Los Angeles Times | ENTERTAINMENT VARIETY The New York Times boston.com

Movie Reviews and Revenues: An Experiment in Text Regression

Mahesh Joshi, Dipanjan Das, Kevin Gimpel and Noah A. Smith

Language Technologies Institute, Carnegie Mellon University

{maheshj,dipanjan,kgimpel,nasmith}@cs.cmu.edu

Carnegie Mellon

lti

I. The Story in Short

- Use metadata and critics' reviews to predict opening weekend revenues of movies
- Feature analysis shows what aspects of reviews predict box office success

II. Data

- 1718 Movies, released 2005-2009
- Metadata (genre, rating, running time, actors, director, etc.): *www.metacritic.com*
- Critics' reviews (~7K): Austin Chronicle, Boston Globe, Entertainment Weekly, LA Times, NY Times, Variety, Village Voice
- Opening weekend revenues and number of opening screens: *www.the-numbers.com*

III. Model

- Linear regression with the elastic net (Zou and Hastie, 2005)

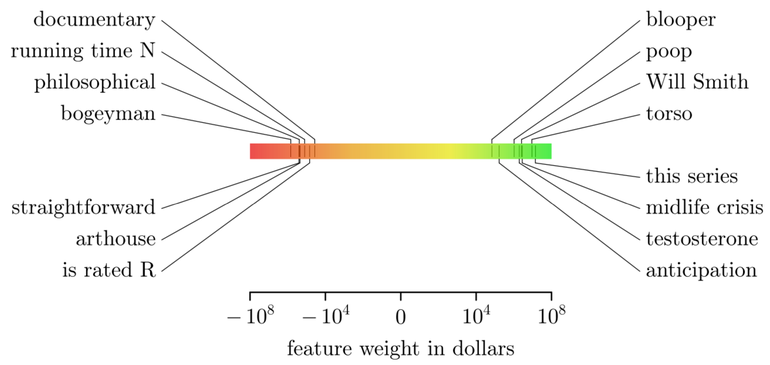
$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}=(\beta_0,\boldsymbol{\beta})}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})\right)^2 + \lambda P(\boldsymbol{\beta})$$

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

IV. Features

I	Lexical n-grams (1,2,3)
II	Part-of-speech n-grams (1,2,3)
III	Dependency relations (nsubj,advmod,...)
Meta	U.S. origin, running time, budget (log), # of opening screens, genre, MPAA rating, holiday release (summer, Christmas, Memorial day,...), star power (Oscar winners, high-grossing actors)

V. What May Have Brought You Here



VI. Results

- Text features can substitute for and improve upon metadata

	Features	Site	Total MAE ($M)	Total r	Per Screen MAE ($K)	Per Screen r
	Predict mean		11.672	–	6.862	–
	Predict median		10.521	–	6.642	–
meta	Best		5.983	0.722	6.540	0.272
text	I	–	8.013	0.743	6.509	0.222
	I	+	7.722	0.781	6.071	0.466
	see Tab. 2	B	7.627	0.793	6.060	0.411
	I ∪ II	–	8.060	0.743	6.542	0.233
	I ∪ II	+	**7.420**	0.761	6.240	0.398
	I ∪ II	B	7.447	0.778	6.299	0.363
	I ∪ III	–	8.005	0.744	6.505	0.223
	I ∪ III	+	7.721	0.785	6.013	**0.473**
	I ∪ III	B	7.595	**0.796**	†**6.010**	0.421
meta ∪ text	I	–	5.921	**0.819**	6.509	0.222
	I	+	5.757	0.810	6.063	0.470
	I	B	5.750	**0.819**	6.052	0.414
	I ∪ II	–	5.952	0.818	6.542	0.233
	I ∪ II	+	5.752	0.800	6.230	0.400
	I ∪ II	B	5.740	**0.819**	6.276	0.358
	I ∪ III	–	5.921	**0.819**	6.505	0.223
	I ∪ III	+	**5.738**	0.812	6.003	**0.477**
	I ∪ III	B	5.750	**0.819**	†**5.998**	0.423

Table 1: 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.

VII. More Cool Features

	Feature	Weight ($M)
rating	pg	+0.085
	New York Times: adult	-0.236
	New York Times: rate_r	-0.364
sequels	this_series	+13.925
	LA Times: the_franchise	+5.112
	Variety: the_sequel	+4.224
people	*Boston Globe*: will_smith	+2.560
	Variety: brittany	+1.128
	^_producer_brian	+0.486
genre	*Variety*: testosterone	+1.945
	Ent. Weekly: comedy_for	+1.143
	Variety: a_horror	+0.595
	documentary	-0.037
	independent	-0.127
sentiment	*Boston Globe*: best_parts_of	+1.462
	Boston Globe: smart_enough	+1.449
	LA Times: a_good_thing	+1.117
	shame_$	-0.098
	bogeyman	-0.689
plot	*Variety*: torso	+9.054
	vehicle_in	+5.827
	with_her_boyfriend	+3.408
	superhero_$	+2.020
release date	summer_movie	+2.671
expectations	it_usually	+4.027
	Boston Globe: blockbuster	+3.694
	anticipation	+0.166
other	*Boston Globe*: of_the_art	+8.700
	and_cgi	+4.106
	Village Voice: canne	-0.112

Table 2: Highly weighted features categorized manually. ^ and $ denote sentence boundaries. "brittany" frequently refers to Brittany Snow and Brittany Murphy. "^_producer_brian" refers to producer Brian Grazer (*The Da Vinci Code*, among others).

VIII. Get the Data!

www.ark.cs.cmu.edu/movie$-data