Extracting Features of Entertainment Products: A Guided LDA Approach Informed by the Psychology of Media Consumption

Olivier Toubia[*], Garud Iyengar[†], Renée Bunnell[‡], and Alain Lemaire[§]

[*]Graduate School of Business, Columbia University.
[†]Industrial Engineering and Operations Research Department, Columbia University.
[‡]Real.org; Real Engagement and Loyalty (REAL); OWEN.AI
[§]Graduate School of Business, Columbia University.

Web Appendices

A Guided LDA Estimation

Following Jagarlamudi et al. (2012), we set α_1=0.01 and α_2=1. In topic K we set the value of α_1 to 60 for the last word W, to capture the fact that the "all other" word is prominent in the baseline topic. Given this specification, the posterior distributions of all variables are given in closed form as follows:

$$Prob(z_i^d = k | w_i^d, \{\phi_k^r\}, \{\phi_k^s\}, \{\pi_k\}, \theta_d) =$$

$$\frac{[Prob(w_i^d | z_i^d = k, x_i^d = 1, \phi_k^s)\pi_k + Prob(w_i^d | z_i^d = k, x_i^d = 0, \phi_k^r)(1-\pi_k)]Prob(z_i^d = k|\theta_d)}{\sum_{k'}[Prob(w_i^d | z_i^d = k', x_i^d = 1, \phi_{k'}^s)\pi_{k'} + Prob(w_i^d | z_i^d = k', x_i^d = 0, \phi_{k'}^r)(1-\pi_{k'})]Prob(z_i^d = k'|\theta_d)}$$

$$= \frac{[\phi_k^s(w_i^d)\pi_k + \phi_k^r(w_i^d)(1-\pi_k)]\theta_d(k)}{\sum_{k'}[\phi_{k'}^s(w_i^d)\pi_{k'} + \phi_{k'}^r(w_i^d)(1-\pi_{k'})]\theta_d(k')} \tag{WA1}$$

$$Prob(x_i^d = 1 | w_i^d, z_i^d, \{\phi_k^r\}, \{\phi_k^s\}, \{\pi_k\}) = \frac{\phi_{z_i^d}^s(w_i^d)\pi_{z_i^d}}{\phi_{z_i^d}^s(w_i^d)\pi_{z_i^d} + \phi_{z_i^d}^r(w_i^d)(1-\pi_{z_i^d})} \tag{WA2}$$

$$Prob(\phi_k^r | \{z_i^d\}, \{x_i^d\}, \{w_i^d\}) =$$

$$Dirichlet(\alpha_1 l_k^r(1) + \sum_{(i,d):z_i^d=k \& x_i^d=0} 1(w_i^d = 1), \ldots, \alpha_1 l_k^r(W) + \sum_{(i,d):z_i^d=k \& x_i^d=0} 1(w_i^d = W)) \tag{WA3}$$

$$Prob(\phi_k^s | \{z_i^d\}, \{x_i^d\}, \{w_i^d\}) =$$

$$Dirichlet(\alpha_1 l_k^s(1) + \sum_{(i,d):z_i^d=k \& x_i^d=1} 1(w_i^d = 1), \ldots, \alpha_1 l_k^s(W) + \sum_{(i,d):z_i^d=k \& x_i^d=1} 1(w_i^d = W)) \tag{WA4}$$

$$Prob(\pi_k|\{z_i^d\},\{x_i^d\},\{w_i^d\}) =$$
$$Dirichlet(1+\sum_{(i,d):z_i^d=k} x_i^d,\ldots,1+\sum_{(i,d):z_i^d=k}(1-x_i^d)) \tag{WA5}$$

$$Prob(\theta_d|\{z_i^d\}) = Dirichlet(\alpha_2+\sum_i 1(z_i^d=1),\ldots\alpha_2+\sum_i 1(z_i^d=K)) \tag{WA6}$$

Equation WA1 simply applies Bayes' rule. The posterior probability that token i in document d belongs to topic k given that it is equal to word w_i^d and given all other parameters is proportional to the prior distribution (given by θ_d, where $\theta_k(d)$ is the d^{th} element of θ_k) multiplied by the probability of drawing word w_i^d given topic k (given by ϕ_k^r, ϕ_r^s, and π_k). A new set of latent variables $\{z_i^d\}$ is drawn at each iteration of the Gibbs sampler, according to this Equation. Equation WA2 similarly follows Bayes' rule, and a new set of latent variables $\{x_i^d\}$ is drawn at each iteration of the Gibbs sampler. Equations WA3 and WA4 follow from the conjugate properties of the Dirichlet distribution. For each topic k, given the set of latent variables $\{z_i^d\}$, we simply count the number of occurrences of each word among tokens that were assigned to each version of topic k across all documents. Equation WA5 also follows from the conjugate properties of the Dirichlet distribution. For each topic, we count the number of times token were assigned to each version. Equation WA6 also follows from the conjugate properties of the Dirichlet distribution. For each document d, given the set of latent variables $\{z_i^d\}$, we simply count the number of tokens that were assigned to each topic in this document. We note that computation time may be improved slightly by using a collapsed Gibbs sampler that integrates out over $\{\phi_k\}$ and $\{\theta_d\}$ (Griffiths and Steyvers, 2004).

B Word Clouds for Topics Reported in Table ??

For each topic in Table **??**, we simulate 500 words based on the exponential of the relevance measure. That is, we draw from all words in the vocabulary (seed as well as non-seed) with replacement, where the probability associated with each word is proportional to the exponent of its relevance for that topic. The word clouds below present a visualization of the simulated sets of words, where the font size of a word in each cloud is proportional to its occurrence among the simulated words for that topic. Note that we remove words that represent first names from the clouds.

Figure WApp 1: "Citizenship 4"

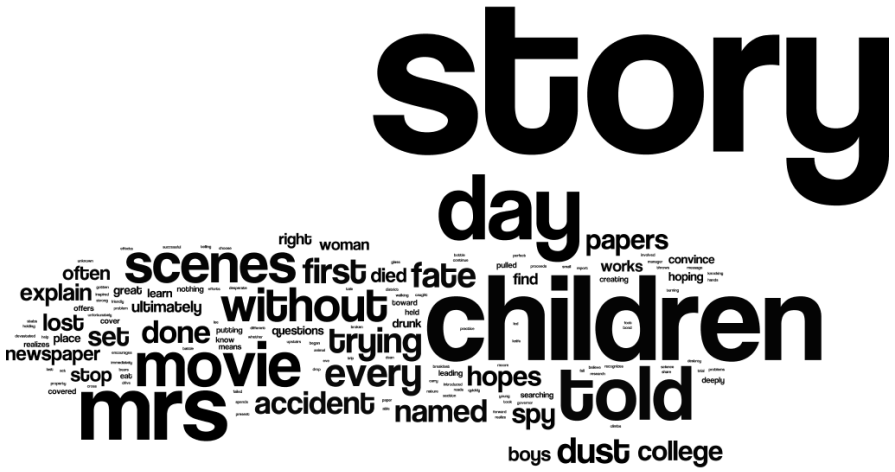

Figure WApp 2: "Creativity 4"

Figure WApp 3: "Fairness 1"

Figure WApp 4: "Leadership 3"

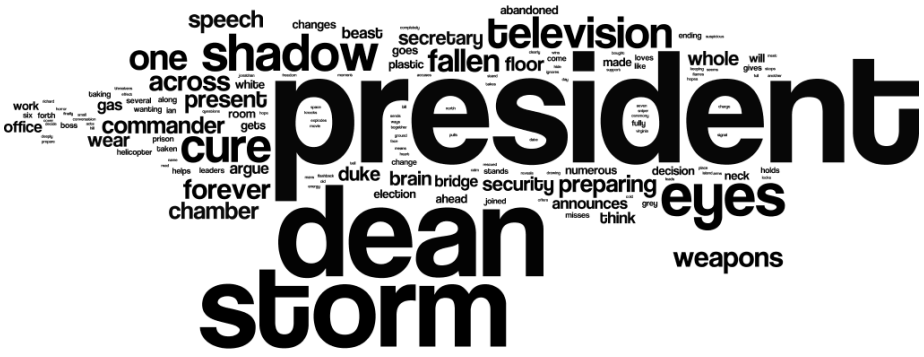

Figure WApp 5: "Leadership 4"

Figure WApp 6: "Love 1"

Figure WApp 7: "Love 3"

Figure WApp 8: "Love 4"

Figure WApp 9: "Love of Learning 3"

Figure WApp 10: "Vitality 3"

C Alternative Estimation Approach Based on Machine Learning

Given the relatively large number of features compared to the number of observations per consumer, we test a machine-learning estimation approach that explicitly controls for complexity. In particular, we estimate the various specifications of the content-based model using the LOG-Het method proposed by Evgeniou et al. (2007). This method was specifically designed for individual-level choice data with a panel structure like ours. We directly use the matlab code made publicly available by these authors. Details of the method are provided in Section 2.2. of Evgeniou et al. (2007).

Preferences are estimated at the individual level (simultaneously for all consumers) by solving the following convex optimization problem:

$$\text{Minimize}_{\{W_c\},W_0,D} \; -\frac{1}{\gamma^*}\Sigma_c\Sigma_m log(Prob(y_{cm}))^{y_{cm}} log(1-Prob(y_{cm}))^{1-y_{cm}} + \Sigma_c(W_c-W_0)^T D^{-1}(W_c-W_0)$$

$$\text{Subject to: } D \text{ is a positive semidefinite matrix scaled to have trace 1}$$

Where $Prob(y_{cm})$ is given by Equation **??** in the present paper. The first term of the objective function captures fit, and the second term captures complexity. The parameter γ^* captures the trade-off between fit and complexity. It is selected (separately for each model specification) using

leave-one-out cross-validation, from the following candidate values: $\{\frac{1}{10}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4, 5, 10\}$. We standardize each feature before estimation, as we found this improves convergence.

Results are provided in Tables WApp 1 and WApp 2, using the same sets of features and as in the main analysis. (Note that because γ^* may be different across versions, in-sample hit rate is not necessarily higher for a version that nests another version.)

Table WApp 1: Study 1 results. Pure content-based choice model estimated using LOG-Het.

Features	**Version 1**	**Version 2**	**Version 3**	**Version 4**	**Version 5**
Intercept	✓	✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
In-sample hit rate	62.21%	69.44%	73.69%	83.17%	81.11%
Out-of-sample hit rate	61.77%	66.32%	68.61%	71.30%	70.53%

Each column corresponds to one set of features. Each column is estimated separately using LOG-Het, i.e., preferences for the features included in the model are estimated at the individual level. Hit rates are averaged across consumers. All pairwise differences in in-sample or out-of-sample hit rates are statistically significant at $p<0.05$.

Table WApp 2: Study 2 results. Pure content-based choice model estimated using LOG-Het.

Features	**Version 1**	**Version 2**	**Version 3**	**Version 4**	**Version 5**
Intercept	✓	✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Sequel		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
Twitter Activity		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
In-sample hit rate	64.04%	72.79%	74.47%	78.54%	80.53%
Out-of-sample hit rate	63.59%	69.05%	69.64%	70.74%	70.74%

Each column corresponds to one set of features. Each column is estimated separately using LOG-Het, i.e., preferences for the features included in the model are estimated at the individual level. Hit rates are averaged across consumers. All pairwise differences in in-sample or out-of-sample hit rates are statistically significant at $p<0.05$, except the difference in out-of-sample hit rate between Version 4 and Version 5 ($p=0.96$).

D Using Movie Spoilers as Text Input

Table WApp 3: Descriptive statistics of movie descriptions (spoilers).

Statistic	Unit of analysis	Mean	St. dev.	Min	Max
Number of words (including "all other")	Movie descriptions (N=429)	2509.09	1303.20	202	7938
Number of occurrences of seed words	Movie descriptions (N=429)	109.09	56.96	8	398
Number of unique seed words	Movie descriptions (N=429)	62.29	27.08	4	191
Number of psychological themes with at least one seed word occurrence	Movie descriptions (N=429)	20.97	3.20	3	24
Total number of occurrences across movie descriptions	Seed words (N=2677)	17.48	58.28	0	888
Proportion of movie descriptions with at least one occurrence	Seed words (N=2677)	0.02	0.06	0	0.65
Total number of occurences across movie descriptions	Seed words with at least one occurrence (N=1662)	28.16	71.91	1	888
Proportion of movie descriptions with at least one occurrence	Seed words with at least one occurrence (N=1662)	0.04	0.07	0.002	0.65
Average number of seed word occurrences per movie description	Psychological Theme (N=24)	6.06	4.16	1.81	19.52
Proportion of movie descriptions with at least one seed word occurrence	Psychological Theme (N=24)	0.87	0.10	0.60	1.00

Table WApp 4: Guided LDA vs. Traditional LDA.

Number of topics per Psychological Theme (n)	Total number of topics	DIC for Guided LDA (*10^3)	DIC for Traditional LDA (*10^3)
1	25	2,992.5	3,050.9
2	49	2,604.9	2,685.3
3	73	2,394.2	2,478.7
4	97	2,251.6	2,316.0

Increasing the number of topics per psychological theme beyond 4 led to convergence issues when estimating viewers's preferences for topics. Therefore we stopped at $n = 4$. Traditional LDA is nested withing Guided LDA: it uses the same vocabulary but each topic has only a regular version, which may load on any word in the vocabulary.

Table WApp 5: Study 1 results. Pure content-based choice model.

Features	**Version 1**	**Version 2**	**Version 3**	**Version 4**	**Version 5**
Intercept	✓	✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
DIC	579.59	475.57	426.53	221.80	271.02
In-sample hit rate	62.09%	71.78%	76.30%	89.06%	85.52%
Out-of-sample hit rate	61.67%	66.44%	67.94%	71.40%	71.19%

Each column corresponds to one set of features. Each column is estimated separately using hierarchical Bayes, i.e., preferences for the features included in the model are estimated at the individual level. Hit rates are averaged across consumers. All pairwise differences in in-sample or out-of-sample hit rates are statistically significant at $p<0.05$, except the difference in out-of-sample hit rate between Version 4 and Version 5 ($p=0.44$).

Table WApp 6: Study 2 results. Pure content-based choice model.

Features	**Version 1**	**Version 2**	**Version 3**	**Version 4**	**Version 5**
Intercept	✓	✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Sequel		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
Twitter Activity		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
DIC	492.91	406.50	371.65	255.73	295.80
In-sample hit rate	64.05%	73.12%	76.54%	85.25%	82.20%
Out-of-sample hit rate	63.60%	68.91%	69.93%	70.92%	71.11%

Each column corresponds to one set of features. Each column is estimated separately using hierarchical Bayes, i.e., preferences for the features included in the model are estimated at the individual level. Hit rates are averaged across consumers. All pairwise differences in in-sample or out-of-sample hit rates are statistically significant at $p<0.05$, except the difference in out-of-sample hit rate between Version 4 and Version 5 ($p=0.44$).

Table WApp 7: Study 1 results. Content-Boosted Collaborative Filtering (CBCF).

Features	**Pure Collaborative Filtering**	**CBCF - Version 2**	**CBCF - Version 3**	**CBCF - Version 4**	**CBCF - Version 5**
Intercept		✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
Out-of-sample hit rate	68.67%	66.83%	68.05%	70.76%	70.62%

Each column corresponds to one set of features in the content-based predictions. For example, the predictions of CBCF in the second column combine the predictions from Version 1 of the content-based model with Collaborative Filtering. Hit rates are averaged across consumers. All pairwise differences in out-of-sample hit rates are statistically significant at $p<0.05$, except the difference between CBCF-Version 4 and CBCF-Version 5 ($p=0.49$).

Table WApp 8: Study 2 results. Content-Boosted Collaborative Filtering (CBCF).

Features	**Pure Collaborative Filtering**	**CBCF - Version 2**	**CBCF - Version 3**	**CBCF - Version 4**	**CBCF - Version 5**
Intercept		✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Sequel		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
Twitter Activity		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
Out-of-sample hit rate	68.27%	68.66%	69.47%	70.26%	70.42%

Each column corresponds to one set of features in the content-based predictions. For example, the predictions of CBCF in the second column combine the predictions from Version 2 of the content-based model with Collaborative Filtering. Hit rates are averaged across consumers. All pairwise differences are statistically significant at $p<0.05$, except the difference between Pure Collaborative Filtering and CBCF-Version 2 ($p=0.10$), and between CBCF-Version 4 and CBCF-Version 5 ($p=0.44$).

E Using Movie Scripts as Text Input

Table WApp 9: Descriptive statistics of movie descriptions (scripts).

Statistic	Unit of analysis	Mean	St. dev.	Min	Max
Number of words (including "all other")	Movie scripts (N=148)	23,327.01	5,517.47	8,224	43,475
Number of occurrences of seed words	Movie script (N=148)	956.01	262.31	343	1,541
Number of unique seed words	Movie scripts (N=148)	295.91	65.42	139	478
Number of psychological themes with at least one seed word occurrence	Movie scripts (N=148)	24	0	24	24
Total number of occurrences across movie scripts	Seed words (N=2,677)	52.85	185.38	0	4,309
Proportion of movie scripts with at least one occurrence	Seed words (N=2,677)	0.11	0.19	0.00	1.00
Total number of occurences across movie scripts	Seed words with at least one occurrence (N=2,004)	70.60	211.33	1	4,309
Proportion of movie scripts with at least one occurrence	Seed words with at least one occurrence (N=2,004)	0.15	0.20	0.01	1.00
Average number of seed word occurrences per movie script	Psychological Theme (N=24)	51.94	25.10	18.49	109.51
Proportion of movie scripts with at least one seed word occurrence	Psychological Theme (N=24)	1	0.00	0.00	1.00

Table WApp 10: Guided LDA vs. Traditional LDA.

Number of topics per Psychological Theme (n)	Total number of topics	DIC for Guided LDA (*10^3)	DIC for Traditional LDA (*10^3)
1	25	6,385.1	6,585.3
2	49	5,764.2	5,818.5
3	73	5,249.3	5,428.0
4	97	5,064.0	5,084.6

Increasing the number of topics per psychological theme beyond 4 led to convergence issues when estimating viewers's preferences for topics. Therefore we stopped at $n = 4$. Traditional LDA is nested withing Guided LDA: it uses the same vocabulary but each topic has only a regular version, which may load on any word in the vocabulary.

Scripts were not available for all movies in our corpus. We train Guided LDA on the set of scripts available, which provides estimates of the topic descriptions, $\{\phi_r^k, \phi_r^s, \pi_k\}$. When constructing Guided LDA features, we estimate θ_d based on the topic descriptions from Guided LDA (trained on movie scripts), using the text of the synopses as input.

Table WApp 11: Study 1 results. Pure content-based choice model.

Features	**Version 1**	**Version 2**	**Version 3**	**Version 4**	**Version 5**
Intercept	✓	✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
DIC	579.59	475.57	426.53	258.07	308.68
In-sample hit rate	62.09%	71.78%	76.30%	87.26%	84.00%
Out-of-sample hit rate	61.67%	66.44%	67.94%	70.18%	70.36%

Each column corresponds to one set of features. Each column is estimated separately using hierarchical Bayes, i.e., preferences for the features included in the model are estimated at the individual level. Hit rates are averaged across consumers. All pairwise differences in in-sample or out-of-sample hit rates are statistically significant at $p<0.05$, except the difference in out-of-sample hit rate between Version 4 and Version 5 ($p=0.53$).

Table WApp 12: Study 2 results. Pure content-based choice model.

Features	**Version 1**	**Version 2**	**Version 3**	**Version 4**	**Version 5**
Intercept	✓	✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Sequel		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
Twitter Activity		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
DIC	492.91	406.50	371.65	265.09	303.44
In-sample hit rate	64.05%	73.12%	76.54%	84.49%	81.59%
Out-of-sample hit rate	63.60%	68.91%	69.93%	70.60%	70.63%

Each column corresponds to one set of features. Each column is estimated separately using hierarchical Bayes, i.e., preferences for the features included in the model are estimated at the individual level. Hit rates are averaged across consumers. All pairwise differences in in-sample or out-of-sample hit rates are statistically significant at $p<0.05$, except the difference in out-of-sample hit rate between Version 4 and Version 5 ($p=0.91$).

Table WApp 13: Study 1 results. Content-Boosted Collaborative Filtering (CBCF).

Features	**Pure Collaborative Filtering**	**CBCF - Version 2**	**CBCF - Version 3**	**CBCF - Version 4**	**CBCF - Version 5**
Intercept		✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
Out-of-sample hit rate	68.67%	66.83%	68.05%	69.82%	69.96%

Each column corresponds to one set of features in the content-based predictions. For example, the predictions of CBCF in the second column combine the predictions from Version 1 of the content-based model with Collaborative Filtering. Hit rates are averaged across consumers. All pairwise differences in out-of-sample hit rates are statistically significant at $p<0.05$, except the difference between CBCF-Version 4 and CBCF-Version 5 ($p=0.55$).

Table WApp 14: Study 2 results. Content-Boosted Collaborative Filtering (CBCF).

Features	**Pure Collaborative Filtering**	**CBCF - Version 2**	**CBCF - Version 3**	**CBCF - Version 4**	**CBCF - Version 5**
Intercept		✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Sequel		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
Twitter Activity		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
Out-of-sample hit rate	68.27%	68.66%	69.47%	70.02%	70.03%

Each column corresponds to one set of features in the content-based predictions. For example, the predictions of CBCF in the second column combine the predictions from Version 2 of the content-based model with Collaborative Filtering. Hit rates are averaged across consumers. All pairwise differences are statistically significant at $p<0.05$, except the difference between Pure Collaborative Filtering and CBCF-Version 2 ($p=0.10$), and between CBCF-Version 4 and CBCF-Version 5 ($p=0.95$).

F True Positive and True Negative Rates

The true positive rate (respectively, true negative rate) for a consumer is defined as the proportion of movies the consumer actually watched (respectively, did not watch), among the movies that the model predicted would be watched (respectively, not watched), i.e., the fitted choice probability was greater than or equal to 0.5 (respectively, lower than 0.5). True positive and true negative rates are averaged across consumers. For each metric (e.g., in-sample true positive rate), results are reported for consumers for whom that metric is defined for all benchmarks (e.g., in the case of in-sample true positive rate, there is at least one positive prediction in each benchmark). This ensures that the underlying sample of consumers is the same across benchmarks within each metric. In all tables, all pairwise differences within a metric are statistically significant at $p<0.05$ in-sample. Given there are only five out-of-sample observations per consumer, there exist several consumers for whom there is no positive or no negative prediction out of sample, i.e., the true positive rate or the true negative rate is undefined. As a result, samples sizes are reduced and many pairwise

comparisons are not statistically significant out-of-sample.

Table WApp 15: Study 1 results. Pure content-based choice model. Using synopses as text input for Guided LDA.

Features	**Version 1**	**Version 2**	**Version 3**	**Version 4**	**Version 5**
Intercept	✓	✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
In-sample true positive rate	66.12%	80.43%	84.12%	97.18%	94.63%
In-sample true negative rate	73.26%	82.35%	86.26%	97.34%	95.30%
Out-of-sample true positive rate	63.56%	74.32%	73.78%	75.92%	75.53%
Out-of-sample true negative rate	74.20%	78.33%	78.82%	80.64%	81.13%

Table WApp 16: Study 2 results. Pure content-based choice model. Using synopses as text input for Guided LDA.

Features	**Version 1**	**Version 2**	**Version 3**	**Version 4**	**Version 5**
Intercept	✓	✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Sequel		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
Twitter Activity		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
In-sample true positive rate	66.57%	80.03%	83.46%	95.01%	92.27%
In-sample true negative rate	75.38%	83.62%	86.65%	95.95%	93.44%
Out-of-sample true positive rate	63.13%	68.61%	72.27%	74.76%	71.87%
Out-of-sample true negative rate	76.10%	80.88%	81.90%	82.47%	82.63%

Table WApp 17: Study 1 results. Pure content-based choice model. Using spoilers as text input for Guided LDA.

Features	**Version 1**	**Version 2**	**Version 3**	**Version 4**	**Version 5**
Intercept	✓	✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
In-sample true positive rate	66.12%	80.43%	84.12%	97.47%	94.88%
In-sample true negative rate	73.26%	82.35%	86.26%	97.62%	95.09%
Out-of-sample true positive rate	63.56%	73.92%	73.38%	76.99%	76.39%
Out-of-sample true negative rate	74.05%	78.22%	78.70%	81.43%	80.65%

Table WApp 18: Study 2 results. Pure content-based choice model. Using spoilers as text input for Guided LDA.

Features	**Version 1**	**Version 2**	**Version 3**	**Version 4**	**Version 5**
Intercept	✓	✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Sequel		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
Twitter Activity		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
In-sample true positive rate	66.57%	80.03%	83.46%	92.37%	90.53%
In-sample true negative rate	75.38%	83.62%	86.65%	95.05%	92.29%
Out-of-sample true positive rate	63.10%	68.99%	72.60%	71.35%	74.05%
Out-of-sample true negative rate	76.14%	80.95%	81.94%	82.30%	82.29%

Table WApp 19: Study 1 results. Pure content-based choice model. Using scripts as text input for Guided LDA.

Features	**Version 1**	**Version 2**	**Version 3**	**Version 4**	**Version 5**
Intercept	✓	✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
In-sample true positive rate	66.12%	80.43%	84.12%	95.72%	92.83%
In-sample true negative rate	73.26%	82.35%	86.26%	96.53%	94.56%
Out-of-sample true positive rate	63.33%	73.92%	73.38%	75.62%	76.11%
Out-of-sample true negative rate	74.11%	78.26%	78.68%	79.36%	80.29%

Table WApp 20: Study 2 results. Pure content-based choice model. Using scripts as text input for Guided LDA.

Features	**Version 1**	**Version 2**	**Version 3**	**Version 4**	**Version 5**
Intercept	✓	✓	✓	✓	✓
Average Critic Rating		✓	✓	✓	✓
Average User Score		✓	✓	✓	✓
Production Budget		✓	✓	✓	✓
Widest Release		✓	✓	✓	✓
Widest Release^2		✓	✓	✓	✓
Domestic Box Office		✓	✓	✓	✓
MPAA Rating		✓	✓	✓	✓
Run Time		✓	✓	✓	✓
Sequel		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star Power		✓	✓	✓	✓
Twitter Activity		✓	✓	✓	✓
DVD Release Timing		✓	✓	✓	✓
DVD Sales Rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-Words variables from LSA			✓		
Guided LDA topic weights				✓	✓
In-sample true positive rate	66.57%	80.03%	83.46%	92.65%	89.21%
In-sample true negative rate	75.38%	83.62%	86.65%	94.68%	91.86%
Out-of-sample true positive rate	63.37%	68.61%	72.67%	73.49%	71.02%
Out-of-sample true negative rate	75.94%	80.74%	81.52%	82.11%	81.51%

References

Evgeniou, Theodoros, Massimiliano Pontil, Olivier Toubia. 2007. A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. *Marketing Science* **26**(6) 805–818.

Griffiths, Thomas L., Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Science* **101** 5228–5235.

Jagarlamudi, Jagadeesh, Hal III Daum, Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* .