zero weights is quite small (estimated in preliminary trials at less than 15%). Phrases that refer to metadata are the more highly weighted and frequent ones. Consistent with previous research, we found some positively-oriented sentiment features to be predictive. Some other prominent features not listed in the table correspond to special effects (“*Boston Globe*: of_the_art”, “and_cgi”), particular movie franchises (“shrek_movies”, “*Variety*: chronicle_of”, “voldemort”), hype/expectations (“blockbuster”, “anticipation”), film festival (“*Variety*: canne” with negative weight) and time of release (“summer_movie”).

	Feature	**Weight ($M)**
rating	pg	+0.085
	New York Times: adult	-0.236
	New York Times: rate_r	-0.364
sequels	this_series	+13.925
	LA Times: the_franchise	+5.112
	Variety: the_sequel	+4.224
people	*Boston Globe*: will_smith	+2.560
	Variety: brittany	+1.128
	ˆ_producer_brian	+0.486
genre	*Variety*: testosterone	+1.945
	Ent. Weekly: comedy_for	+1.143
	Variety: a_horror	+0.595
	documentary	-0.037
	independent	-0.127
sentiment	*Boston Globe*: best_parts_of	+1.462
	Boston Globe: smart_enough	+1.449
	LA Times: a_good_thing	+1.117
	shame_$	-0.098
	bogeyman	-0.689
plot	*Variety*: torso	+9.054
	vehicle_in	+5.827
	superhero_$	+2.020

Table 3: Highly weighted features categorized manually. ˆ and $ denote sentence boundaries. “brittany” frequently refers to Brittany Snow and Brittany Murphy. “ˆ_producer_brian” refers to producer Brian Grazer (*The Da Vinci Code*, among others).

5 Conclusion

We conclude that text features from pre-release reviews can substitute for and improve over a strong metadata-based first-weekend movie revenue prediction. The dataset used in this paper has been made available for research at http://www.ark.cs.cmu.edu/movie$-data.

References

J. Friedman, T. Hastie, and R. Tibshirani. 2008. Regularized paths for generalized linear models via coordinate descent. Technical report, Stanford University.

A. Ghose, P. G. Ipeirotis, and A. Sundararajan. 2007. Opinion mining using econometrics: A case study on reputation systems. In *Proc. of ACL*.

D. Klein and C. D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in NIPS 15*.

S. Kogan, D. Levin, B. R. Routledge, J. Sagi, and N. A. Smith. 2009. Predicting risk from financial reports with regression. In *Proc. of NAACL*.

G. Mishne and N. Glance. 2006. Predicting movie sales from blogger sentiment. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. of EMNLP*.

R. Sharda and D. Delen. 2006. Predicting box office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2):243–254.

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

N. Terry, M. Butler, and D. De’Armond. 2005. The determinants of domestic box office performance in the motion picture industry. *Southwestern Economic Review*, 32:137–148.

K. Toutanova and C. D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proc. of EMNLP*.

W. Zhang and S. Skiena. 2009. Improving movie gross prediction through news analysis. In *Proc. of Web Intelligence and Intelligent Agent Technology*.

H. Zou and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal Of The Royal Statistical Society Series B*, 67(5):768–768.