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[§]

Abstract

The authors propose a quantitative approach for describing entertainment products, in a way that allows improving the predictive performance of consumer choice models for these products. Their approach is based on the media psychology literature, which suggests that the consumption of entertainment products by individuals is influenced by the psychological themes featured in these products. They classify psychological themes based on the "Character Strengths" taxonomy from the positive psychology literature (Peterson and Seligman, 2004). They develop a natural language processing tool, Guided LDA, that automatically extracts a set of features of entertainment products based on their descriptions. Guided LDA is flexible enough to allow features to be informed by psychological themes, while allowing other relevant dimensions to emerge. They apply this tool to movies. They show that Guided LDA features help better predict movie-watching behavior at the individual level. They find this result both with award-winning movies and blockbuster movies. They illustrate the potential of the proposed approach in pure content-based predictive models of consumer behavior, as well as in hybrid predictive models that combine content-based models with collaborative filtering. They also show that Guided LDA can improve the performance of models that predict aggregate outcomes.

[*]Glaubinger Professor of Business, Graduate School of Business, Columbia University, ot2107@gsb.columbia.edu.

[†]Professor, Industrial Engineering and Operations Research Department, Columbia University, garud@ieor.columbia.edu.

[‡]Real.org; Real Engagement and Loyalty (REAL); OWEN.AI, rb@real.org

[§]Graduate Student, Graduate School of Business, Columbia University, ALemaire18@gsb.columbia.edu.

The revenue of the global entertainment and media industry was estimated at $1.8 trillion in 2016 (Statista, 2017b). One important trend in this industry is the increasing use of digital services such as streaming, video-on-demand, e-readers, etc. For example, the over-the-top streaming market (including Netflix, Hulu, etc.) in the US alone is expected to grow from $4.67B in 2013 to $12.64B in 2019 (Statista, 2016a), and it is predicted that 28% of the US population will own an e-Reader by 2020 (Statista, 2017a). Importantly for marketers, these technologies increase the availability of *panel* data in which consumers are observed making decisions over time. In addition, many brick-and-mortar distributors of entertainment products now offer loyalty programs to their customers (e.g., Regal Crown Club, AMC Stubs Card, Cinemarks' CineMode for movies, B&N Membership for books, etc.), which provide panel data of a similar nature.

Approaches for leveraging panel data in the media and entertainment industry have been classified into three categories (Adomavicius and Tuzhilin, 2005): pure collaborative approaches, where the behavior of a user is predicted based on the past behavior of similar users; content-based approaches, where the behavior of a user is predicted based on their own past behavior; and hybrid methods, which combine collaborative and content-based methods. Popular collaborative approaches include variants of neighborhood-based Collaborative Filtering (Breese et al., 1998; Linden et al., 2003) and latent factor models (Koren et al., 2009). Content-based approaches often use various types of regressions, decision trees, or neural networks to link product features to consumption. Popular hybrid approaches include Content-Boosted Collaborative Filtering (Melville et al., 2002),and the Bayesian approach proposed by Ansari et al. (2000). In the marketing literature, most consumer choice models for entertainment products have been content-based or hybrid (e.g., Ansari et al., 2000; Bo-

dapati, 2008; Eliashberg and Sawhney, 1994; Rust and Alpert, 1984; Shachar and Emerson, 2000; Ying et al., 2006).

Content-based and hybrid approaches rely on estimating a set of weights on a pre-existing set of product *features*. As such, the performance of content-based and hybrid methods, including the ones developed in the marketing literature, is a direct function of the quality and relevance of these features. Relevant features are easy to generate for many types of products and services outside of the media and entertainment industry. For example, a digital camera may be defined based on its memory, shutter speed, size, brand, price, etc.

However, when it comes to *entertainment* products, defining a feature set is not as straightforward. The most common features of entertainment products used in content-based and hybrid approaches are *genres*. For example, Eliashberg and Sawhney (1994) and Möller and Karppinen (1983) use the liking for different genres of movies as a predictor for movie enjoyment. Ansari et al. (2000) and Ying et al. (2006) include genres in their hybrid recommendation models. Rust and Alpert (1984) and Shachar and Emerson (2000) include genres in their models of television viewing behavior. In the news industry, Chu et al. (2009) quantify Yahoo! users' preferences over (manually coded) types of news article, using an approach inspired by conjoint analysis.

Despite their convenience, genre classifications suffer from some limitations. First, genres tend to be category-specific. For example, the International Movie Database classification (www.imdb.com) contains 22 genres such as "action," "comedy," etc. However, these genres are not necessarily completely relevant in other industries such as books. A taxonomy that would be relevant across categories would allow merging data from the same consumers across categories. Second and perhaps more importantly, there appears to be consensus today in

the industry that traditional genre classifications are not enough to describe entertainment products with adequate granularity and richness. For example, Netflix developed its own proprietary system of over 76,000 genres or "tags" (Madrigal, 2014). Examples include "spy action & adventure movies from the 1930s" and "time travel movies starring William Hartnell." This approach is not only prohibitively costly (Netflix has been reported to hire human raters to tag content), it embodies the traditional tradeoff between fit and complexity in data analytics. Complex models tend to fit better in-sample, but their interpretation may be less obvious, and their out-of-sample fit may not be as high.[1,2]

The difficulty of generating a set of features of entertainment products that are both relevant and parsimonious may partly explain why many practitioners have favored collaborative methods over content-based or hybrid methods when working with individual-level panel data. In particular, Collaborative Filtering is a very popular approach for predicting consumption at the individual level and recommending entertainment products to consumers (Koren et al., 2009; Linden et al., 2003). Collaborative filtering approaches offer the benefit of not requiring a set of explicit features describing products. However, this comes with at least two main limitations (Su and Khoshgoftaar, 2009). First, it becomes challenging to develop insights and reach interpretable results in the absence of a set of features that predict consumer choices. Second, collaborative approaches suffer from the "new item cold start" problem. That is, if products cannot be defined by a common set of features, every

[1]Note that efforts to develop a comprehensive taxonomy of genres have also been pursued in the public domain, also leading to complex systems. For example, the Library of Congress started a project in 2007 to "develop a dynamic, multi-disciplinary body of genre/form terms that is cohesive, unified, intuitive, and user-friendly." As of January 2015, the genre/form list contained 847 terms, and new terms continue to be added regularly (Library of Congress, 2015).

[2]Inspired by the Information Retrieval literature, some researchers (e.g., Mooney and Roy, 2000) have developed content-based models that use individual words as features (e.g., whether a book contains a given word). However, this approach also leads to a very large set of features, leading to a curse of dimensionality (Adomavicius and Tuzhilin, 2005).

product is "unique" and it becomes challenging to make predictions for new products for which little or no consumption data are available. In contrast, content-based and hybrid approaches allow making predictions for new products, based on consumers' preferences for the features that describe the content of these products.

In sum, a wide range of collaborative, content-based, and hybrid approaches have been proposed over the years to leverage individual-level panel data in the entertainment and media industry. While much effort has been spent on developing new and more powerful methods, less effort has been spent developing better *input* for content-based and hybrid methods, i.e., sets of features that are objectively defined, predictive of consumers' decisions, and not excessively complex. With higher quality input, the content-based and hybrid methods developed in the marketing literature and elsewhere might have the potential to gain even more popularity among practitioners.

In this paper, we propose a new way of describing entertainment products. Our objective is *not* to develop new methods that predict consumer choices conditional on a set of features, but rather to develop a new method for constructing the set of features, which can be used as input into any existing content-based or hybrid model that attempts to predict the behavior of consumers based on past behavior. Our taxonomy is inspired by the psychology behind the consumption of entertainment products. The starting point of our theoretical development is the media psychology literature, which suggests that a consumer's preferences for an entertainment product are driven at least in part by the alignment of their psychological profile with the psychological themes featured in the product. Accordingly, we construct features that have the ability to reflect the psychological themes in entertainment products. We borrow from the positive psychology literature and use Peterson and

Seligman (2004)'s taxonomy of psychological themes. We adapt the approach proposed by Jagarlamudi et al. (2012) to develop a Natural Language Processing (NLP) tool, Guided LDA, that automatically extracts features of entertainment products based on their descriptions. Descriptions of entertainment products are generally publicly available, in the form of synopses, summaries, etc. Our Guided LDA approach is flexible enough to allow features to be informed and guided by psychological themes, while allowing other relevant dimensions to emerge from the descriptions. We apply this tool to a dataset of 429 movies. The output is a set of features describing each product, to be included in content-based or hybrid predictive models of consumer behavior. In two online studies, we show that Guided LDA features improve our ability to predict movie consumption at the individual level, above and beyond standard features such as genres. We find this result both with award-winning movies and blockbuster movies. We illustrate the potential of Guided LDA both in a pure content-based model (hierarchical Bayes logistic regression - see also Web Appendix C for a machine learning-based approach) and in a hybrid model (Content-Boosted Collaborative Filtering). Although Guided LDA was developed primarily to produce input for models that predict behavior at the individual level, we also illustrate its use with models that predict aggregate outcomes such as box office performance or return on investment.

Our contribution is both substantive and managerial. Substantively, past empirical research in marketing and related fields has shed much light on the link between the consumption of entertainment products and social factors such as word of mouth (d'Astous and Touil, 1999; Dellarocas et al., 2007; Duana et al., 2008; Liu, 2006) or joint decision making (De Silva, 1998), and individual differences in demographic or personality variables (Austin, 1986; Cuadrado and Frasquet, 1999; De Silva, 1998; Eliashberg and Sawhney, 1994). How-

ever, the literature has not focused as much on providing a rich, theory-driven taxonomy of entertainment products that would allow predicting individual-level behavior, and that would be generalizable across categories.

Managerially, because our Guided LDA method is automated and scalable, Guided LDA features may be used as input into any existing content-based or hybrid "big data" analytics tools, including the ones developed in the marketing literature. Based on our encouraging empirical results, we hope that our research will increase the adoption of these methods among practitioners in the media and entertainment industry. In addition, our research makes marketing models that were initially developed for other industries (e.g., based on conjoint analysis or scanner data), more relevant and applicable to the media and entertainment industry.

The rest of the paper is organized as follows. We first present our theoretical argument. Next, we introduce our Guided LDA approach, and apply it to a movie dataset. Next, we report on two studies that explore the value of the proposed approach in predicting movie consumption at the individual level. Finally,we explore the use of Guided LDA features as input into aggregate predictive models of performance.

Relevant Literatures

Media Psychology

Media psychology is a sizable subfield of psychology which studies how people perceive, interpret, respond, and interact with media. This literature suggests that people prefer entertainment products that satisfy psychological needs. For example, Rentfrow et al. (2011)

argue that "people seek out entertainment that reflects and reinforces aspects of their personalities" (p. 251), and find that preferences for an entertainment product are driven at least in part by the alignment of the consumer's psychological profile with the psychological themes featured in the product. These psychological themes are reflected in entertainment products by the characters in a story, the setting of the story, the type of challenges faced by characters, etc.

Empirically, the traditional approach for exploring the psychology of media consumption in this literature has been to use a survey to measure both the psychological profiles of a sample of consumers and their preferences for different genres of entertainment products, and then explore the link between the two sets of variables. For example, Weaver (1991, 2003) found that viewers who score high on neuroticism have less preference for adventure movies, while viewers who score high on psychoticism express stronger preferences for horror movies. Kraaykamp and Van Eijck (2005) linked the Big Five personality factors (McCrae and Costa, 1999) of a sample of Dutch consumers to their media preferences. They found for example that people who scored higher on "Openness to Experiences" had stronger preferences for cultural programs but weaker preferences for soap programs. Rentfrow and Gosling (2003) found similar types of correlation between consumers' psychological profiles and preferences, in the domain of music.

The approach of measuring consumers' psychological profiles and linking them to their media preferences has been useful in demonstrating that psychological factors are important predictors of media preferences. However, this approach is not scalable as it requires surveying all consumers under consideration. In contrast, in this paper we focus on describing the entertainment *products* themselves based on the psychological themes they feature, rather

than describing *consumers* based on their own psychological profiles. In particular, we develop an approach for weighing entertainment products along relevant dimensions inspired by the positive psychology literature, producing features to be incorporated into models that learn consumers' preferences through their behavior, without explicitly measuring consumers' psychological profiles.

Positive Psychology

The media psychology literature suggests a link between the psychological themes featured in an entertainment product and preferences for that product. This raises the question of how psychological themes may be described and classified. Several of the media psychology studies reviewed above used the Big Five personality dimensions (McCrae and Costa, 1999) as a taxonomy of consumers' psychological traits.

We adopt instead a taxonomy of psychological themes based on the positive psychology literature. Positive psychology is the branch of psychology that focuses on the achievement of a satisfactory life (see Seligman and Csikszentmihalyi, 2000; Seligman et al., 2005, for an introduction to positive psychology). Positive psychology has become a major subfield of psychology in the recent years, but its applications in the marketing literature have been rare. A significant milestone in the advent of positive psychology was the *Character Strengths and Virtues Handbook* by Peterson and Seligman (2004), which identified and classified 24 psychological themes, labeled "Character Strengths." These "Character Strengths" include Bravery, Integrity, Citizenship, Humility, Prudence, Gratitude, and Hope. A complete list is provided in Table 1.

<INSERT TABLE 1 ABOUT HERE>

Our choice to base our taxonomy on the positive psychology literature was driven by two main factors. First, with 24 dimensions (vs. 5 for example in the Big Five framework), this framework is fairly granular and appears likely to allow subtle distinctions between entertainment products. Second, the positive psychology literature has had a strong focus on various ways to achieve life satisfaction through pleasure, meaning and engagement (Peterson et al., 2005b; Seligman et al., 2005). Therefore, adopting this framework opens the door for future research that would explore and exploit the link between the consumption of entertainment products and life satisfaction. For example, one could envision recommendation engines that would take the user's well-being into consideration. In this particular paper, we do not make any claims related to the link between the consumption of entertainment products and life satisfaction, and leave such endeavor to future research.

It is important to note that the term "Character" in "Character Strengths" is *unrelated* to the concept of characters (i.e., protagonists) in entertainment products. In other words, the term "character" has different meanings in the media literature (where it refers to one of the protagonists in a story) and in the positive psychology literature (where "Character Strengths" refer to psychological themes). In this paper we use "psychological themes" and "Character Strengths" interchangeably, and favor the former in an attempt to reduce confusion. As noted above, entertainment products reflect psychological themes not only through their characters, but also through the challenges faced by characters, the setting, etc.

It is also important to note that the definition of each "Character Strength" (as provided by Peterson and Seligman, 2004) is sometimes broader than the common English definition of the term used to label it. For example, "Citizenship" includes social responsibility, loyalty,

and teamwork, and is defined as "identification with and sense of obligation to a common good that includes the self but that stretches beyond one's own self-interest" (Peterson and Seligman, 2004, Page 370).

Clinical psychologists have previously attempted to establish connections between the positive psychology literature and the media and entertainment literature. Niemiec and Wedding (2014) show how movies may be used to study, teach and practice positive psychology. Their book is targeted toward educators and practitioners of positive psychology who are interested in using movies as a vehicle for teaching and practicing positive psychology. As a result, their work is purely qualitative and descriptive in nature. In particular, these authors manually identified movies that illustrate each "Character Strength." In contrast, our target audience is modelers interested in applying content-based or hybrid predictive models of consumer behavior to the media and entertainment industry. As a result, our work is much more quantitative in nature. In contrast to Niemiec and Wedding (2014), we focus on the consumption of entertainment products, and we propose a scalable tool for automatically classifying products, without relying on human input.

Screenwriting

The screenwriting literature (e.g., Blacker, 1988; Field, 2007; Hauge, 2011; McKee, 1997) has identified factors that describe movies and influence the quality of a script. This literature is more prescriptive in nature, and many of these factors may be viewed as reflecting "best practices" in screenwriting. Eliashberg et al. (2007, 2014) integrate and summarize this literature to construct a set of criteria that capture "how a story should be told and what kind of stories would resonate with audience" (Eliashberg et al., 2007, Page 884). For

example, the story should follow a logical, causal relationship, each scene description should advance the plot and be closely connected to the central conflict, etc. As shown by Eliashberg et al. (2007, 2014), these factors are good predictors of a movie's *aggregate* box office performance. In contrast, our primary focus in this paper is on predicting *individual-level* behavior captured by panel data. Accordingly, the features we extract from entertainment products are meant to reflect "horizontal" rather than "vertical" differentiation, i.e., they are meant to reflect differences in consumer tastes rather than differences in the overall quality of a story. For example, we are more concerned with predicting which movie will appeal to which consumers (controlling for each movie's overall appeal), rather than with predicting the aggregate number of consumers to which a movie will appeal. For completeness, we include variables developed by Eliashberg et al. (2007, 2014) both in our individual-level and aggregate analyses.

We also note that some authors in the screenwriting literature have discussed aspects of stories that reflect horizontal rather than vertical differentiation across movies, and that these aspects are not inconsistent with the idea that movies may be described based on the psychological themes that they feature. For example, McKee (1997) introduces the concept of "story values," which he describes as "the soul of storytelling" and defines as "the universal qualities of human experience"(p. 34). One may argue that our taxonomy of psychological themes captures these values at least to some extent. Indeed, examples of story values provided by McKee (1997) include love/hate, courage, cowardice, loyalty/betrayal, wisdom/stupidity, which may all be linked to some of the psychological themes in our taxonomy. Similarly, Hauge (2011) defines the *theme* of a movie as "the universal statement the screenplay makes about the human condition" (Page 82). Hauge gives the example of

the movie "Wedding Crashers," and argues that "Beyond the hilarity, clever plot, terrific dialogue, and sexual shenanigans, the theme of *Wedding Crashers* ... speaks to the need for honesty and emotional risk" (Page 82). "Integrity" is indeed one of the psychological themes in our taxonomy.

Natural Language Processing

We have hypothesized that the consumption of entertainment products may be linked to the psychological themes featured in these products, and that the positive psychology literature provides a useful taxonomy of psychological themes. At the same time, we acknowledge that other factors may help predict choices. Accordingly, we develop an approach that is flexible enough to allow features to be informed and guided by our taxonomy of psychological themes, while allowing other relevant dimensions to emerge. This approach is based on the Natural Language Processing (NLP) literature.

Given the changing nature of the types of data collected in many marketing contexts, NLP has become increasingly relevant to the marketing literature. Many of the marketing applications to date have focused on the analysis of user-generated content (e.g. Archak et al., 2011; Ghose et al., 2012; Lee and Bradlow, 2011; Netzer et al., 2012; Tirunillai and Tellis, 2014). Applications of NLP to the entertainment industry include Eliashberg et al. (2007, 2014). These authors use Latent Semantic Analysis (LSA) (Deerwester et al., 1990) to characterize the text of movie descriptions. They show that the box office performance of a movie may be predicted based on variables coming from analyzing the textual description of the movie, combined with other types of input. As noted above, our primary focus here is on *individual-level* behavior captured by panel data, rather than aggregate outcomes.

One additional key difference between Eliashberg et al. (2007, 2014)'s work and the present paper is that our NLP analysis is grounded in the media psychology and positive psychology literatures, that is, the descriptors we consider are not only driven by data, but they are also informed by theory. In addition, while these authors use LSA, our Guided LDA approach is an extension of Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Blei, 2012; Tirunillai and Tellis, 2014). In this paper we empirically compare Guided LDA features to features obtained from LSA.

LDA is a Bayesian learning algorithm that extracts "topics" from text based on co-occurrence. It is a probabilistic version of LSA, thus enabling likelihood-based inference. Topics may be viewed as groups of words that are semantically related to each other, i.e., they tend to appear together in the corpus of text. A more detailed description of traditional LDA is presented in the next section. In a marketing context, Tirunillai and Tellis (2014) use LDA to identify dimensions of quality and valence expressed in online reviews. In Traditional LDA or in LSA, topics emerge strictly from the data and need to be labeled by the researcher, i.e., learning is unsupervised. The labeling of topics in Traditional LDA is similar to the labeling of components in Principal Component Analysis. In our context, topics should be informed by psychological themes. One approach would be to constrain each topic to reflect exactly one psychological theme, by constraining the vocabulary in each topic to consist of a set of words that are known to be associated with a particular psychological theme. However, such approach would not give the opportunity for other relevant topics to emerge. Indeed, while the literature suggests that preferences for entertainment products are linked to the psychological themes featured in these products, other factors are likely to help predict consumer choices. Hence, we use an approach that is flexible enough to allow the definition

of topics to be informed by theory, while allowing topics to emerge freely from the data and to capture other, unrelated constructs. In particular, our approach is based on the method proposed by Jagarlamudi et al. (2012).[3] This approach allows us to specify "seed words" that are believed to be representative of each psychological theme, based on the positive psychology literature.[4] Topics are guided by these seed words, i.e, learning is supervised; yet topics are at the same time allowed to deviate from seed words. We describe our NLP approach in the next section.

Guided LDA

Our Guided LDA approach takes the following input: a dictionary of seed words associated with psychological themes, and the textual description of a set of entertainment products. The two main outputs of the analysis are: (i) a set of topic-word distributions, i.e., each topic k is defined by a multinomial distribution over the words in the vocabulary; (ii) a set of document-topic distributions, i.e., each document d (that describes one entertainment product) is associated with a set of weights that capture a multinomial distribution over topics. These weights are meant to be used as features describing each product.

The simplest version of our Guided LDA approach has $K = 25$ topics: one topic per psychological theme, plus one topic that controls for the baseline occurrence of words (more details below). We also developed versions in which each psychological theme is assigned multiple topics. Indeed, the psychological themes defined by Peterson and Seligman (2004) tend to be quite broad and may have sub-themes. For example, there may be different sub-themes of Love (one may be related to friendship, one to romantic relationships, etc.). If each

[3] Our Guided LDA is based on Model 1 in Jagarlamudi et al. (2012).

[4] Tirunillai and Tellis (2014) used seed words to measure the valence of online reviews.

psychological theme is assigned n topics, then the total number of topics is $K = 24 * n + 1$ (n topics per psychological theme plus the baseline topic).

As explained further below, each topic itself has two versions: a "seeded" version that is constrained to map onto a set of seed words associated with the corresponding psychological theme, and a "regular" version that is unconstrained and has positive weights on all the words in the dictionary.

Compiling the set of seed words

In order to add supervision to the LDA learning process, we define a set of seed words associated with each psychological theme. An initial set of seed words were obtained from the descriptions of the "Character Strengths" in Peterson and Seligman (2004) and Seligman et al. (2005), as well as the scales developed by Peterson et al. (2005a). Our seed words come from all parts of speech (mainly nouns, verbs and adjectives) and include single words as well as short phrases (e.g., "look forward" for Hope). For simplicity we refer to all these entries as "seed words." Because we do not stem words (stemming is not always performed in topic modeling, e.g., Jagarlamudi et al., 2012), seed words include both singular and plural nouns, as well as different conjugations of the same verbs.

In order to augment our initial list of seed words, we asked Amazon Mechanical Turk participants to suggest additional words associated with each topic. We selected ten common seed words for each of the 24 psychological themes, based on our preliminary analysis. For each theme, we showed participants the ten seed words and asked them to propose three new words that would complement the list well. We received complete responses from $N = 106$ respondents, who were screened for being based in the US and who were each paid \$1. We

went through the list of words suggested by participants manually to identify new seed words. These respondents were not invited to our main studies, and they had no input other than proposing new seed words (e.g., they were *not* asked to rate any movie on any dimension). Finally, a media psychologist with expertise in positive psychology reviewed our list of seed words and suggested additions and edits.

Our final dictionary of seed words contains 2,677 unique seed words. These seed words reflect a variety of vehicles through which each psychological theme may be featured in an entertainment product. For example, some seed words relate primarily to characters (e.g., "artist" is a seed word for Creativity and for Appreciation of Beauty and Excellence, "patriotic" is a seed word for Bravery and for Citizenship), others relate to the setting (e.g., "school" is a seed word for Love of Learning), the problems faced by characters (e.g., "divorce" is a seed word for Love), the solutions to these problems (e.g., "reconciliation" is a seed word for Forgiveness and Mercy), etc. Each psychological theme has on average 136.33 seed words (standard deviation=22.09).

The sets of seed words may overlap between psychological themes. For example, "clever" is a seed word for Creativity, Open-Mindedness, Wisdom, and Social Intelligence. The average overlap between the sets of seed words corresponding to any two psychological themes is 2.92 (standard deviation=4.23), and 30.80% of all pairs of psychological themes have non-overlapping sets of seed words. As will become clear, our Guided LDA approach is able to handle seed words associated with multiple topics.

We note that our seed words may have either positive or negative valence. For example, "heaven" and "hell" are both seed words for Spirituality. We combined both types of seed words because the media psychology literature suggests that consumers may be attracted to

both positive or negative expressions of psychological themes (Ang, 1985; Cohen, 2001, 2006; Hoffner and Cantor, 1991). Therefore, our primary focus is identifying which psychological themes are featured in an entertainment product, not *how* they are featured.[5] Note that in the versions in which we assign more than one topic to each psychological theme, different sub-themes may load more heavily on positive vs. negative words. We leave a more detailed treatment of the valence of seed words to future research. To the extent that more information may be learned by distinguishing positive vs. negative seed words more specifically, the results presented in this paper present a lower bound of the potential of the proposed approach.

Creating the vocabulary

Our vocabulary contains a mix of seed words and other relevant words which are not seed words. In addition to the dictionary of seed words, we extract all words that appear at least 10 times in the corpus under consideration. We select our vocabulary among seed and non-seed words, using the standard term frequency-inverse document frequency (tf-idf) metric (Manning et al., 2008). For each word in the corpus, we compute the term frequency (tf) as the total number of occurrences of this word in the corpus, and the document frequency (df) as the number of documents (e.g., movie synopses) in which the word appears at least once. Term frequency - inverse document frequency (tf-idf) is then defined as $tf \times log(N/df)$, where N is the number of documents in the corpus. Following standard practice, we construct our vocabulary by selecting the words with the highest tf-idf. In particular, we keep the 2,000

[5]For example, a movie like "The Hangover" may be found to have a large weight on a topic related to the theme of Prudence, because prudence is a dominant theme in the movie, expressed by the *lack* of prudence shown by the movie's protagonists. A large weight on such topic may result from the presence of negative seed words such as "careless," "accident," or "danger" in the movie description.

words with the highest $tf\text{-}idf$. We apply the same screening criterion, i.e., the same $tf\text{-}idf$ threshold, to both seed and non-seed words. Finally, we complete our vocabulary by adding one "all other" word that captures any word that appears in any document but that is not in the vocabulary. This "word" allows us to control for the length of the documents. More details are provided in the next section.

Guided LDA Specification

We assume that each document in the corpus has been tokenized, i.e., broken down into individual words or phrases (tokens). Tokens represent the smallest unit of observation in our data, i.e., a document is represented as a collection of tokens. Each token may be thought of as a "slot" in the document that is "filled" with a word. Traditional LDA (Blei et al., 2003; Blei, 2012; Tirunillai and Tellis, 2014) assumes the text corpus comes from the following data generating process. First, each token in each document is independently assigned to a topic according to a multinomial distribution that captures how topics are distributed within that document. Second, the token is assigned to a particular word according to another multinomial distribution that captures how words are distributed within that topic. The assignment of tokens to topics is captured by a set of latent variables.

Mathematically, we index documents (where each document describes an entertainment product) by $d = 1, ...D$, topics by $k = 1, ...K$, and words in the vocabulary by $w = 1, ...W$, where the last word W is the "all other" word. For each topic k, we define ϕ_k as a $1 * W$ vector that we estimate, that contains the topic-word set of probability weights for topic k, that is, the probability that a token is assigned to each word given that it is assigned to topic k. For each document d, we define θ_d as a $1 * K$ vector that we estimate, that contains

the document-topic set of probability weights for document d, that is, the probability that a token is assigned to each topic given that it is in document d. These weights may be used as product features in content-based or hybrid consumer choice models. The i^{th} token in document d belongs to topic $z_i^d \in \{1, ...K\}$. The variable z_i^d is an unobserved, latent variable, which is also estimated. We denote by $w_i^d \in \{1, ...W\}$ the index of the word associated with the i^{th} token in document d.

Guided LDA nests Traditional LDA by allowing each topic to have two versions: a "regular" version defined as in Traditional LDA that has positive weights on all words in the dictionary (seed and non-seed), and a "seeded" version that has positive weights only on the seed words for the corresponding psychological theme. The seeded version ensures that topics are guided by seed words, while the regular version allows other relevant dimensions to emerge.

We denote by l_k^s the $1 * W$ vector of binary variables that capture the set of seed words on which the seed version of topic k is allowed to have positive weights, where $l_k^s(w) = 1$ if and only if word w is a seed word for topic k. The regular version of the topic is allowed to have positive weights on all words (except the "all other" word): $l_k^r(w) = 1$ for all $w < W$. The data generating process assumed by Guided LDA is as follows, where 1_K is a vector of 1's:

1. For each topic $k = 1, ...K$,

    - Draw regular topic: $\phi_k^r \sim Dirichlet(\alpha_1 l_k^r)$
    - Draw seed topic: $\phi_k^s \sim Dirichlet(\alpha_1 l_k^s)$
    - Draw weight on seeded topic: $\pi_k \sim Beta(1, 1)$

2. For each document $d = 1, ...D$,

   - Draw topic distribution: $\theta_d \sim Dirichlet(\alpha_2 1_K)$
   - For each token i:
     - Draw a topic: $z_i^d \sim Multinomial(\theta_d)$
     - Draw an indicator: $x_i^d \sim Binomial(\pi_{z_i^d})$
     - If indicator $x_i = 0$, draw a word from regular topic: $w_i^d \sim Multinomial(\phi^r_{z_i^d})$
     - If indicator $x_i = 1$, draw a word from seeded topic: $w_i^d \sim Multinomial(\phi^s_{z_i^d})$

Each topic is a mixture between a seeded topic and a regular topic. For each topic, the difference between the regular and seed versions lies in the supports l_k^s and l_k^r. The seeded version of the topic is allowed to have positive weights only on the corresponding seed words, while the regular version is allowed to have positive weights on all words (except the "all other" word). When $n > 1$ topics are associated with each psychological theme, $K = 24n+1$ and all seed topics associated with a given psychological theme have the same value of l_k^s.

The last topic, K, is a baseline topic, for which both the regular and seed versions have $l_K^r(w) = l_K^s(w) = 1$ for all w, i.e., this topic may have positive weights on all words, including the "all other" word. This topic allows us to control for the baseline occurrence of words, as well as the length of the documents. In particular, as mentioned earlier, the last word in our vocabulary (indexed by W) is an "all other" word that captures any word that appears in any document but that is not in our vocabulary. We use this "word" to account for the total number of words in documents, and constrain it to have a positive weight only on the last topic K, i.e., $l_k^r(W) = l_k^s(W) = 0$ for all $k < K$, and $l_K^r(W) = l_K^s(W) = 1$. That is, the number of tokens associated with word W in topic K for document d is equal to the number

of tokens in the document that are not equal to any word in our vocabulary.

Guided LDA Estimation

The priors on the topic-word probabilities $\{\phi_k^r\}$ and $\{\phi_k^s\}$ and the document-topic probabilities $\{\theta_d\}$ are given as follows: $\phi_r^s \sim Dirichlet(\alpha_1 l_k^r)$; $\phi_k^s \sim Dirichlet(\alpha_1 l_k^s)$; $\theta_d \sim Dirichlet(\alpha_2 1_K)$. Given this specification, parameters may be estimated using Gibbs sampling, based on the posterior distributions of all variables, which are given in closed form, as specified in Web Appendix A. We estimate the model using MCMC with 5,000 iterations, using the first 1,000 as burn-in and saving one in 10 iterations thereafter. We estimate four versions of Guided LDA, where we vary the number of topics per psychological theme between $n = 1$ and $n = 4$. Increasing the value of n beyond 4 raised issues of convergence in the next step of our analysis, where we estimate consumer choice models based on data from studies 1 and 2. In order to inform model selection, we compute the Deviance Information Criterion (DIC), based on Celeux et al. (2006).[6]

Application to movies

Movie Descriptions

In this paper we apply Guided LDA to movies, which are probably the type of entertainment products that have received the most attention in the marketing literature. In our main analysis, we use synopses of movies, available on imdb.com, as input into our Guided LDA analysis. Synopses offer several benefits. First, they are not unique to the movie indus-

[6] We use a formulation of the DIC that is specific to models with latent variables (based on DIC_7 in Celeux et al. (2006)).

try and they are available for most entertainment products. Second, compared to reviews, synopses have the benefit of being objective descriptions rather than subjective evaluations. Subjective evaluations would be problematic in our case, because the language used to express these evaluations tends to overlap with the language used to describe psychological themes. For example, the fact that a reviewer wrote that he or she "loved" a particular aspect of a movie does not imply that Love is a theme featured in the movie. Finally, synopses have the benefit of being publicly available.

We assembled a dataset of 429 movie descriptions.[7] This set is the union of the 39 movies that received one of the "big five" oscars between 2004 and 2014 (which were used in Study 1, see Table 2), the top 40 movies in terms of US domestic box office performance in 2013 (which were used in Study 2, see Table 3), as well as all movies that were manually assigned to "Character Strengths" by Niemiec and Wedding (2014). Selecting movies that were manually assigned to various "Character Strengths" increases the chance that all psychological themes be represented in the sample, and improves our ability to define topics related to each psychological theme.

We preprocessed all movie descriptions following standard practice, using the R tm package. We eliminate non-English characters and words and tokenize the text. Following the standard "bag of words" approach, after preprocessing each movie description is treated as an unordered set of tokens. Table 4 presents some descriptive statistics of movie descriptions (i.e., synopses).

<INSERT TABLES 2, 3, 4 ABOUT HERE>

[7]There were 39 movies for which the synopsis was not available on imdb.com. For these movies, we used the plot summary instead of the synopsis (available either on imdb.com or on wikipedia.org).

For robustness, we repeat the analysis with two other data sources: movie spoilers and scripts. Spoilers, also used by Eliashberg et al. (2007), provide extensive summaries of movies. However, spoilers present at least two potential limitations, compared to synopses. First, they tend to vary across movies in quality and style. Second, spoilers are fairly unique to the movie industry, and we would like to ensure that our approach is applicable to any entertainment product. Like Eliashberg et al. (2007), we access movie spoilers from the publicly available resource www.themoviespoiler.com. Scripts were obtained from the Internet Movie Script Database (www.imsdb.com). We report the results based on spoilers in Web Appendix D, and the results based on scripts in Web Appendix E. We find that our results are not sensitive to the use of spoilers vs. synopses vs. scripts. When using Guided LDA features as input into predictive models of aggregate performance, we use spoilers as input to Guided LDA in order to improve the comparison with the LSA features created based on Eliashberg et al. (2007).

Guided LDA Results

Table 5 reports the DIC for Guided LDA when n, the number of topics per psychological theme, is varied from one to four. In addition, for each version of Guided LDA, we run Traditional LDA with the same number of topics. First, we see that the DIC favors Guided LDA over Traditional LDA. Next, comparing different versions of Guided LDA, we see that there is value in allowing $n > 1$, and that $n = 4$ gives rise to the lowest DIC. As noted above, increasing n further led to convergence issues when estimating choice models on the data collected in studies 1 and 2. Therefore we stopped at four topics per psychological theme.

Table 6 reports the 10 topics with the highest total weight on seed words,[8] excluding the baseline topic. Table 6 also reports examples of movies that have high weights on each topic, along with words that have high relevance for the topic and that appear in that movie's description.[9] Web Appendix B reports word clouds that reflect the most relevant words for each of these topics. These figures illustrate the benefits of allowing multiple topics per psychological theme. For example, the topic "Leadership 3" appears to capture leadership in the context of sports, while the topic "Leadership 4" appears to capture leadership in the context of national crises. Similarly, "Love 1" tends to relate to family relationships with an emphasis on mothers, while "Love 3" tends to capture romantic relationships among younger people, and "Love 4" tends to capture romantic relationships among adults. (The word clouds reported in Web Appendix B are provided for illustration purposes only, and are not used in any other part of the paper.)

<INSERT TABLES 5 AND 6 ABOUT HERE>

Using Guided LDA Features as Input into Predictive Consumer Choice Models

We now explore whether describing entertainment products based on topics estimated by Guided LDA may improve the performance of predictive content-based and hybrid consumer choice models for these products.

[8] The total weight of seed words on topic k is equal to: $\pi_k + (1-\pi_k)\Sigma_w \phi_k^r(w) l_k^s(w)$.

[9] The relevance of word w to topic k is a measure of the weight of this word on that topic ($\phi_k(w) = \pi_k \phi_k^s(w) + (1-\pi_k)\phi_k^r(w)$), controlling for the average weight of the word across topics, i.e., this measure identifies words that are more uniquely identified with each topic. More precisely, we measure relevance as: $\lambda log(\phi_k(w)) + (1-\lambda) log(\frac{\phi_k(w)}{\overline{\phi}(w)})$, where $\phi_k(w)$ is the weight of word w on topic k, $\overline{\phi}(w)$ is the average weight of word w across topics, and λ is the weight placed on the weight $\phi_k(w)$ relative to its lift $\frac{\phi_k(w)}{\overline{\phi}(w)}$. Following Liu and Toubia (2016), we set $\lambda = 0.6$.

Empirical framework

We continue with our application to movies. We focus on individual-level consumption. That is, our dependent variable is whether a particular consumer chose to watch a particular movie. We focus on consumption data in this paper because they are managerially relevant in every entertainment industry.

We note that by definition, the decision by a consumer to watch a movie is based on information they collect *before* watching the movie. In our main analysis, we use synopses from IMDB as movie descriptions. These are available to consumers before watching a movie, hence the input to Guided LDA does not include any information that was unavailable to consumers at the time at which they decided to watch a movie. Note that we do not assume however that all consumers actually read a movie's synopsis before deciding to watch it. Rather, we treat synopses as one source of information on the content of the movie. Consumers may base their decision to watch a movie on this or any other information that also reflects the movie's content (e.g., trailers, previews, billboards, reviews, word of mouth).

We consider data that capture binary viewing decisions made by C consumers on M movies. We do not assume that data are available for all consumers on all movies, but rather that we have at least some movie watching data for each consumer. We index consumers by c and movies by m. We specify a simple predictive content-based model that links product features to movie consumption. In particular, we adopt approach that is standard in the marketing literature and that is well suited for statistical inference. We simply assume a linear additive utility function with binomial logistic choice probabilities:

$$Prob(y_{cm} = 1) = \frac{exp(X_m W_c)}{1 + exp(X_m W_c)} \tag{1}$$

where y_{cm} is a binary variable that captures whether consumer c watched movie m, X_m is a row vector of covariates (features) that describe movie m, and W_c is a column vector of weights on each feature for consumer c.

Design of the studies

In Study 1 we tested the use of Guided LDA features in predicting the consumption of movies that may be considered of "high quality." This study focused on movies that won one of the "big five" Oscars between 2004 and 2014 (Best Picture, Best Actor, Best Actress, Best Director, Best Original Screenplay). The list of movies in Study 1 is included in Table 2. We recruited participants from Amazon Mechanical Turk's online panel, screened for being based in the US. We asked each respondent to indicate whether they had watched each of the movies in the set. We received complete data from N=599 participants, who were each paid $1 for their participation. Each movie had been watched by an average of 33.19% of the participants (standard deviation 16.57%), and each participant had watched on average 12.94 of the movies in the sample (standard deviation 7.84).

In Study 2 we explored whether our results generalize to "blockbuster" movies. We selected the top 40 movies based on US domestic box office performance in 2013 (the study was run in the Summer of 2014). The list of movies included in Study 2 is available in Table 3. We recruited participants from Amazon Mechanical Turk's online panel, screened for being based in the US. Again, we asked each respondent to indicate whether they had watched each of the movies in the set. We received complete data from N=542 respondents, who were each paid $1 for their participation. Each movie had been watched by an average of 30.33% of the participants (standard deviation 11.11%), and each participant had watched

on average 12.13 of the movies in the sample (standard deviation 8.30).

We recognize that relying on respondents' recollection of which movies they watched is likely to induce some noise in the dependent variable. However, our comparisons hold the dependent variable constant, and explore different sets of features that may be used to describe movies and predict this dependent variable. Therefore, any measurement error in the dependent variable would only reduce our ability to differentiate between sets of features, which makes our results more conservative.

Movie features

Our dependent variable is y_{cm} (whether consumer c watched movie m), which was collected in the survey. We consider three sets of predictive variables (features) that may be used to describe movies and predict this dependent variable at the consumer level. The list of variables is summarized in Table 7. These variables were collected for all the movies included in studies 1 and 2 (with the exception of the "sequel" variable in Study 1, in which only one of the 40 movies was a sequel and the "number of tweets" variable in Study 1, as many movies in this set were released before social media became significant). Survey respondents had no input into any of the movie features, they only provided us with the dependent variable y_{cm}.

The first set of features capture information about movies that is commonly considered in academic studies on movies (Baek et al., 2017; Eliashberg et al., 2006; Ghiassi et al., 2015; Litman and Ahn, 1998; Narayan and Kadiyali, 2015; Ravid, 1999; Sharda and Delen, 2006; Zufryden, 1996). For each movie in each study, we collect the average critic rating (from metacritic.com); the average user score (from metacritic.com); the production budget

(from imdb.com, adjusted for inflation using the tool available at http://data.bls.gov/cgi-bin/cpicalc.pl); the maximum number of screens on which the movie was shown in the US throughout the course of its run in theaters, known as "widest release" (available from boxofficemojo.com), for which we also include a square term; the domestic box office performance (from imdb.com, adjusted for inflation using the tool available at http://data.bls.gov/cgi-bin/cpicalc.pl); the MPAA rating (from imdb.com);[10] the movie's run time in minutes (from boxofficemojo.com); a dummy variable equal to 1 if the movie was a sequel; the degree of competition faced by the movie at the time of its release, captured by two dummy variables (following Sharda and Delen, 2006): a "High Competition" variable is equal to 1 for movies released in the months of June and November, and a "Medium Competition" equal to 1 for movies released in the months of May, July and December (release month was obtained from imdb.com); "star power," measured as the power of the highest rated star in the movie at the time of its release (following Elberse and Eliashberg, 2003), where power is measured using the starmeter rating provided by IMDB; a measure of activity on Twitter, based on the publicly available MovieTweetings database of Dooms et al. (2013) (we use the total number of tweets about each movie in the database as a cumulative measure of activity); the time elapsed between the release of the movie in theatres (obtained from IMDB) and the release of the DVD (obtained from Amazon); the sales rank of the movie's DVD as of December 2017 (obtained from Amazon).

The second set of features capture specifically the content of each movie, and are based on the work of Eliashberg et al. (2007, 2014). First, we extract genre and content variables.

[10] In Study 1 we capture MPAA rating with one dummy variable indicating whether the movie is R rated (there are no G-rated movie and only one PG-rated movie in this study, so we combine G, PG, and PG-13 ratings as the baseline). In Study 2 we use two dummies variables indicating whether the movie is rated R or PG-13 (there is only one G-rated movie in this study, so we combine PG and G ratings as the baseline)

We asked two independent readers trained in film studies to read the script of each movie (when available, otherwise we used the spoiler) and answer the same questionnaire as in Eliashberg et al. (2014) (Section 2.1., Page 2640). The level of agreement between the two judges, 84.38%, is similar to the one reported by these authors. We average the two readers' binary responses for each question.[11] Second, we extract "semantic variables" on each movie based on their spoilers, following Eliashberg et al. (2007). Using MS Word, we extract the number of characters, the number of words, the number of sentences, and the average number of characters per word, for each movie spoiler.[12] Third, we extract "Bag-of-Words" variables. In order to increase the precision of these features and make them more comparable to Guided LDA features, we base this analysis on the same set of 429 movies on which Guided LDA was run, although only the features created for the movies in studies 1 and 2 are needed for this analysis. Because scripts were not available for all movies under study, we use spoilers like Eliashberg et al. (2007). We use the same approach as Eliashberg et al. (2014) (see Section 2.3, Page 2640-2642). That is, we first eliminate all punctuations, standard English names, stop words, and stem the words. Next, we compute an importance index (similar to *tf-idf*) for each word using the same formula as Eliashberg et al. (2007, 2014) (see Equation (1) in Eliashberg et al., 2014), and keep the top 100 most important words. We perform Latent Semantic Analysis (LSA) on the word-document matrix, like Eliashberg et al. (2014). Similar to Eliashberg et al. (2007, 2014), we find an "elbow" at the two singular-value solution, and hence extract two features for each movie.

[11] One of the judges was unable to answer questions on 17 of the movies. For these movies, we use the responses from the other judge only.

[12] Eliashberg et al. (2007) also extract the proportion of passive sentences. We were unable to do so as this function appears to be unavailable on more recent versions of MS Word. Note also that we extract semantic variables based on spoilers like Eliashberg et al. (2007), rather than scripts like Eliashberg et al. (2014), because scripts were not available for all the movies under study.

The third and final set of features consists of the weights θ_m estimated by Guided LDA, capturing the extent to which movie m features each topic. We drop the baseline topic and are left with 96 weights for each movie.

We stress again that our focus in this paper is on *predictive* models, and we do not make any claim of causality between any of these features and the dependent variable. In particular, there may exist additional, "omitted" variables that correlate both with the features considered here and the dependent variable.

<INSERT TABLE 7 ABOUT HERE>

Leveraging Guided LDA Features in Content-Based Choice Models

We start by illustrating the use of Guided LDA features in content-based models of consumer behavior. We cannot consider all content-based models that have been proposed in the literature. Instead, we use a hierarchical Bayes logistic choice model based on Equation 1. We assume a normal prior on the weight vectors: $W_c \sim N(W_0, D)$, where W_c is a set of individual-level weights for consumer c (i.e., the model is estimated at the individual level). As an alternative estimation approach, we also consider the LOG-Het method proposed by Evgeniou et al. (2007). Log-Het is a machine learning-based approach that explicitly controls for complexity; it was designed specifically for individual-level choice data with a panel structure like ours. Details are provided in Web Appendix C. We focus here on the hierarchical Bayes approach, which gives rise to similar conclusions.

We assume a diffuse improper prior on W_0 and an inverse-Wishart prior on D: $D^{-1} \sim Wishart(0.001I, npar+3)$, where $npar$ is the dimensionality of W_c. We estimate all versions of the model using hierarchical Bayes MCMC (Rossi et al., 2012) with 100,000 iterations,

using the first 50,000 as burn-in and saving one in ten iterations. We measure goodness of fit using DIC (Celeux et al., 2006).

We randomly select five movies as holdouts for each respondent in each study, i.e., the identity of the holdout movies varies across respondents. We compute a hit rate for each observation, i.e., for each consumer-movie pair for which we collected y_{cm}. The hit rate is defined as the average posterior probability of the value of y_{cm} that was observed in the data. We also report an analysis of true positive and true negative rates in Web Appendix F.

Our main focus in this section is on comparing the value of Guided LDA features to that of other features based on the content of movies, and in particular to features based on an unsupervised LSA approach. To that effect, we test whether Guided LDA may complement or replace some of the features developed by Eliashberg et al. (2007, 2014). We start with a specification of the choice model with an intercept only as a baseline (Version 1). Next, we consider the inclusion of the basic movie features (average critic rating, production budget, etc. - Version 2). Next, we consider the addition of the features based on Eliashberg et al. (2007, 2014), i.e., genres, content variables, semantic variables and bag-of-words LSA variables (Version 3). We consider replacing the bag-of-words variables created using LSA with Guided LDA features (Version 4), as both of these sets of features are based on some natural language processing of movie descriptions. Finally, we consider replacing all of the features based on Eliashberg et al. (2007, 2014) with Guided LDA features (Version 5). For each version of the model, we compute the average in-sample and out-of-sample hit rates for each consumer. We compare hit rates across versions of the model using standard paired t-tests (i.e., the number of observations for the t-tests is the number of consumers).

Tables 8 and 9 present the results for studies 1 and 2 respectively. We see that including

the basic movie features in the model (average critic rating, production budget, etc.) significantly improves both in-sample and out-of-sample fit ($p < 0.05$), compared to a version of the model with an intercept only (Version 2 vs. Version 1). Adding features based on Eliashberg et al. (2007, 2014) that describe the actual content of the movie improves in-sample and out-of-sample fit further (Version 3 vs. Version 2). Replacing the unsupervised featured based on LSA with Guided LDA features gives rise to a significant improvement in in-sample and out-of-sample fit (Version 4 vs. Version 3). Replacing *all* the features based on Eliashberg et al. (2007, 2014) with Guided LDA features also significantly improves in-sample and out-of-sample fit (Version 5 vs. Version 3). Conditioning on the presence of Guided LDA features (i.e., comparing Versions 4 and 5), in-sample fit is significantly improved with the presence of genres, content variables and semantic variables, but out-of-sample fit is not improved by the presence of these features.

Based on these comparisons, it appears that Guided LDA topic weights have the potential to increase the ability of content-based choice models to predict the consumption of entertainment products by individual consumers. In our data, including Guided LDA features in addition or instead of other features gives rise to improvements in in-sample hit rates in the order of 10%, and improvements in out-of-sample hit rates in the order of 1-3%. Given the size of the filmed entertainment industry ($88.3 billion globally in 2015, Statista, 2016b), such improvements in hit rates might be worth millions of dollars to companies involved in the production and/or distribution of content.

<INSERT TABLES 8 AND 9 ABOUT HERE>

Leveraging Guided LDA Features in Hybrid Models: Content-Boosted Collaborative Filtering

Our analysis so far has focused on traditional content-based choice models. In this subsection, we explore the use of Guided LDA features in hybrid approaches that combine content-based and collaborative methods. Like with pure content-based approaches, we cannot test Guided LDA with all hybrid methods that have been proposed in the literature. We focus here on Content-Boosted Collaborative Filtering (Melville et al., 2002), which has become particularly popular and has shown consistently high performance (Burke, 2007). Future research may incorporate Guided LDA features in other hybrid approaches, such as the ones proposed by Ansari et al. (2000), Bodapati (2008), or Ying et al. (2006). Note that we do not test the effectiveness of recommendations directly, but rather explore whether the proposed features may improve the predictive validity of hybrid methods, which is a pre-requisite for improving recommendations.

We first describe the pure neighborhood-based Collaborative Filtering (CF) framework (interested readers are referred to Breese et al. (1998) and Linden et al. (2003) for more detail). For each consumer c, each observation in the training sample is considered as a "vote" against or in favor of that movie m based on whether movie m was watched by consumer c. Positive votes are given a weight proportional to the number of consumers in the sample who watched that movie, to capture that similarity on less-watched movies is more predictive (Linden et al., 2003). That is, consumer c's weighted vote on movie m is $v_{c,m} = y_{cm}/\bar{y}_m$, where $y_{cm} = 1$ if consumer c watched movie m and $\bar{y}_m$ is the proportion of consumers who watched movie m among consumers for whom that movie was in the training set. (Similar results were obtained without this weighting). The distance between

each pair of consumers c and c' is computed as the cosine between their vectors of weighted votes: $w(c,c') = \sum_{m \in I_c \cap I_{c'}} \frac{v_{c,m} v_{c',m}}{\sqrt{\sum_{m \in I_c} v_{c,m}^2 \sum_{m \in I_{c'}} v_{c',m}^2}}$, where I_c is the set of training movies for consumer c.

In a pure neighborhood-based CF framework, the predicted probability that consumer c would watch an out-of-sample movie m is given by: $\hat{y}_{c,m} = \bar{y}_c + \frac{\sum_{c':m \in I_{c'}} w(c,c')(y_{c',m} - \bar{y}_{c'})}{\sum_{c':m \in I_{c'}} w(c,c')}$, where $\bar{y}_c$ is the proportion of movies in c's training sample for which $y_{c,m} = 1$. In other words, the prediction for consumer c and movie m is equal to that consumers' average propensity to watch movies, adjusted up or down based on other consumers' data, where more weight is given to consumers whose profile is more similar to c's.

Content-Boosted CF extends this framework by mixing pure neighborhood-based CF predictions with pure content-based predictions (Melville et al., 2002). The pure content-based predictions may come from any model; we use the hierarchical Bayes choice model from the previous subsection. Let $p_{c,m}$ be the content-based predicted probability that consumer c would watch movie m. Such predictions are available for all out-of-sample observations. Observations $y_{c,m}$ for consumer c are augmented as follows:

$$z_{c,m} = \begin{cases} y_{c,m} \text{ if } m \in I_c \\ p_{c,m} \text{ if } m \notin I_c \end{cases}$$

That is, this approach "fills the holes" in the out-of-sample observations using pure content-based predictions. The predicted probability that consumer c would watch an out-of-sample movie m is given as follows for Content-Boosted CF: $\hat{y}_{c,m} = \bar{z}_c + \frac{w_s \times (z_{c,m} - \bar{z}_c) + \sum_{c' \neq c} w(c,c')(z_{c',m} - \bar{z}_{c'})}{w_s + \sum_{c' \neq c} w(c,c')}$, where w_s is the weight on the pure content-based prediction for that movie and that user, versus the prediction based on the other users. For illustration, we set this weight equal

to the number of consumers in the sample minus one, i.e., the content-based and the CF predictions are weighted similarly. Similar results are obtained with different weights (results are available from the authors).

Neither pure CF nor Content-Boosted CF produce in-sample hit rates and the DIC is not available for these methods. We compare the pure CF model to different versions of Content-Boosted CF, in which the content-based predictions come from Versions 2 to 5 of the content-based model tested in tables 8 and 9. That is, for each version of the content-based model (except for the intercept-only version), we test a corresponding version of Content-Boosted CF. Results are reported in tables 10 and 11. We see that the comparisons are similar to those with the pure content-based model. That is, the introduction of Guided LDA features has a significant positive impact on predictive validity, above and beyond the other features. This suggests that Guided LDA features may be used to improve the predictive performance of hybrid methods that combine content-based predictions with a collaborative filtering framework.

<INSERT TABLES 10 AND 11 ABOUT HERE>

Using Guided LDA Features as Input into Predictive Models of Aggregate Performance

We developed Guided LDA for use with content-based and hybrid predictive models that leverage panel data in which individual consumers are observed making decisions over time. Nevertheless, in this section we explore the use of Guided LDA features in predictive models of *aggregate* demand for entertainment products. In order to do that, we replicate (to the

best of our abilities) the main models proposed by Eliashberg et al. (2007) and Eliashberg et al. (2014). Following Eliashberg et al. (2014) (Equation 2, Page 2643), we measure the aggregate performance of movie i as: $y_i = log(\frac{BOX\ OFFICE_i}{BUDGET_i})$, where $BOX\ OFFICE_i$ is the box office performance of the movie, and $BUDGET_i$ is its production budget. We focus on the combined set of movies from studies 1 and 2.

We use again the genres, content, semantic, and bag-words variables based on Eliashberg et al. (2007, 2014), as described in the section "Movie features." We test two specific models, based respectively on Eliashberg et al. (2007) and Eliashberg et al. (2014). The first model is a Bagged-CART model based on the Bag-CART model of Eliashberg et al. (2007), which we replicated to the best of our ability based on the information contained in the paper.[13] The second model is a Kernel-Based model based on the Kernel-II (optimized feature weights) model of Eliashberg et al. (2014), which we again replicated to the best of our ability based on the information contained in the paper and Appendix.[14] Our implementation code for both models is available upon request.

Like Eliashberg et al. (2007, 2014), we measure performance using the Mean Square Error (MSE) between the actual and predicted dependent variable on a holdout sample of movies. We split the sample into 65 movies for calibration and 14 movies for validation. Because our sample size is smaller than that of Eliashberg et al. (2007, 2014), we reduce the sensitivity to the set of calibration vs. validation movies by replicating the analysis 100 times, each time

[13] In order to make the results comparable to the other model, we use $y_i = log(\frac{BOX\ OFFICE_i}{BUDGET_i})$ as dependent variable rather than $log(ROI + 1)$. Due to the constraints of our programming environment (Matlab), we constrain the number of splits in each tree to be no greater than 14, rather than constraining the number of layers to be no greater than 4 (a tree with 4 layers can have a maximum of 14 splits). Like Eliashberg et al. (2007), we average over 1,000 bootstrap trees.

[14] We calibrated the tuning parameter θ, the feature weights v, and the complexity penalty λ using the same approach as Eliashberg et al. (2014). We found it necessary to adjust the range of possible values for λ given our data.

with a different random split between calibration and validation movies.

As in the previous section, we consider replacing the bag-of-words variables created using LSA with Guided LDA features, and we consider replacing all of the features based on Eliashberg et al. (2007, 2014) with Guided LDA features. We report the average MSE for each version of each model, in Table 12. We see that both for the Bagged-CART model based on Eliashberg et al. (2007) and for the Kernel-based model based on Eliashberg et al. (2014), performance is improved when Guided LDA features are included instead of the LSA "Bag-of-Words" variables. We see that performance is also improved when all of the features based on Eliashberg et al. (2007, 2014) are replaced with Guided LDA features. However, this time, conditioning on the inclusion of Guided LDA features, performance is improved when genres, content variables, and semantic variables are included.

This analysis suggests that although Guided LDA was developed for use in individual-level predictive models of consumption, it appears to be also useful for constructing features to be used in aggregate predictive models, such as the ones proposed by Eliashberg et al. (2007, 2014). This exercise also further illustrates that Guided LDA is an approach for constructing features to be incorporated into various extant models, rather than a new model designed to "compete" with extant models.

<INSERT TABLE 12 ABOUT HERE>

Conclusions

In this paper we bridge the media psychology literature, the positive psychology literature, the natural language processing literature, the choice modeling literature, and the

collaborative filtering literature. We propose a new set of descriptors of entertainment products, theoretically founded in the media psychology literature and the positive psychology literature. We rely on the natural language processing literature to develop a method for tagging entertainment products in an automated and scalable manner. In the context of movies, we first show that the proposed features improve our ability to predict consumption at the individual level. We find this result both with award-winning movies and blockbuster movies. We illustrate the use of Guided LDA features in pure content-based models as well in hybrid models that combine content-based predictions with collaborative filtering. We also show that Guided LDA features have the potential to improve the performance of models that predict aggregate performance outcomes rather than individual-level consumption.

Managerially, the proposed feature extraction approach may be implemented in an automated and scalable way, to provide features that may be included into any existing content-based or hybrid choice model. Accordingly, our research makes these models more attractive to practitioners in the media and entertainment industry. This may improve even further the use and impact of the content-based and hybrid approaches developed in the marketing literature for that industry (e.g., Ansari et al., 2000; Bodapati, 2008; Ying et al., 2006). This may also make marketing models that were initially developed for other industries (e.g., based on conjoint analysis or scanner data), more relevant and applicable to the media and entertainment industry.

We close by highlighting several opportunities for future research. First, while our two studies cover "blockbuster" movies as well as "high-quality" movies, future research may test our approach on other sets of movies. Second, while our theoretical development is relevant to entertainment products in general, our current analysis is based on movies only. Our

results may be replicated with other types of entertainment products. Third, given the focus in the positive psychology literature on improving well-being and life satisfaction, it would be worthwhile to study the link between psychological themes in entertainment products and well-being. In particular, recommendation engines may be developed based on the proposed approach, that would be designed to increase not only consumption, but also well-being.

References

Adomavicius, Gediminas, Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on* **17**(6) 734–749.

Ang, Ien. 1985. *Watching Dallas: Soap Opera and the Melodramatic Imagination*. Methuen (London, UK).

Ansari, Asim, Skander Essegaier, Rajeev Kohli. 2000. Internet recommendation systems. *Journal of Marketing research* **37**(3) 363–375.

Archak, Nikolay, Anindya Ghose, Panagiotis G. Ipeirotis. 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Science* **57**(8) 1485–1509.

Austin, Bruce A. 1986. Motivations for movie attendance. *Communication Quarterly* **34**(2) 115–126.

Baek, Hyunmi, Sehwan Oh, Hee-Dong Yang, JoongHo Ahn. 2017. Electronic word-of-mouth, box office revenue and social media. *Electronic Commerce Research and Applications* **22** 13–23.

Blacker, Irwin R. 1988. *The elements of screenwriting: a guide for film and television writers*. Collier Books.

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM* **55**(4) 77–84.

Blei, David M., Andrew Y. Ng, Michael I. Jordan, John Lafferty. 2003. Latent dirichlet allocation. *Machine Learning* **3**(4/5) 993–1022.

Bodapati, Anand V. 2008. Recommendation systems with purchase data. *Journal of marketing research* **45**(1) 77–93.

Breese, John S., David Heckerman, Carl Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. *Proc. 14th Conf. Uncertainty in Artificial Intelligence* 43–52.

Burke, Robin. 2007. Hybrid web recommender systems. *The adaptive web*. Springer, 377–408.

Celeux, Gilles, Florence Forbes, Christian P. Robert, D. Michael Titterington. 2006. Deviance information criteria for missing data models. *Bayesian Analysis* **1**(4) 651–673.

Chu, Wei, Seung-Taek Park, Todd Beaupre, Nitin Motgi, Amit Phadke, Seinjuti Chakraborty, Joe Zachariah. 2009. A case study of behavior-driven conjoint analysis on yahoo! front page today module. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM* .

Cohen, Jonathan. 2001. Defining identification: A theoretical look at the identification of audiences with media characters. *Mass Communications and Societe* **4**(3) 245–264.

Cohen, Jonathan. 2006. *Audience Identification With Media Characters, in Psychology of Entertainment*. Lawrence Erlbaum Associates, 183–197.

Cuadrado, Manuel, Marta Frasquet. 1999. Segmentation of cinema audiences: An exploratory study applied to young consumers. *Journal of Cultural Economics* **23** 257–267.

d'Astous, Alain, Nadia Touil. 1999. Consumer evaluations of movies on the basis of critics' judgments. *PsychologyMarketing* **16**(8) 677–694.

De Silva, Indra. 1998. *The Motion Picture Mega-Industry*, chap. Consumer Selection of Motion Pictures. Allyn and Bacon, 144–171.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* **46**(6) 391–407.

Dellarocas, Chrysanthos, Xiaoquan (Michael) Zhang, Neveen F. Awad. 2007. Exploring the value of online product reviews in forecasting sales: the case of motion pictures. *Journal of Interactive Marketing* **21**(4) 23–45.

Dooms, Simon, Toon De Pessemier, Luc Martens. 2013. Movietweetings: a movie rating dataset collected from twitter. *Workshop on Crowdsourcing and human computation for recommender systems, CrowdRec at RecSys*, vol. 2013. 43.

Duana, Wenjing, Bin Gub, Andrew B. Whinston. 2008. The dynamics of online word-of-mouth and product salesan empirica linvestigation of the movie industry. *Journal of Retailing* **84**(2) 233–242.

Elberse, Anita, Jehoshua Eliashberg. 2003. Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures. *Marketing Science* **22**(3) 329–354.

Eliashberg, Jehoshua, Anita Elberse, Mark A.A.M. Leenders. 2006. The motion picture industry: Critical issues in practice, current research, and new research directions. *Marketing Science* **25**(6) 638–661.

Eliashberg, Jehoshua, Mohanbir S. Sawhney. 1994. Modeling goes to hollywood: Predicting individual differences in movie enjoyment. *Management Science* **40**(9) 1151–1173.

Eliashberg, Joshua, Sam K. Hui, John Z. Zhang. 2007. From story line to box office: A new approach for green-lighting movie scripts. *Management Science* **53**(6) 881–893.

Eliashberg, Joshua, Sam K. Hui, Z. John Zhang. 2014. Assessing box office performance using movie scripts: A kernel-based approach. *IEEE Transactions on Knowledge and Data Engineering* **26**(11) 2639–2648.

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

Field, Syd. 2007. *Screenplay: The foundations of screenwriting*. Delta.

Ghiassi, M, David Lio, Brian Moon. 2015. Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Systems with Applications* **42**(6) 3176–3193.

Ghose, Anindya, Panagiotis G. Ipeirotis, Beibei Li. 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science* **31**(3) 493–520.

Hauge, Michael. 2011. *Writing screenplays that sell*. A&C Black.

Hoffner, Cynthia, Joanne Cantor. 1991. *Perceiving and Responding to Mass Media Characters in Responding to the Screen: Reception and Reaction Processes*. Lawrence Erlbaum Associates (Hillsdale, NJ), 63–103.

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* .

Koren, Yehuda, Robert Bell, Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* (8) 30–37.

Kraaykamp, Gerbert, Koen Van Eijck. 2005. Personality, media preferences, and cultural participation. *Personality and individual differences* **38**(7) 1675–1688.

Lee, Thomas Y., Eric T. Bradlow. 2011. Automated marketing research using online customer reviews. *Journal of Marketing Research* **48** 881–894.

Library of Congress. 2015. *Introduction to Library of Congress Genre/Form Terms for Library and Archival Materials*.

Linden, Greg, Brent Smith, Jeremy York. 2003. Amazon.com recommendations - item-to-item collaborative filtering. *IEEE Internet Computing* **7**(1) 76–80.

Litman, Barry R, Hoekyun Ahn. 1998. Predicting financial success of motion pictures. br litman the motion picture mega-industry.

Liu, Jia, Olivier Toubia. 2016. A semantic approach for estimating consumer content preferences from online search queries. *working paper, Columbia Business School* .

Liu, Yong. 2006. Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing* **70** 74–89.

Madrigal, Alexis C. 2014. How netflix reverse engineered hollywood. *The Atlantic* .

Manning, Christopher D, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge.

McCrae, Robert R., Paul T. Costa. 1999. *Handbook of personality: Theory and research 2 (2nd edition)*, chap. A five-factor theory of personality. New York: Guilford, 139–153.

McKee, Robert. 1997. *Substance, Structure, Style, and the Principles of Screenwriting*. New York: HarperCollins.

Melville, Prem, Raymond J Mooney, Ramadass Nagarajan. 2002. Content-boosted collaborative filtering for improved recommendations. *AAAI/IAAI*. 187–192.

Möller, K.E.Kristian, Pirjo Karppinen. 1983. Role of motives and attributes in consumer motion picture choice. *Journal of Economic Psychology* **4**(3) 239–262.

Mooney, Raymond J, Loriene Roy. 2000. Content-based book recommending using learning for text categorization. *Proceedings of the fifth ACM conference on Digital libraries*. ACM, 195–204.

Narayan, Vishal, Vrinda Kadiyali. 2015. Repeated interactions and improved outcomes: An empirical analysis of movie production in the united states. *Management Science* .

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, Moshe Fresko. 2012. Mine your own business: Market-structure surveillance through text mining. *Marketing Science* **31**(3) 521–543.

Niemiec, Ryan M., Danny Wedding. 2014. *Positive Psychology at the Movies (2nd ed.)*. Hogrefe, Boston MA.

Peterson, Christopher, Nansook Park, Martin EP. Seligman. 2005a. *Assessment of character strengths, in Psychologist desk reference, 2nd ed.*. New York: Oxford University Press, 93–98.

Peterson, Christopher, Nansook Park, Martin E.P. Seligman. 2005b. Orientations to happiness and life satisfaction: the full life versus the empty life. *Journal of Happiness Studies* **6** 25–41.

Peterson, Christopher, Martin EP Seligman. 2004. *Character Strengths and Virtues: A Handbook and Classification*. American Psychological Association / Oxford University Press.

Ravid, S Abraham. 1999. Information, blockbusters, and stars: A study of the film industry. *The Journal of Business* **72**(4) 463–492.

Rentfrow, Peter J, Lewis R Goldberg, Ran Zilca. 2011. Listening, watching, and reading: The structure and correlates of entertainment preferences. *Journal of personality* **79**(2) 223–258.

Rentfrow, Peter J, Samuel D Gosling. 2003. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology* **84**(6) 1236.

Rossi, Peter, Greg Allenby, Robert McCulloch. 2012. *Bayesian Statistics and Marketing*. John Wiley Sons.

Rust, Roland T, Mark I Alpert. 1984. An audience flow model of television viewing choice. *Marketing Science* **3**(2) 113–124.

Seligman, Martin E.P., Mihaly Csikszentmihalyi. 2000. Positive psychology: An introduction. *American Psychologist* **55**(1) 5–14.

Seligman, Martin EP., Tracy A. Steen, Nansook Park, Christopher Peterson. 2005. Positive psychology progress: Empirical validation of interventions. *American Psychologist* **60**(5) 410–421.

Shachar, Ron, John W Emerson. 2000. Cast demographics, unobserved segments, and heterogeneous switching costs in a television viewing choice model. *Journal of Marketing Research* **37**(2) 173–186.

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

Statista. 2016a. Electronic home video revenue in the united states from 2010 to 2019, by source (in billion u.s. dollars). *http://www.statista.com/statistics/259993/electronic-home-video-revenue-in-the-us-by-source/* .

Statista. 2016b. Filmed entertainment revenue worldwide from 2015 to 2019. *http://www.statista.com/statistics/259985/global-filmed-entertainment-revenue/* .

Statista. 2017a. Percentage of population who own an e-reader in the united states from 2014 to 2020. *https://www.statista.com/statistics/190283/penetration-rate-of-ereaders-in-the-united-states-since-2009/* .

Statista. 2017b. Value of the global entertainment and media market. *https://www.statista.com/statistics/237749/value-of-the-global-entertainment-and-media-market/* .

Su, Xiaoyuan, Taghi M Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* **2009** 4.

Tirunillai, Seshadri, Gerard J. Tellis. 2014. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research* **51** 463–479.

Weaver, James B. 1991. Exploring the links between personality and media preferences. *Personality and Individual Differences* **12**(12) 1293–1299.

Weaver, James B. 2003. Individual differences in television viewing motives. *Personality and individual differences* **35**(6) 1427–1437.

Ying, Yuanping, Fred Feinberg, Michel Wedel. 2006. Leveraging missing ratings to improve online recommendation systems. *Journal of marketing research* **43**(3) 355–365.

Zufryden, Fred S. 1996. Linking advertising to box office performance of new film releases-a marketing planning model. *Journal of Advertising Research* **36**(4) 29–42.

Tables

Table 1: List of psychological themes and examples of seed words.

Psychological Theme	**Examples of seed words**
Creativity	idea, original, novel
Curiosity	discover, question, interested
Open Mindedness	examine, considerate, impartial
Love of Learning	school, course, professor
Wisdom	experience, knowledge, advisor
Bravery	battle, hero, courage
Persistence	goal, effort, sacrifice
Integrity	truth, promise, genuine
Vitality	energy, peppy, enthusiastic
Love	relationship, marriage, friend
Kindness	gift, favor, compassion
Social Intelligence	psychologist, mindful, insightful
Citizenship	loyal, society, duty
Fairness	justice, law, rule
Leadership	team, captain, president
Forgiveness and Mercy	apologize, peace, repent
Humility and Modesty	humble, discrete, timid
Prudence	careful, responsible, safety
Self Regulation	abstain, restrain, virgin
Appreciation of Beauty and Excellence	wonder, awe, beautiful
Gratitude	gift, grateful, blessed
Hope	dream, opportunity, confidence
Humor	joke, laugh, funny
Spirituality	church, faith, heaven

Table 2: Movies included in Study 1 (winners of "big five" academy awards).

Year of award(s)	**Movie title**	**Year of award(s)**	**Movie title**
2004	Lost in Translation	2010	The Hurt Locker
2004	Monster	2010	The Blind Side
2004	Mystic River	2010	Crazy Heart
2004	The Lord of the Rings: The Return of the King	2011	The King's Speech
2005	Million Dollar Baby	2011	Black Swan
2005	Ray	2012	The Artist
2005	Eternal Sunshine of the Spotless Mind	2012	The Iron Lady
2006	Crash	2012	Midnight in Paris
2006	Capote	2013	Argo
2006	Brokeback Mountain	2013	Lincoln
2007	The Departed	2013	Silver Lining Playbook
2007	The Last King of Scotland	2013	Life of Pi
2007	The Queen	2013	Django Unchained
2007	Little Miss Sunshine	2014	12 Years as a Slave
2008	No Country for Old Men	2014	Dallas Buyer Club
2008	There Will be Blood	2014	Blue Jasmine
2008	La Vie en Rose	2014	Gravity
2008	Juno	2014	Her
2009	Slumdog Millionaire		
2009	Milk		
2009	The Reader		

Table 3: Movies included in Study 2 (top box office performers of 2013).

Box office rank	**Movie title**	**Box office rank**	**Movie title**
1	The Hunger Games: Catching Fire	21	Grown Ups 2
2	Iron Man 3	22	Anchorman 2: The Legend Continues
3	Frozen	23	Lone Survivor
4	Despicable Me 2	24	G.I. Joe: Retaliation
5	Man of Steel	25	Cloudy with a Chance of Meatballs 2
6	Gravity	26	The Wolf of Wall Street
7	Monsters University	27	The Butler
8	The Hobbit: The Desolation of Smaug	28	The Hangover Part III
9	Fast & Furious 6	29	The Wolverine
10	Oz The Great and Powerful	30	Now You See Me
11	Star Trek Into Darkness	31	Epic
12	Thor: The Dark World	32	Captain Phillips
13	World War Z	33	Bad Grandpa
14	The Croods	34	Pacific Rim
15	The Heat	35	This is the End
16	We're the Millers	36	Olympus Has Fallen
17	American Hustle	37	42
18	The Great Gatsby	38	Elysium
19	The Conjuring	39	Planes
20	Identity Thief	40	The Lone Ranger

Table 4: Descriptive statistics of movie descriptions (synopses).

Statistic	**Unit of analysis**	**Mean**	**St. dev.**	**Min**	**Max**
Number of words (including "all other")	Movie descriptions (N=429)	1446.65	1226.82	42	5817
Number of occurrences of seed words	Movie descriptions (N=429)	72.19	55,73	3	397
Number of unique seed words	Movie descriptions (N=429)	45.25	27.97	3	167
Number of psychological themes with at least one seed word occurrence	Movie descriptions (N=429)	18.43	4.26	3	24
Total number of occurrences across movie descriptions	Seed words (N=2677)	11.57	36.77	0	624
Proportion of movie descriptions with at least one occurrence	Seed words (N=2677)	0.02	0.04	0	0.61
Total number of occurrences across movie descriptions	Seed words with at least one occurrence (N=1608)	19.26	45.86	1	624
Proportion of movie descriptions with at least one occurrence	Seed words with at least one occurrence (N=1608)	0.03	0.05	0.002	0.61
Average number of seed word occurrences per movie description	Psychological Theme (N=24)	4.03	2.71	1.25	13.08
Proportion of movie descriptions with at least one seed word occurrence	Psychological Theme (N=24)	0.77	0.12	0.52	0.97

Table 5: 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	2,043.9	2,073.3
2	49	1,781.0	1,819.4
3	73	1,659.3	1,697.2
4	97	1,554.9	1,594.9

Increasing the number of topics per psychological theme beyond 4 led to convergence issues when estimating viewers' preferences for topics. Therefore we stopped at $n = 4$. Traditional LDA is nested within 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 6: Examples of topics from Guided LDA.

Topic	**Average document-topic weight * 10^{-3}**	**Example of movie with large weight**	**Examples of words with high relevance present in movie description**
"Citizenship 4"	4.07	My Big Fat Greek Wedding	family, daughter, time
"Creativity 4"	4.30	The Golden Compass	children, told, dust
"Fairness 1"	3.47	Robin Hood	king, sword, lady
"Leadership 3"	5.02	Glory Road	team, coach, players
"Leadership 4"	3.33	G.I. Joe: Retaliation	president, storm, shadow
"Love 1"	4.24	The Secret Life of Bees	mother, growing, bed
"Love 3"	3.47	Kissing Jessica Stein	friend, night, girl
"Love 4"	5.41	Sex and the City	wedding, marriage, affair
"Love of Learning 3"	4.61	Freedom Writers	students, school, class
"Vitality 3"	4.80	Eat, Pray, Love	life, returns, experience

The 10 topics with the highest total weights on seed words are presented in alphabetical order (baseline topic omitted).

Table 7: Variables in Studies 1 and 2.

Variables	Type	Description	Source
Movie Watching	Dependent	Dummy variable $y_{cm} = 1$ if consumer c watched movie m	Survey
Average Critic Rating	Predictive	Continuous variable between 0 and 100	Metacritic.com
Average User Score	Predictive	Continuous variable between 0 and 10	Metacritic.com
Production Budget (in $M)	Predictive	Continuous variable (inflation adjusted)	IMDB
Widest Release (in thousands of theatres)	Predictive	Continuous variable	boxofficemojo.com
Widest Release (in thousands of theatres)^2	Predictive	Continuous variable	boxofficemojo.com
Domestic Box Office (in $M)	Predictive	Continuous variable (inflation adjusted)	IMDB
MPAA rating	Predictive	Dummy variable(s)	IMDB
Run Time (in minutes)	Predictive	Count variable	boxofficemojo.com
Sequel	Predictive	Dummy variable	IMDB
Competition	Predictive	2 dummy variables	IMDB
Star Power	Predictive	Discrete variable	IMDB
Twitter Activity	Predictive	Discrete variable	MovieTweetings
DVD Release Timing	Predictive	Discrete variable (time elapsed between theatre and DVD release, in days)	IMDB+Amazon
DVD Sales Rank	Predictive	Discrete variable	Amazon
Genres	Predictive	8 variables	Independent raters (following Eliashberg et al. (2014)
Content variables	Predictive	24 variables	Independent raters (following Eliashberg et al. (2014)
Semantic variables	Predictive	4 variables	Word processing of spoilers (following Eliashberg et al. (2007)
Bag-of-Words variables from LSA	Predictive	2 continuous variables	LSA on spoilers (following Eliashberg et al. (2007)
Guided LDA topic weights	Predictive	96 continuous variables between 0 and 1	Guided LDA

The production budget and the domestic box office performance are adjusted for inflation using the tool available at http://data.bls.gov/cgi-bin/cpicalc.pl. MPAA Rating is captured by one dummy variable in Study 1 (R rated) and two dummy variables in Study 2 (R rated, PG-13 rated), as there is no G-rated movie and only one PG-rated movie in Study 1, and no G-rated movie in Study 2. Twitter Activity is available for Study 2 only, as many movies in Study 1 were released before social media became significant. Following Sharda and Delen (2006), Competition is captured by creating two dummy variables: "High Competition" is equal to 1 for movies released in the months of June and November, and the "Medium Competition" is equal to 1 for movies released in the months of May, July and December.

Table 8: 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	492.91	406.50	371.65	232.02	280.28
In-sample hit rate	62.09%	71.78%	76.30%	88.21%	85.08%
Out-of-sample hit rate	61.67%	66.44%	67.94%	70.32%	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 $p<0.05$.

Table 9: 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	232.02	280.28
In-sample hit rate	64.05%	73.12%	76.54%	86.35%	83.28%
Out-of-sample hit rate	63.60%	68.91%	69.93%	71.00%	70.89%

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 ($p=0.65$).

Table 10: 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%	69.90%	70.60%

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 $p<0.05$.

Table 11: 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%	70.31%	70.24%

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 out-of-sample hit rates are statistically significant at $p<0.05$, except the difference in out-of-sample hit rate between Pure Collaborative Filtering and CBCF-Version 2 ($p=0.10$), and between CBCF-Version 4 and CBCF-Version 5 ($p=0.71$).

Table 12: Using Guided LDA Features in Models that Predict Aggregate Performance.

Features	**Bagged CART**	**Bagged CART**	**Bagged CART**	**Kernel-Based**	**Kernel-Based**	**Kernel-Based**
Genres	✓	✓		✓	✓	
Content variables	✓	✓		✓	✓	
Semantic variables	✓	✓		✓	✓	
Bag-of-Words variables from LSA	✓			✓		
Guided LDA topic weights		✓	✓		✓	✓
Out-of-sample MSE	0.5186	0.4590	0.4657	0.5176	0.4703	0.4806

The Bagged CART model is based on the Bag-CART model of Eliashberg et al. (2007); the Kernel-Based model is based on the Kernel II (optimized feature weights) model of Eliashberg et al. (2014). We report the Mean Square Error (MSE) between the observed and predicted $log(\frac{BOX\ OFFICE}{BUDGET})$ on a set of holdout movies.