International Journal of Forecasting 32 (2016) 243–256

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting sales of new and existing products using consumer reviews: A random projections approach

Matthew J. Schneider [a,1], Sachin Gupta [b,*]

[a] Northwestern University, Medill School of Journalism, Media, Integrated Marketing Communications, 1845 Sheridan Road, Evanston, IL 60208-2101, United States

[b] Cornell University, Samuel Curtis Johnson Graduate School of Management, 452 Sage Hall, 14853 Ithaca, NY, United States

ARTICLE INFO

Keywords:
Big data
Forecasting
Consumer reviews
Textual data
Random projections

ABSTRACT

We consider the problem of predicting sales of new and existing products using both the numeric and textual data contained in consumer reviews. Many of the extant approaches require considerable manual pre-processing of the textual data, making the methods prohibitively expensive to implement and difficult to scale. In contrast, our approach uses a bag-of-words method that requires minimal pre-processing and parsing, making it efficient and scalable. However, a key implementation challenge with the bag-of-words approach is that the number of predictors can quickly outstrip the number of degrees of freedom available. Furthermore, the method can require impractically large computational resources. We propose a random projections approach for dealing with the curse-of-dimensionality issue that afflicts bag-of-words models. The random projections approach is computationally simple, flexible and fast, and has desirable statistical properties. We apply the proposed approach to the forecasting of sales at Amazon.com using consumer reviews with an attributes-based regression model. The model is applied to produce of one-week-ahead rolling horizon sales forecasts for existing and newly-introduced tablet computers. The results show that the predictive performance of the proposed approach for both tasks is strong and significantly better than those of either models that ignore the textual content of consumer reviews, or a support vector regression machine with the textual content. Furthermore, the approach is easy to repeat across product categories, and readily scalable to much larger datasets.

1. Introduction

User-generated online product reviews are an important source of market research information for firms. Since such reviews are a voluntary expression of consumers' experiences and beliefs about the quality of products and services, there is a lot that the interested firm can learn about the market by monitoring reviews closely. The literature has identified several managerial uses of this information, including gaining an understanding of the market structure (Netzer, Feldman, Goldenberg, & Fresko, 2012), identifying influential reviewers (Ghose & Ipeirotis, 2011), and learning about product attributes from a consumer perspective (Lee & Bradlow, 2011). Since consumers rely on online product reviews when making their own purchasing decisions (Chen & Xie, 2008; Zhao, Yang, Narayan, & Zhao, 2013), it seems natural that the content and valence of consumer reviews should help to predict consumer behavior. Accordingly, several studies have examined the

[*] Corresponding author.
E-mail addresses: matt.schneider@northwestern.edu (M.J. Schneider), sg248@cornell.edu (S. Gupta).
[1] Tel.: +1 847 467 1784.

http://dx.doi.org/10.1016/j.ijforecast.2015.08.005
0169-2070/© 2015 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

impact of consumer reviews on product sales (e.g., Archak, Ghose, & Ipeirotis, 2011 and Hu, Koh, & Reddy, 2013).

Consumer-provided online reviews are very voluminous and highly dynamic. In most cases, they are available without cost to any organization that can devote the relatively small resources that are needed to scrape the web. Most reviews contain not only numerical ratings of product opinions, but also textual content that adds considerable richness to the data. Collectively, consumer reviews represent a vast store of word-of-mouth information that consumers rely on in their purchase deliberations. All of this suggests that the information contained in consumer reviews should help to predict product sales.

Many of the characteristics of reviews that make them attractive as market research data also constitute challenges and limitations. For instance, there may be an inherent sample selection bias, because consumers who post a review are only a small subset of those consumers who buy the product or service because they think that they will like it. Li and Hitt (2008) point out that this may be the case with early buyers in particular, implying that reviews will tend to become more negative over time. Posted reviews are a result not only of consumers' objective product evaluations, but also of social dynamics of opinions (e.g., Moe & Trusov, 2011). The distributions of numeric ratings are typically available on a small number of discrete points, often five, and are also skewed towards the most positive scale point. The textual content of reviews can be an important source of additional, nuanced information about consumer perceptions and evaluations. However, from a modeling perspective, it takes effort to extract information from textual content, because the data are largely qualitative. Examples of text mining include studies by Archak et al. (2011), Ghose and Ipeirotis (2011), and Hu et al. (2013). Using human reviewers to code is slow and prohibitively expensive (Liu, 2006). Mechanical analysis using natural language processing (NLP) tools is faster and cheaper, but still requires considerable manual intervention. As a result, repeatability across contexts (e.g., product categories, markets, languages) and scalability as the number of reviews grows remain difficult challenges. The fact that the set of consumer reviews changes constantly makes the data especially valuable in dynamic markets, such as those with many product entries and exits, but also implies that largely automated and highly scalable approaches are critical in order to permit frequent analyses.

In this paper, we develop a model for forecasting product sales by a major online retailer using historical data on sales ranks for existing products, prices and consumer reviews. Products are represented as bundles of attributes, and the information contained in consumer reviews is conceptualized as stock variables, similarly to advertising or goodwill stock. The model is used to predict sales of new products that are introduced into the category, as well as to make rolling horizon sales predictions of existing products. To deal with the challenge of incorporating the textual content of customer reviews, we propose using a bag-of-words with a random projections model. This 'bag-of-words' is a feature extraction approach in which text (such as a sentence or a document) is represented as the collection of its words, ignoring grammar and even word order but keeping multiplicity. Thus, in its simplest form, the count of each word in the bag-of-words becomes a predictor in a forecasting model. In a more complicated form, related approaches use sequential or non-sequential sequences of words in a document as predictors, and our forecasting model is scalable and extends easily to this form.

A key characteristic of the bag-of-words approach is that the parsing procedure is very simple, allowing the preprocessing of the data to be almost entirely automated. In the preprocessing step, terms are filtered and manipulated automatically in order to remove terms that do not contain content, such as stop words, numbers, punctuation marks, or very small words, and to remove endings based on declination or conjugation by applying stemming. The ease of this pre-processing is a particularly significant advantage in the context of new products, because early information received via customer reviews can be included in the forecasting model rapidly in order to improve the accuracy and speed of forecasts. This advantage is also important when scaling a forecasting model across languages (countries, for instance) for a given product category, or to new categories.

Many of the extant approaches to the textual analysis of consumer reviews rely on the extraction of product features and evaluative phrases (e.g., in their analysis of digital cameras, Archak et al., 2011, extract pre-specified features such as battery life, design, and picture quality, and popular opinion phrases for these product features, like good, very good and great), whereas a considerable part of consumer reviews contains very little or no text about product features. For instance, the following two sentences from a customer review would largely be disregarded by many extant approaches because product features are not referred to explicitly: "This product is without doubt the best purchase I have made in years. I love it". In contrast, a bag-of-words method would consider such text in the model.

However, a key concern with the bag-of-words and related approaches is the fact that the number of words or sequences of words in the bag can reach tens of millions, thus exploding the dimensionality of the predictor matrix. The count of unique words in the English language is in the hundreds of thousands, while commonly used words number in the tens of thousands. When commonly used unique words and their combinations, such as sequential and non-sequential pairs, are counted, the number of predictors can exceed one trillion (i.e., the number of commonly-used words squared), each of which is probably very sparse, and much greater than the sample size of data. Returning to the two-sentence review referenced above, when parsed, this text contains 15 unique words, 15 sequential pairs, 105 non-sequential pairs (i.e., ${}_{15}C_2$), 14 sequential triplets, and 455 non-sequential triplets (i.e., ${}_{15}C_3$). This explosion leads to multiple problems. One, as has been noted, the number of predictors exceeds the available degrees of freedom (which is the number of reviews or the number of observations of product-sales). Two, even if an adequate number of degrees of freedom are available, it becomes computationally infeasible to estimate standard models because of such a high-dimensional predictor matrix. Third, forecasting out-of-sample data based on predictors that exist in only one or a few reviews would be likely to lead to over-fitting issues.

To overcome these problems, we propose using random projections to reduce the dimensionality of the bag-of-words in the sales forecasting model. In random projections, the original high-dimensional data are projected onto a lower-dimensional subspace using a random matrix with specific characteristics (Bingham & Mannila, 2001). In its assessment of alternative approaches to dimensionality reduction, the Committee on the Analysis of Massive Data (National Research Council of the National Academies, 2013) labels random projections approaches "*data oblivious*", in that the dimensionality reduction mapping can be computed without any knowledge of or use of the data. This is in contrast to "*data aware*" methods such as principal components analysis and its refinements, where the mapping is dependent on a given dataset. The report also identifies key benefits of random projections as follows (p. 77): "…the projection is guaranteed to work (in the sense that it preserves the distance structure or other properties) for arbitrary point-sets. In addition, generating such projections requires very little resources in terms of space and/or time, and it can be done before the data are even seen. Finally, this approach leads to results with provable guarantees". In Section 3, we discuss these "provable guarantees" in relation to preserving the distances in the original high-dimensional predictor matrix. The benefits of random projections highlighted in this quote are quite consistent with our goal of creating a fast, scalable, and efficient sales forecasting model, especially for new products.

The use of random projections in predictive models is just emerging in the machine learning literature (e.g., Fard, Grinberg, Pineau, & Precup, 2012 and Maillard & Munos, 2012), and, to the best of our knowledge, this paper is its first application to business forecasting. Although this paper uses random projections to reduce the dimensionality of the predictor matrix in a regression-based forecasting model, other statistical models such as support vector machines, neural networks, and k-nearest neighbors could also be used when appropriate. As a comparison with our proposed model, we implement a support vector machine with a bag-of-words approach. We also discuss aspects of alternative models in the concluding section of the paper.

Our use of a bag-of-words approach combined with random projections, while advantageous for the particular application at hand, is not without its costs. Since the effects of the textual content of consumer reviews on sales are not specified based on theory or prior knowledge, and are also transformed via random projections, they are not interpretable like most dimension reduction approaches. One defense of this approach lies in the managerial motivation for our model, which is primarily forecasting, not description or understanding. A related concern arises because of the fear of over-fitting to the estimation sample, given the manner of model specification (discussed later in the paper). To overcome this concern, we use holdout samples carefully and rigorously for model selection, trying to ensure that the best predictive model is retained in the process.

The rest of this paper is organized as follows. In Section 2, we describe the tablet computer dataset that we use for our empirical application. Section 3 presents the proposed model. In Section 4, we discuss results, including model validation. We conclude in Section 5 with conclusions and ideas for further research.

2. Data

We use the dataset titled "Market Dynamics and User Generated Content about Tablet Computers" that is available via *Marketing Science* (Wang, Mai, & Chiang, 2013). The dataset contains product information and consumer review data collected from Amazon.com and from the websites of various tablet manufacturers.[2] While a full description of the data gathering methodology is provided by Wang et al. (2013), we provide some relevant information here. Market dynamics data (e.g., sales ranks, prices, consumer reviews) were gathered via a Java web crawler once a week over the 24-week period from February 1 to July 11, 2012, for a total of 2163 products. Product characteristics were gathered for the subset of 794 products for which at least one consumer review was posted. The market dynamics data are not always complete (there are occasional missing prices and sales ranks) for a variety of reasons. One reason is that since "tablet computers" was a relatively new product category in early 2012, Amazon introduced many new products during the 24-week time period; similarly, several existing products became unavailable on Amazon during this window.

For our analysis, we focus on a subset of 231 tablet computers (out of 794) for which sales ranks and prices (discussed subsequently) are available for each week of the 24-week period. As an additional validation sample, we also choose 73 tablet computers that were introduced during weeks 7–24 of the 24-week period. Descriptive statistics of the data are provided in Table 1.

The tablet computer dataset that we use contains a sales rank for each of the 2163 tablet computers for each week; this variable is defined in the documentation (Wang et al., 2013, Table 1, Panel A) as "Sales Rank in the Tablets and Tablet PCs category". While Amazon.com does not report actual sales of individual products on its website for competitive reasons, Amazon does report sales ranks. This is a very important feature of the retailer's service and has been used in a large number of academic research studies; it has also spawned a mini-industry around the tracking and reporting of sales ranks for use by authors, publishers, manufacturers and resellers.

A careful scrutiny of the sales ranks available in the tablet data and a study of Amazon.com's definitions of the sales rank information revealed that the sales ranks displayed can be interpreted as follows. Individual tablet computers on Amazon are categorized primarily in the hierarchy Electronics → Computers and Accessories → Tablets. At any point in time, consider all tablet products

[2] As defined by Wang et al. (2013), "a tablet computer is a one-piece mobile computer. Devices typically have a touchscreen, with finger or stylus gestures replacing the conventional computer mouse".

Table 1
Descriptive statistics of the tablet computer data (existing and new products).

	Observations	Mean	Std. Dev.	Min	Max
Log sales rank	6,190	4.6	0.96	0	5.7
Price ($)	6,190	$310.30	$288.76	$49.99	$1925.00
Consumer reviews:	33,507				
Numerical rating (stars)		3.87	1.4	1	5
Number who found review helpful		5.19	29.0	0	998

at Amazon as being ordered by rank in descending order of sales numbers. For the highest selling r_1 tablet products, the displayed sales ranks are in Tablets; for the next r_2 tablet products, the displayed ranks are in Computers and Accessories; and for the lowest selling r_3 products the displayed ranks are in the broadest category, Electronics. r_1, r_2, and r_3 are not revealed. Thus, the sales ranks are not comparable across tablet products directly, since they may be based on different comparison groups. As an illustration, the lowest-ranked tablet product in our entire dataset had a sales rank of 2064,329, which is obviously not a legitimate rank within the Tablets category, but is a legitimate rank in Electronics.[3]

For our analysis, we chose to convert the sales ranks monotonically into modified sales ranks using the following transformation. In each week, we sorted the 304 products (231 existing plus 73 newly introduced products) based on the sales ranks available in the dataset. New products that had not yet been introduced or did not have a sales rank were omitted. We then assigned a modified sales rank sequentially, starting with rank 1 being assigned to the highest ranked product. As an example, imagine that in week t, 231 existing products and 50 new products had sales ranks in the dataset. We therefore assigned these products modified ranks of 1–281. Note that this transformation is monotonic and preserves the rank-ordering in the sales ranks displayed. Furthermore, the transformed ranks are comparable across products and are more reasonable measures of the rankings within the tablets category. For robustness, we also estimated our models using the displayed sales ranks, and found that, while the magnitudes of the parameter estimates changed, the substantive conclusions did not.

For each product, the data provide multiple prices, including a retail price suggested by the manufacturer (list price), the current selling price of the product if purchased from Amazon (Amazon price), and the prices of new and used versions of the product provided by merchants for whom Amazon serves as a marketplace. We focus only on new (i.e., not used) versions of products, and, taking the perspective of the consumer, we elect to use the lower of the Amazon price and the lowest new price quoted by merchants in each week. This weekly price is used in our model only if the product had a published sales rank in the same period. Table 1 shows that tablet prices ranged from $49.99 to $1925.00 across all products and weeks.

An important aspect of the data is the full set of consumer-generated reviews of each product, as of the date of data collection. The total number of reviews in the dataset is 33,507. Each review is characterized by a numerical rating of between one and five stars, a title, and the textual content of the review. Table 1 shows that the average rating across reviews is as high as 3.87 stars, exemplifying a positively skewed distribution, as has also been reported in the previous literature (for instance, Ho-Dac, Carson, & Moore, 2013, report mean ratings of 3.88 for Blu-Ray players and 3.62 for DVD players). In addition, a count of the number of readers of the review who clicked “Yes” in answer to the binary question “Was this review helpful to you?” is displayed as “X out of Y people found the following review helpful”. The mean number of helpful votes per review is 5.19 and ranges from 0 to almost 1000, showing a significant degree of variance across reviews.

The raw data in the title and textual content of reviews is pre-processed by a relatively simple parsing program that includes a filter for special characters (e.g., “.” or “,”) and a corpus for the English language. First, all textual content in the title and body of reviews is converted to lowercase, and special characters are removed. Then, after removing certain words from a stop list (e.g., “a” or “I”), the parsing program counts the number of times that each word occurs in a review. After pre-processing, the total number of unique words across all reviews in our dataset is 20,068. As was noted previously, adding word combinations would increase this number many-fold; the current application does not do so.

We consider each tablet computer to be a bundle of attributes, and represent it as such in the model. This is especially important because of our goal of using the model to forecast sales of new products, each of which can be described as an attribute bundle. This approach is consistent with the large body of literature on conjoint analysis (e.g., Green & Rao, 1971) and SKU-level choice modeling (e.g., Fader & Hardie, 1996). Using the manufacturer-provided product characteristics information included in the dataset, we create a schema where four attributes are represented as discrete and four as continuous. This process requires category knowledge and some subjective judgment. In Table 2, we show the attributes and levels for discrete attributes and the numbers of products in our data that possess each of the attribute levels, as well as descriptive statistics of the continuous attributes.

While we construct a relatively simple categorization for the discrete attributes manually, Lee and Bradlow

[3] Previous studies that have used sales rank data displayed at Amazon.com appear to have experienced this same concern, although it has not been highlighted. For instance, Table 2 of Ghose and Ipeirotis (2011) reports the maximum sales rank of audio and video players in their dataset as 2090,308. Since all of the products included in their study were picked from Amazon’s “top-100” list (p. 1500), this rank must be in a higher-level category than audio and video players. The same phenomenon holds for digital cameras and DVD players in their data, where the maximum sales ranks are shown as 38,353 (Table 3) and 49,868 (Table 4) respectively, though the products were top-100 products.

Table 2a
Discrete product attributes, their representation, and numbers of products in the sample.

Discrete attributes	Attribute levels Number of products in sample possessing attribute level Dummy variable representation, and parameter				
Operating system	Android 247	Apple 23 X^{OS}_{Apple}; δ^{OS}_{Apple}	Blackberry 3 X^{OS}_{Black}; δ^{OS}_{Black}	HP 2 X^{OS}_{HP}; δ^{OS}_{HP}	Windows 29 X^{OS}_{Win}; δ^{OS}_{Win}
Battery life (hours)	1–3.99 31	4–6.99 119 X^{BL}_{4+}; δ^{BL}_{4+}	7–9.99 97 X^{BL}_{7+}; δ^{BL}_{7+}	10 or more 57 X^{BL}_{10+}; δ^{BL}_{10+}	
Screen size (diagonal in inches)	Low (4–7.99) 143	Medium (8–9.99) 62 X^{SS}_{Med}; δ^{SS}_{Med}	High (10+) 99 X^{SS}_{High}; δ^{SS}_{High}		
RAM	Low (256 MB, 512 MB) 193	High (1 GB or more) 111 X^{RAM}_{High}; δ^{RAM}_{High}			

Table 2b
Continuous product attributes and descriptive statistics in the sample.

Continuous attributes	Variable and parameter	Mean	Std. Dev.	Min.	Max.
Processor speed (GHZ)	X^{Proc}; δ^{Proc}	1.068	0.331	0.300	2.700
Storage (GB)	$X^{Storage}$; $\delta^{Storage}$	27.2	67	0.5	640
Weight (pounds)	X^{Weight}; δ^{Weight}	1.39	0.84	0.40	8.00
Pixels per inch (PPI)	X^{PPI}; δ^{PPI}	137.80	21.47	84.77	293.70

(2011) present an approach that automates this process and constructs more detailed schema. Since some of the attributes may be considered as vector attributes (i.e., "more is better" attributes), we are able to form expectations of the signs of the estimated parameters in order to assess the face validity of the model. The vector attributes are battery life (positive effect on sales; that is, sales are expected to increase as the battery life increases), screen size (positive), RAM (positive), processor speed (positive), storage (positive), weight (negative), and pixels per inch (positive). Due to the difficulty of developing a parsimonious coding scheme, we do not include the brand as an attribute (the number of brands is very large), and also omit minor attributes like webcam resolution and wireless type. Since only a subset of attributes is included in the model, an omitted variable bias may have an untoward effect on the estimated coefficients of the included variables.

3. Model

Our model is specified to explain the variation in $-\log(s_{jt})$, where s_{jt} is the sales rank of product j in week t, for products $j = 1, \ldots, J$ and over weeks $t = 1, \ldots, T$:

$$-\log(s_{jt}) = \alpha + \mathbf{X}_j'\boldsymbol{\delta}_j + \beta_{trend}X_{time} + \beta_{price}\ln(p_{jt}) + \beta_{RS}RS_{jt} + \boldsymbol{\beta}_{BOWS}\mathbf{BOWS}_{jt} + \varepsilon_{jt}. \qquad (1)$$

As was noted earlier, the use of sales ranks instead of sales is necessitated by the lack of availability of sales data from Amazon. The use of sales ranks instead of unit sales relies on the assumption that sales ranks follow a Pareto distribution. This assumption was used cleverly by Chevalier and Goolsbee (2003) to derive an empirically estimable log-linear relationship of the form $\log(sales) = a + b\log(sales\ rank) + \varepsilon$. Since then, multiple studies have shown that this relationship fits the data well in categories such as books, software and electronics (e.g., Brynjolfsson, Hu, & Smith, 2003). Although the parameters a and b are unknown for the tablet computers category, to the extent that $\log(sales)$ is simply a linear transformation of $\log(sales\ rank)$, we can use the latter as the dependent variable in a predictive model as a substitute for the former (Archak et al., 2011). However, this implies that the estimated coefficients of predictor variables in the model in Eq. (1) cannot be interpreted directly as marginal effects on $\log(sales)$, since the coefficients and standard errors are scaled by the unknown parameter[4] $1/b$. Since higher ranks mean lower sales, our use of a negative sign in the dependent variable in Eq. (1) facilitates easier interpretation of the signs of parameter estimates. For instance, a positive estimated coefficient in Eq. (1) implies a negative effect on the sales rank, or a positive effect on sales.

The term $\mathbf{X}_j'\boldsymbol{\delta}_j$ captures the effects of the characteristics of product j. $\boldsymbol{\delta}_j$ is a (14×1) parameter vector that represents the effects of discrete and continuous attributes, and $\mathbf{X}_j$ is a (14×1) vector that represents information on product attributes, as defined in Table 2. The four continuous attributes (Table 2b) are introduced into the model in log form to allow for the coefficients to be interpreted as elasticities. The base level of the four discrete attributes (leftmost level in Table 2) is absorbed into the intercept α. We include X_{time} as a linear time-trend. Turning to product- and week-specific explanatory variables, p_{jt} is the price,

[4] Since the unknown scaling parameter applies to all coefficients, ratios of coefficients are meaningful. For instance, a willingness to pay for a product feature can be estimated as the price coefficient divided by the coefficient of the product feature.

RS_{jt} is the rating stock based on customer reviews, and $\mathbf{BOWS}_{jt}$ is a d-dimensional vector of bag-of-words stock of customer reviews. Including the price as $\ln(p_{jt})$ allows its coefficient β_{price} to be interpreted as the price elasticity of the sales rank. Next, we discuss the variables that capture the information contained in customer reviews.

In week t, prospective customers in the category have access to a stock of information contained in customer reviews. This includes *all* reviews that were ever written on Amazon for products in the category, which can be a very voluminous body of information. As a result, it is unlikely that all reviews for a given product will be weighted and used by consumers in their decision making equally. We conceptualize that prospective customers weight reviews differently based on two factors: (i) the helpfulness of the review as measured by the number of helpful votes it received, and (ii) the recency of the review.[5] For instance, customers on Amazon have the opportunity to sort reviews based on the numbers of helpful votes received. Furthermore, to account for the effects of past and current reviews, we employ unobservable "stock" variables that are similar to goodwill or advertising stock, consistent with the large econometric literature on the modeling of advertising effects. Our conceptualization differs from that of most previous research, which emphasizes the role of average ratings (e.g. Chevalier & Mayzlin, 2006) or the variance of the ratings (e.g., Archak et al., 2011 and Sun, 2012). In particular, we use the numerical ratings in customer reviews to construct a Ratings Stock (RS) and the textual content of reviews to construct a Bag-of-Words Stock (BOWS). The values of these stock variables in any given week result from current week reviews[6] and a depreciated value of the stock from the previous week.

Define A_{j0} as the set of user-generated reviews about product j that have been posted up to week 0, where week 0 is the week just before the beginning of the 24-week time window. Define $\tilde{A}_{jt}$, $t = 1, \ldots, T$ as the set of reviews that are posted any day in week t. Each review $r \in A_{j0}$ or $r \in \tilde{A}_{jt}$ can be characterized by the vector $X_r^1 = \{rating_r, Text_r\}$ and $\#\,helpful_{rt}$, where $rating_r$ is the 5-star numeric rating of the product in review r, $Text_r$ is a bag-of-words representation of the title and text of review r, and $\#\,helpful_{rt}$ is the number of helpful votes received by review r till week t. The Ratings Stock for product j in week t is based on the exponential decay process of Nerlove and Arrow (1962), as follows:

$$RS_{jt} = \lambda\left(RS_{j(t-1)}\right) + (1-\lambda)\sum_{r\in\tilde{A}_{jt}} rating_r \cdot w_{rt},$$

$$t = 1, \ldots, T$$

$$RS_{j0} = \sum_{r\in A_{j0}} rating_r \cdot w_{r0},$$

where $w_{rt} = \frac{\#\,helpful_{rt}}{\sum_{r\in\tilde{A}_{jt}} \#\,helpful_{rt}}$, $w_{r0} = \frac{\#\,helpful_{rt}}{\sum_{r\in A_{j0}} \#\,helpful_{rt}}$, and $0 < \lambda < 1$ is a carryover parameter to be estimated.

We construct the Bag-of-Words Stock as follows. We define a set of words $L = \{l_i | i = 1, \ldots, d\}$. $Text_r$ is then used to build a d-dimensional vector V_r for review r such that $v_r(i) = \#\,(Text_r, l_i)$, $i = 1, \ldots, d$, where $\#\,(Text_r, l_i)$ is the number of times that word l_i appears in $Text_r$. Analogous to the definition of Ratings Stock, we can now define $\mathbf{BOWS}_{jt}$ as a d-dimensional vector whose ith element is as follows:

$$\begin{aligned} bows_{jt}(i) &= \lambda(bows_{j(t-1)}(i)) + (1-\lambda) \\ &\quad \times \sum_{r\in\tilde{A}_{jt}} v_r(i)\cdot w_{rt}, \quad t = 1, \ldots, T \end{aligned}$$

$$bows_{j0} = \sum_{r\in A_{j0}} v_r(i)\cdot w_{r0}.$$

3.1. Random projections

In typical applications, d is counted in the tens of thousands or even millions. Let's represent the number of predictor variables by $d^+ \geq d$. The number of observations of sales ranks is $n = J * T$, where J is the number of products and T is the number of weeks. Since typically $n \ll d^+$, it is not possible to regress sales on the predictor variables directly. In our application to tablet computers, if we use the full estimation sample of 231 computers, and the full 24 weeks of data, $n = 231 \times 24 = 5544$. The number of unique words in the textual content of consumer reviews is $d = 20,068$. In fact, as we demonstrate subsequently, we use at most 12 of the 24 available weeks of data in our model estimation, in order to have an adequate validation sample, which further exacerbates the problem of not having sufficient data to estimate the full model.

Even if the sample size based on the numbers of products and weeks allows sufficient degrees of freedom, it is often not feasible to estimate the full model because of the implied memory or computational resource requirements. For instance, OLS estimation of the model in Eq. (1) on our data would require the inverting of $(\mathbf{X}'\mathbf{X})$, which is of dimensionality d^+ by d^+. We found that it was not possible to perform this computation on a 64-bit operating system with 16 GB of RAM after artificially increasing the number of observations above d^+ for non-singularity in $\mathbf{X}$. Furthermore, if we were to extend the bag-of-words model to include non-sequential word pairs, this issue would grow exponentially.

We propose the use of low-dimensional random projections (Bingham & Mannila, 2001) in order to overcome these dimensionality problems in a computationally simple, fast and flexible manner. The basic idea in dimensionality reduction is to project the data onto a lower-dimensional subspace that captures as much of the variation as possible. While Principal Component Analysis (PCA) has been the traditional approach for accomplishing this, the computational complexity of the technique is $O(d^3)$ and the memory complexity of the technique is $O(d^2)$, which make the technique infeasible for the size of d involved in our application (van der Maaten, Postma, & van den Herik, 2009). In random projections, the original d-dimensional $\mathbf{X}$ matrix is projected onto a k-dimensional subspace, $k \ll d$, using a random $d \times k$ matrix $\mathbf{R}$ whose columns have unit lengths, to obtain $\mathbf{X}^{RP}$ as follows:

$$\mathbf{X}^{RP}_{n\times k} = \mathbf{X}_{n\times d}.\mathbf{R}_{d\times k}. \qquad (2)$$

[5] These are also the two factors that are used by Amazon to determine its rankings of reviewers, in addition to the number of reviews written by each reviewer. See http://www.amazon.com/gp/customer-reviews/guidelines/top-reviewers.html.

[6] We discuss the concern of potential endogeneity of reviews subsequently.

The key idea of random projections is based on the Johnson–Lindenstrauss lemma (Johnson & Lindenstrauss, 1984), which says that if the points in a vector space are projected onto a randomly selected subspace of a suitably high dimension, then the distances between the points are approximately preserved.

There is a substantial theoretical body of literature on the value of random projections for dimensionality reduction, as well as some empirical applications, primarily non-predictive. Recent work in computer science has explored the theoretical properties of least squares estimators in reduced dimension subspaces generated via random projections. For instance, Fard et al. (2012) theoretically compute the worst-case prediction error of the OLS estimator in the space induced by random projections when the **X** matrix is sparse. An important finding (Theorem 4, p. 1056) is that the bound on the error reduces to a bias–variance trade-off as a function of the projection size. Increasing the size of the projection reduces the approximation error of the OLS estimator but introduces error due to an increase in the variance of the estimated parameters. In numerical experiments, they find that, under certain conditions on the variance of the model error, it is better to use random projections of an intermediate size in order to optimize the bias–variance trade-off. This concern is similar to that of over-fitting for forecasting applications (De Gooijer & Hyndman, 2006); hence, relying on out-of-sample validation instead of in-sample model fit is critical.

The highly dynamic and voluminous nature of web-based customer review data and the consequent need for repeated applications of predictive models make the benefits of computational speed and simplicity that are afforded by random projections particularly valuable. In addition, the random projections approach can also be updated dynamically as each new textual review is introduced, since the computation is a series of dot products and the dot product has a distributive property (i.e., $(Text_r + Text_{r*})R = Text_rR + Text_{r*}R$). Therefore, product managers can revise their forecasts with a minimal cost each time a new textual review is introduced, which is especially important for high velocity data.

As is evident from Eq. (2), the computation of the random projections matrix $\mathbf{X}^{RP}$ is very simple. The choice of k is guided by considerations of computational costs, as well as by the choice of an appropriate random distribution for **R**. For computational costs, the expense of inverting $\mathbf{X}^{RP\prime}\mathbf{X}^{RP}$ through Cholesky factorization is $O(k^3)$, multiplying $\mathbf{X}^{RP\prime}$ by **Y** gives $O(n \times k)$, and multiplying $\mathbf{X}^{RP\prime}$ by $\mathbf{X}^{RP}$ gives $O(n \times k^2)$. All of these computations are necessary in order to generate least squares estimates, and if k is chosen to be small enough, the Cholesky factorization will no longer impose the greatest computational burden, which is a key advantage of the random projections approach.

The choice of the distribution for **R** affects the ability of the random projections to preserve the pairwise distances of **X**, and also influences the processing cost of multiplying **X** by **R**. A standard normal **R** with independent and identically distributed entries not only preserves the Euclidean pairwise distances between any two rows and the Euclidean norms of individual rows with probabilities that depend[7] on k, but also ensures that the expectation of the projected Euclidean norm of a row in $\mathbf{X}^{RP}$ is equal to the original Euclidean norm of that same row in **X** (Arriaga & Vempala, 2006). A Cauchy **R** preserves absolute pairwise distances (i.e., the Manhattan distance) (Li, Samorodnitsky, & Hopcroft, 2013). In order to speed up processing by a factor of $\sqrt{d}$, Li et al. (2006a) recommend using a sign random distribution that generates entries $\{-1, 0, 1\}$ with probabilities $\left\{1/2\sqrt{d},\, 1-1/\sqrt{d},\, 1/2\sqrt{d}\right\}$. This projection preserves Euclidean pairwise distances with only a small accuracy loss compared to standard normal random projections. Sign random distributions are especially important in applications where d is in the tens of millions (e.g., word pairs), because of the savings in processing costs.

3.2. Simplified example

We use a highly simplified example to illustrate the use of random projections for reducing the dimensionality of the bag-of-words matrix and for predictions using regression. Consider two products – a good tablet (Product A, rank 1) and a bad tablet (Product B, rank 2) – that have three consumer reviews each. Assume that each review contains only the words "tablet", "good", or "not", as generated from our corpus of the English language. The bag-of-words representation, $\mathbf{X}_{6\times3}$, shows the number of times that a given word appears in a particular review.

To reduce the dimensionality of $\mathbf{X}_{6\times3}$, we multiply it by a standard normal random matrix $\mathbf{R}_{3\times2}$ to generate $\mathbf{X}^{RP}_{6\times2}$. To illustrate the effect of randomness in this example, we consider two randomly generated values of $\mathbf{R}_{3\times2}$, to obtain two different values of $\mathbf{X}^{RP1}_{6\times2}$ and $\mathbf{X}^{RP2}_{6\times2}$ (not shown).

Entries of $\mathbf{X}_{6\times3}$			
Review	"tablet"	"good"	"not"
Product A	1	1	0
Product B	1	1	1
Product A	2	2	0
Product B	2	1	2
Product B	0	0	4
Product A	0	4	0

When the actual ranks of the products $\mathbf{Y}_{6\times1}$ shown in Table 3 are regressed on $\mathbf{X}^{RP1}_{6\times2}$ and $\mathbf{X}^{RP2}_{6\times2}$ respectively, widely different estimates of the regression coefficients may be obtained (also not shown). However, Table 3 shows that the predicted ranks, which are linear combinations of these coefficients and the specific projected values in $\mathbf{X}^{RP1}_{6\times2}$ and $\mathbf{X}^{RP2}_{6\times2}$, are consistent with both the actual ranks and the predicted ranks obtained from the original higher-dimensional $\mathbf{X}_{6\times3}$. In fact, the rank order of the predicted ranks would match the rank order of the actual sales ranks perfectly, thereby creating a perfect gains chart. This example was repeated using log(ranks), and exactly the same substantive results were obtained.

[7] See Eq. (9) of Li, Hastie, and Church (2006a) for the worst-case probability of this assurance failing among any two rows.

Table 3
Predictions of sales rank using original and reduced dimension X.

	Actual sales ranks $\mathbf{Y}_{6\times1}$	Predicted ranks $\hat{\mathbf{Y}}_{6\times1}$ using $\mathbf{X}_{6\times3}$	Predicted ranks $\hat{\mathbf{Y}}_{6\times1}$ using $\mathbf{X}^{RP1}_{6\times2}$	Predicted ranks $\hat{\mathbf{Y}}_{6\times1}$ using $\mathbf{X}^{RP2}_{6\times2}$
Product A	1	1.20	1.44	1.29
Product B	2	1.47	1.57	1.51
Product A	1	1.29	1.18	1.25
Product B	2	1.87	1.62	1.80
Product B	2	2.20	2.21	2.22
Product A	1	0.97	0.98	0.94

3.3. Endogeneity concerns

A few different kinds of endogeneity concerns may arise from the use of consumer reviews as exogenous predictors of product sales in our model specification. One is that both the volume and valence of consumer reviews may be influenced by product sales. As the number of product buyers grows, the set of possible reviewers, and hence the number of consumer reviews, grows. Similarly, the content of reviews may be affected by the degree of product success, as reflected in sales. A second kind of endogeneity concern may be due to omitted variables. Say that there are product-specific, time-varying influences on both sales and consumer reviews that are omitted from the model; one example of such influences is product advertising. Both of these forms of endogeneity would bias the estimated effects of consumer reviews.

To examine the effects of possible endogeneity, we estimated a model in which the Ratings Stock and Bag-of-Words Stock variables were formulated using only ratings lagged by one week. Thus, contemporaneous reviews were not included. We found that our model estimates were essentially unaffected. This is not too surprising, since only a small number of new reviews are written in any given week, and as we report later, we find a high carry-over effect of past reviews on current sales.

3.4. Estimation strategy

Our goal is to identify separate models that provide the best rolling horizon one-week-ahead forecasts of sales ranks for existing products and new products respectively, and we develop our estimation strategy accordingly. Importantly, forecasting for new products using historical data on existing products is a significantly more challenging task than one-week ahead forecasting for existing products. Thus, we expect systematic differences between the best models for the two tasks, as well as between their predictive performances.

The model in Eq. (1) can be estimated using Ordinary Least Squares (OLS), conditional on:

(i) an estimate of the carryover parameter λ, which enters the model non-linearly; and
(ii) the choice of k, which is the number of columns in $\mathbf{R}$.

For the analysis reported in this paper, we specify $\mathbf{R}$ to be standard normal, but discuss alternative specifications in the concluding section of the paper. As has been noted, the machine learning literature provides conflicting guidance as to the best choice of k. Across values of k, there is a tradeoff between theoretically preserving the properties of $\mathbf{X}$ and practically over-fitting the model in the forecasting task. We select optimal values of λ and k based on a grid search in which we evaluate the forecasting performances of the model in predictive holdout tasks using three different measures of the predictive error.

All three measures of the predictive error are based on comparisons of the actual sales ranks with the predicted sales ranks. It is important to point out that since the parameters of the models are estimated in logarithmic space, predictions are obtained in logarithmic space initially, then transformed to sales ranks by exponentiation. Goldberger (1968) pointed out that such predictions are biased downwards, but can be corrected by multiplying the prediction by $\exp(1/2 * MSE)$, where MSE is the estimated mean square error of the model. The three measures we use are described next.

First, we use the root mean squared error (RMSE) which is on the same scale as the data inputted into the model and is of theoretical relevance to the statistical model (i.e., OLS minimizes the squared error, which has a monotonic relationship with RMSE) but is more sensitive to outliers (Hyndman & Koehler, 2006). For example, a forecasted sales rank of 2 for a product with an actual sales rank of 1 would have a squared error of 1, while a forecasted sales rank of 4 for a product with an actual sales rank of 1 would have a squared error 9 times greater, thus contributing a greater share to the RMSE. Second, we use the mean absolute percentage error (MAPE), which is scale independent (Kolassa & Schütz, 2007) and has the ability to compare forecast performances across data sets (Hyndman & Koehler, 2006), but is bad with forecasts that are below the actual values, due to a maximum deviation of 100% (Armstrong & Collopy, 1992). For example, a forecasted sales rank of 2 for a product with an actual sales rank of 8 would have a 75% error, while a forecasted sales rank of 8 for a product with an actual sales rank of 2 would have a 300% error. Note that MAPE is more consistent with an estimator that estimates the conditional median, whereas we use OLS, which estimates the conditional mean. Third, we use the correlation, which measures the strength and direction of the relationship between the forecasted and actual values.

We divide the sample dataset of 24 weeks and 304 products (231 existing products and 73 new products) into five subsamples, as depicted in Table 4, and identify the primary use of each sample in the last row of the table. Since new products enter the market only in week 7 or later, these products have no initialization sample (IS). IS is first used to initialize the values of the Ratings Stock and BOW Stock variables, then combined with the

Table 4
Five subsamples for model estimation and validation.

	Weeks 1–6	Weeks 7–12	Weeks 13–24
Existing products (EP)	Initialization sample (IS)	Estimation sample (ES_EP)	Validation sample (VS_EP)
New products (NP)		Estimation sample (ES_NP)	Validation sample (VS_NP)
Primary use of samples	Initialize ratings stock and BOW stock variables	Select model based on predictive performance in running horizon, one-week-ahead task	Assess forecasting performance of full model relative to baseline model

Estimation Sample–Existing Products (ES_EP) to serve as the estimation sample. For each value of λ and k in a grid, we estimate the model in Eq. (1) iteratively to produce rolling horizon one-week-ahead predictions.

The model selection is done as follows:

1. Model for existing products: we estimate the model six times successively using data for existing products for weeks 1 to T, where $T = 6, 7, 8, 9, 10$, and 11. Each set of model estimates is then used to predict sales ranks for existing products for week $T+1$ using $\mathbf{X}_{j(T+1)}$. Thus, six sets of forecasts are obtained, for weeks 7, 8, 9, 10, 11, and 12 respectively. The three measures of prediction error are computed and used for model selection (discussed further in Section 4).
2. Model for new products: the same estimation process is followed as for existing products, but predictions are made for the set of new products that are available in the Estimation Sample—New Products (ES_NP) subsample. Importantly, sales ranks for new products are never used in model estimation, but $\mathbf{X}_{j(T+1)}$ for the new products is used in forecasting. Once again, the three error measures are used for model selection.

Once we have selected the best forecasting models for existing and new products, we also estimate in each case both a "baseline" model and a competitor model, which is a Support Vector Machine (SVM). The baseline model is the model in Eq. (1) excluding $BOWS_{jt}$. The SVM is a regression machine and includes all of the covariates in Eq. (1) and $BOWS_{jt}$. The SVM is intended to serve as a comparison to our full model, since it uses the same data but a different analysis approach. The performances of the models selected for existing and new products are assessed on the validation samples VS_EP and VS_NP respectively, through a comparison with the performances of the baseline models and SVMs.

We considered the use of alternative competitor methods such as sentiment analysis, but did not find them suitable for this study because they typically require additional data such as a lexicon of sentiment words and phrases, parts-of-speech tagging, and so forth. While these supplemental data can provide additional information, particularly in the form of the valence of words contained in the bag of words, they are often domain- and context-specific, and therefore costly to gather. This would compromise some of the key advantages of our approach: fast, scalable, efficient, low resource requirements. Moreover, it would mean that the results were not comparable to ours. In contrast, SVM is comparable to the proposed method from the perspective of the cost to a business, in that no domain-specific data need to be collected. It is a technique that is cited widely in the literature on textual analysis, and also allows for a sparse data matrix. It should be noted that we had to upgrade the RAM on our computer from 16 to 32 GB to enable us to implement the SVM computations, and that the fitting of the SVM model took over an hour, compared to a few minutes for our random projections model for a given k.

4. Results

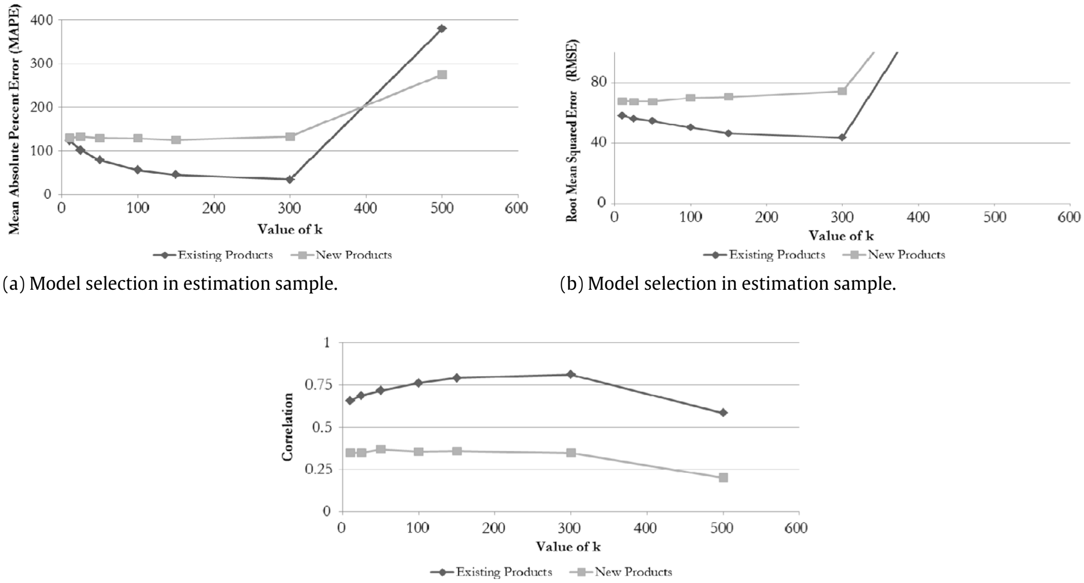
4.1. Model selection

The computational cost of a grid search in the joint space of λ and k is higher for individual values of λ than for individual values of k because the bag-of-words rating stock changes for each value of λ. Consequently, we used a coarse grid for λ and a finer grid for k. The optimal value of λ was found to be 0.90, implying that 90% of the previous week's rating stock carries over into the current week. In Fig. 1, we show the performance of the full model in Eq. (1) for forecasting existing and new products in ES_EP and ES_NP respectively, for different values of k. All results were averaged across 10 simulations of each k.

In accordance with our previous intuition, forecast performances were better for existing products than for new products across all values of k in our estimation sample. The values of k that had the best overall predictive performances in ES_EP and ES_NP were $k = 300$ and $k = 50$, respectively. The only exception was the MAPE of $k = 50$ (129.3%), which came in a close third place after $k = 150$ (124.9%) and $k = 100$ (128.6%) for ES_NP, as Fig. 1(a) shows. We also ran the full model in Eq. (1) using values of $k = 750$ and 1000; however, we found that the model could not be computed due to near-singularities in the $\mathbf{X}^{RP}$ matrix. For validation, we used $k = 300$ and $k = 50$ for VS_EP and VS_NP, respectively.

4.2. Parameter estimates

In Table 5, we show parameter estimates for the baseline model, the full model for existing products for $k = 300$, and the full model for new products for $k = 50$, estimated on the full 24 weeks of data. We do not show results for the SVM because the parameters lose their interpretations due to the penalization term when fitting the parameters. The results for existing and new products are similar, although the model fit is better for existing products and more of the coefficient estimates are statistically significant. This is not entirely unexpected, due to the considerably greater number of parameters in the model for existing products. Focusing on the full model results for



(a) Model selection in estimation sample.

(b) Model selection in estimation sample.

(c) Model selection in estimation sample.

Fig. 1. Measures of error between actual and forecasted sales ranks for existing (ES_EP) and new (ES_NP) products.

Table 5
OLS parameter estimates of the baseline and full models for new and existing products.

Parameters	Baseline model		Full model ($BOWS_{jt}$ coefficients not shown) **Existing** products, $k = 300$		Full model($BOWS_{jt}$ coefficients not shown) **New** products, $k = 50$	
	Estimate	Std. error	Estimate	Std. error	Estimate	Std. error
Intercept α	-8.609^{***}	0.532	-6.945^{***}	0.410	-7.865^{***}	0.440
δ^{OS}_{Apple}	0.923^{***}	0.057	1.056^{***}	0.052	0.938^{***}	0.051
δ^{OS}_{Black}	1.467^{***}	0.102	2.173^{***}	0.183	0.705^{***}	0.105
δ^{OS}_{HP}	1.825^{***}	0.117	0.707^{***}	0.152	1.461^{***}	0.102
δ^{OS}_{Win}	-0.123^{*}	0.061	0.087^{****}	0.052	0.080	0.054
δ^{BL}_{4+}	-0.067^{****}	0.039	-0.072^{*}	0.028	-0.006	0.033
δ^{BL}_{7+}	-0.167^{***}	0.04	-0.173^{***}	0.03	-0.028	0.034
δ^{BL}_{10+}	0.299^{***}	0.051	0.005	0.041	0.133^{**}	0.044
δ^{SS}_{Med}	0.261^{***}	0.033	0.127^{***}	0.025	0.133^{***}	0.028
δ^{SS}_{High}	0.428^{***}	0.083	0.348^{***}	0.063	0.284^{***}	0.070
δ^{RAM}_{High}	0.114^{***}	0.035	0.033	0.028	0.002	0.030
$\delta^{\Pr oc}$	0.201^{***}	0.055	0.011	0.042	0.095	0.047
$\delta^{Storage}$	0.081^{***}	0.017	0.123^{***}	0.013	0.078^{***}	0.014
δ^{Weight}	-0.221^{***}	0.035	-0.056^{*}	0.026	-0.078^{**}	0.029
δ^{PPI}	0.913^{***}	0.113	0.440^{***}	0.083	0.612^{***}	0.093
β_{trend}	-0.015^{***}	0.002	0.001^{****}	0.000	-0.009^{***}	0.001
β_{price}	-0.131^{***}	0.031	-0.258^{***}	0.022	0.153^{***}	0.026
β_{RS}	0.076^{***}	0.022	0.177^{***}	0.031	0.073^{***}	0.022
Observations	5544		5544		5245	
R^2	0.342		0.862		0.619	

Lambda is fixed at 0.90.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.
**** $p < 0.10$.

existing products, the signs of most of the estimated attribute effects are as expected, but with some exceptions. The attribute values that are preferred include larger screen sizes, higher RAM, faster processors (not statistically significant), more storage, lower weight, and more pixels per inch (PPI). The signs of the battery life coefficients are not consistent with the expectation that a greater battery life will be preferred. While the large number of face-valid coefficient signs is encouraging, we are not too surprised at having a few unintuitive coefficients.

The model is likely to be misspecified in terms of product characteristics because of omitted features such as the brand, which may be correlated with the features included. Furthermore, the features included are collinear, as would be expected with real world data; higher quality tablet computers have better levels of most features, and also tend to be higher priced. Finally, we deal with missing data on product features by data imputation, which will also introduce error. The incidence of such imputations was less than 1% of observations for most product features except for the battery life, for which we had to impute values for about 13% of observations. The estimated effect of log(*price*) is negative, as expected. Interestingly, the magnitude of the coefficient in the full model is almost double that in the baseline model. As expected, the effect of numeric ratings in consumer reviews is positive.

4.3. Predictive performances in validation samples

Table 6 shows the validation results of VS_EP and VS_NP using the optimal values of k relative to the baseline model without the bag-of-words representation of the textual content of reviews. All of the results were averaged across 10 simulations except for the SVM, which had a single result.[8] The full model had better out-of-sample predictive performances than the baseline model across all measures, for both existing and new products. In addition, the full model had better out-of-sample predictive performances than the SVM across all measures for existing products. The SVM's performance for existing products was in between those of the full model and the baseline model, which indicates that the textual information added value, but more so when combined with random projections. For new products, the full model had better RMSEs and MAPEs than the SVM, and a correlation similar to the SVM. Therefore, we suggest the use of our random projections model with a bag-of-words representation for both scenarios.

As a side note, the in-sample forecast correlations for existing products were always over 0.90 for the full model with $k = 300$, and 0.58 for the baseline model. This exemplifies the difference between out-of-sample forecasting and in-sample fitting. Although simply increasing the value of k always improved the in-sample performance (a greater number of predictors always increases the R^2 value), our results showed that this is not desirable in the forecasting task. There is a trade-off between the number of predictors and the out-of-sample performance, due to over-fitting. In fact, if we had used $k = 500$ to forecast new products in VS_NP, our results show that the performance would have been worse than that of the baseline model, with a RMSE of 142.2, a MAPE of 336.3% and a correlation of 0.210.

To illustrate the difference in performance between the full model and the baseline model for the validation sample further, Figs. 2a and 2b show scatter plots of predicted versus actual ranks for existing products using VS_EP. This plot contains 12 weeks × 231 products = 2772 observations. In Fig. 2a, the better performance of the full model relative to the baseline model, especially at low actual ranks, can be discerned from the significantly greater closeness of the forecasts from the full model to the 45° line, relative to forecasts from the baseline model. Similarly, Fig. 2b shows that the SVM systematically overestimates sales ranks at low ranks, and underestimates sales ranks at high ranks.

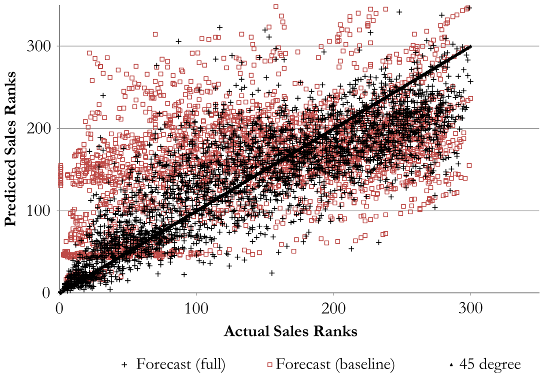


Fig. 2a. Forecasts of sales ranks of existing products in validation sample: Full model versus baseline model.

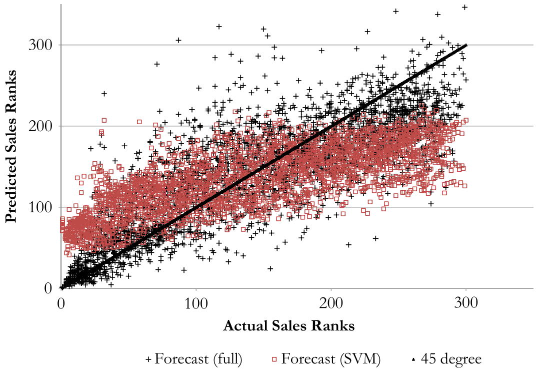


Fig. 2b. Forecasts of sales ranks of existing products in validation sample: Full model versus SVM.

4.4. "Hot" versus "not"

For new products, one issue that is of interest to managers is the predictive identification of the top performers, in order to ensure the appropriate allocation of scarce resources such as management attention. Colloquially, the question may be phrased as whether a new product will be "hot" or "not". Say that we assume that management is interested in predictively identifying the 54 highest ranked (based on *actual* sales ranks) observations, or 10% of the 545 observations in VS_NP. In Fig. 3, we display a gains chart that shows the percentage of the 54 observations that are "hit" as successively higher percentages of the 545 observations are "examined", based on the rank orderings predicted by the full model and the baseline model

[8] Similarly to the proposed model, predictions from the SVM were made in log-space and then converted back to the original space by exponentiation. However, the correction factor referred to earlier was not applied, since the errors in the SVM are not known to be log-normally distributed.

Table 6
Predictive performances in validation samples: full model versus baseline models.

Existing products (VS_EP)	*K*	RMSE	MAPE	Correlation
Full model	300	48.4	37.2%	0.813
Baseline model	–	80.0	163.5%	0.514
SVM		60.6	122.7%	0.741
New products (VS_NP)				
Full model	50	72.6	243.9%	0.464
Baseline model	–	80.4	288.6%	0.253
SVM		77.9	360.3%	0.471

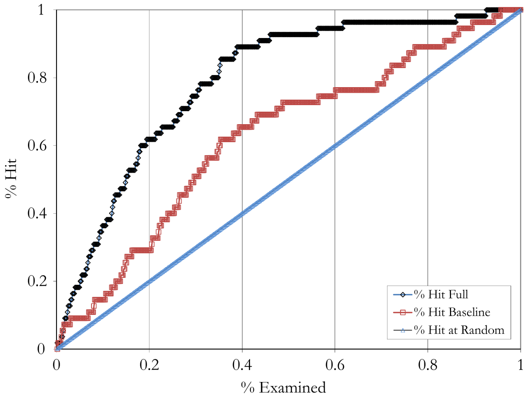


Fig. 3. Identifying top 10% performers in new product validation sample.

respectively. The dominance of the full model over the baseline model is quite clear. If management examines the top 20% of the 545 observations (about 110) based on predictions from the full model, that set would include 60% or 32 of the top performers, whereas the baseline model would only lead to 31% or 15 of the top performers being included. Furthermore, when only 60% of the observations are examined, virtually all 54 of the top performers are hit by the full model.

In order to measure and test the statistical significance of whether a new product will be "hot or not", we calculated the probability that a randomly chosen "hot" product will have a higher forecast (i.e., a better forecasted rank) than a randomly chosen "not" product. Naive random guessing would result in a 50% probability. This probability is equivalent to the non-parametric Wilcoxon–Mann–Whitney test statistic, or the area under the curves in Fig. 3 after the *x*-axis has been adjusted to exclude the "hot" products (i.e., including only the % examined of the bottom 95%, 90%, or 80%, respectively). Since the forecasts from our baseline and full models were correlated, we used a paired test of comparisons to test the null hypothesis that the probabilities were equal (Delong, DeLong, & Clarke-Pearson, 1988). Table 7 shows that, for all Top A% that management considers, the full model has statistically greater probabilities than the baseline model of identifying a "hot" new product at any significance level. The full model forecasts a "hot" new product as having a better rank than a "not" new product more than 80% of the time. Table 7 also shows that, when forecasting new products, the full model and SVM are statistically the same for Top 5% and Top 10%, but the full model outperforms the SVM at the Top 20% level.

5. Conclusions and future research

The rapid emergence of e-commerce and massive search engines has meant that business is confronted with an explosion of data that potentially holds a huge promise of informing better business decisions. The textual content of online consumer reviews is one such form of data that is almost entirely qualitative and very voluminous, but could be extremely helpful for sales forecasting. The challenge for business is to benefit from the data without becoming overwhelmed by its complexity, volume and rapidity of change.

This paper examines an approach to handling very high-dimensional data that offers considerable benefits in regard to scalability, speed, and flexibility, with relatively low resource requirements. In our view, the last benefit is especially important, because, while the access to data at sites like Amazon is virtually free for anyone who can scrape the web, the computational resources and statistical expertise then required to mine the data are prohibitive for many businesses. As a result, textual data are often omitted from predictive models. The approach proposed in this paper makes it easy and inexpensive to include textual data in a largely automated form. Furthermore, our results demonstrate that textual data add substantially to models' predictive abilities. We also found that the model incorporating the random projections model performed substantially better than a competitor model for the forecasting of existing products, and slightly better than the competitor model for the forecasting of new products.

There are several directions in which we believe that the research presented here can be extended fruitfully. Since our forecasting task involved an (assumed-to-be) interval-scaled variable, it was appropriate to use regression analysis. However, in many predictive problems the dependent variable of interest is categorical, and the task involves classification. We believe that classification techniques like *k*-nearest neighbors (*k*-NN) can be beneficial because *k*-NN is already used commonly as a computationally efficient big data technique, and predicts categorical classifications based upon exactly the same pairwise distance metrics that random projections theoretically preserve. Thus, businesses can choose an appropriate distance metric (e.g., Euclidean, Manhattan, or Mahalanobis) depending on their managerial objective, whereas our regression analysis was restricted to minimizing the mean squared error of the residuals. In addition, we feel that a useful approach for addressing our concern about overfitting in the forecasting task would be to use random projections with a ridge regression. Ridge regressions also

Table 7
Probabilities of better ranking of "hot" versus "not" new products.

Top A% new products	Baseline model Pr(Hot > Not)	Full model Pr(Hot > Not)	SVM Pr(Hot > Not)	p-value (baseline vs. full)	p-value (SVM vs. full)
Top 5%	0.714	0.829	0.849	0.004	0.161
Top 10%	0.642	0.820	0.787	0.000	0.072
Top 20%	0.657	0.830	0.784	0.000	0.016

minimize the mean squared error, like regressions, but shrink the estimated coefficients toward zero, which reduces the variance relative to a regression. Similarly, we implemented a SVM on the bag-of-words data directly, but found that the forecasting performance suffered, perhaps due to an overfitting or overweighting of some textual content. Future research may consider the combination of random projections and a model that uses a shrinkage parameter.

However, none of these more advanced techniques comes without a limitation relative to regression, which seems to be a consistent theme for "Big Data" techniques. Regression analysis preserves the interpretability of the estimated relationships between the predictors and the dependent variable (e.g., an increase in storage size is associated with an increase in product sales), which is an important requirement for marketing planning and decision making. Ridge regression, on the other hand, inserts a bias into the regression coefficients, which makes them more difficult to interpret. It also introduces an additional parameter to be optimized, which makes it more computationally expensive in the forecasting scheme. Finally, although k-NN classifies an observation based upon its "nearest neighbors" with minimal pairwise distances (which certain random projections preserve), it disregards inference and only outputs predictions.

Another important question that we do not explore in depth here is the choice of the distribution for the random matrix **R**. We expect additional gains from varying the parameters and type of the random distribution, depending on the application. In many applications, the predictor matrix **X** is extremely large, with dimensions in the tens or hundreds of millions. Enormous computational savings can be achieved by using sparse random projections, in the sense that only a few entries of each row of **R** are populated by non-zeroes. By simply storing the locations of these non-zero entries instead of the entire matrix, the processing cost for multiplying **X** by **R** decreases in proportion to the sparseness of **R**. The savings in processing costs appear because any elements in the dot product with a zero entry do not need to be multiplied and can be ignored; therefore, focusing only on the non-zero elements of **X** and **R** increases the processing speed and scalability of the method greatly when the sparseness is high. This is especially important when considering non-sequential word pairs or word triplets, which may have zero entries in over 99% of the rows in **X**. However, one limitation is that the sample sizes (n) need to be very large under this setup, to ensure that certain word pairs or triplets are represented in both the estimation and validation data sets. Otherwise, we expect the problem of over-fitting to unique pairs and triplets to degrade the forecasting performance, which is one reason why we limited our analysis to single words, given the limited sample size (products and time periods) in the current empirical analysis. We also did not investigate dimension reduction using random projections on the sample size (n) of the predictor matrix, which can be in the tens of millions for some marketing applications. Theoretical work with random projections for OLS estimators exists on this sampling issue, but little empirical work was found. Finally, the use of marginal norms in a maximum likelihood estimator has been shown to improve the performance of random projections (Li, Hastie, & Church, 2006b). The benefits for the forecasting performance can be examined in future work.

In summary, in this paper we have proposed an approach that uses the textual content of consumer-generated reviews for the sales forecasting of new and existing products. The data in this situation have many of the typical characteristics of "big data"—they are voluminous and dynamic, and contain a mix of structured and unstructured formats. The proposed method has the significant merits of being largely context-independent and highly automatable, both of which make it attractive for dealing with such data. Our empirical application demonstrated the great value of the proposed approach relative to both a baseline model in which the textual content of reviews is ignored and a competitor model in which the textual content was included in the same structure as in the proposed approach.

References

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

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, *8*(1), 69–80.

Arriaga, R. I., & Vempala, S. (2006). An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, *63*(2), 161–182.

Bingham, E., & Mannila, H. (2001). Random projections in dimensionality reduction: applications to image and text data. In *Proceedings of KDD, San Francisco, CA* (pp. 245–250).

Brynjolfsson, E., Hu, Y., & Smith, M. (2003). Consumer surplus in the digital economy: estimating the value of increased product variety. *Management Science*, *49*(11), 1580–1596.

Chen, Y., & Xie, J. (2008). Online consumer review: word-of-mouth as a new element of marketing communication mix. *Management Science*, *54*(3), 477–491.

Chevalier, J. A., & Goolsbee, A. (2003). Measuring prices and price competition online: Amazon.com and Barnes and Noble.com. *Quantitative Marketing and Economics*, *1*, 203–222.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: online book reviews. *Journal of Marketing Research*, *48*, 345–354.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, *22*(3), 443–473.

Delong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*, 837–845.

Fader, P. S., & Hardie, B. G. S. (1996). Modeling consumer choice among SKUs. *Journal of Marketing Research*, *33*(4), 442–452.

Fard, M. M., Grinberg, Y., Pineau, J., & Precup, D. (2012). Compressed least squares regression on sparse spaces. In *Proceedings of the 26th AAAI conference on artificial intelligence* (pp. 1054–1060).

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, *23*(10), 1498–1512.

Goldberger, A. S. (1968). The interpretation and estimation of Cobb–Douglas functions. *Econometrica*, *35*, 464–472.

Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research*, *8*, 355–363.

Ho-Dac, N. N., Carson, S. J., & Moore, W. L. (2013). The effect of positive and negative customer reviews: do brand strength and category maturity matter? *Journal of Marketing*, *77*, 37–53.

Hu, N., Koh, N. S., & Reddy, S. K. (2013). Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. *Decision Support Systems*, *57*(1), 42–53.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipshitz mapping into Hilbert space. In *Contemporary mathematics: conference in modern analysis and probability. Vol. 26* (pp. 189–206). American Mathematical Society.

Kolassa, S., & Schütz, W. (2007). Advantages of the MAD/MEAN ratio over the MAPE. *Foresight: The International Journal of Applied Forecasting*, *6*, 40–43.

Lee, T., & Bradlow, E. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research*, *48*(5), 881–894.

Li, P., Hastie, T.J., & Church, K.W. (2006a). Very sparse random projections. In *KDD* (pp. 287–296).

Li, P., Hastie, T. J., & Church, K. W. (2006b). Improving random projections using marginal information. In *Learning theory: lecture notes in computer science. Vol. 4005* (pp. 635–649). Springer.

Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, *19*(4), 456–474.

Li, P., Samorodnitsky, G., & Hopcroft, J. (2013). Sign stable projections, sign Cauchy projections and chi-square kernels. Working paper, Cornell University.

Liu, Y. (2006). Word-of-mouth for movies: its dynamics and impact on box-office revenue. *Journal of Marketing*, *70*(3), 74–89.

Maillard, O.-A., & Munos, R. (2012). Linear regression with random projections. *Journal of Machine Learning Research*, *13*, 2735–2772.

Moe, W., & Trusov, M. (2011). The value of social dynamics in online product ratings forums. *Journal of Marketing Research*, *48*, 444–456.

National Research Council (2013). *Frontiers in massive data analysis*. Washington DC: The National Academies Press.

Nerlove, M., & Arrow, K. J. (1962). Optimal advertising policy under dynamic conditions. *Economica*, *29*, 129–142.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: market-structure surveillance through text mining. *Marketing Science*, *31*(3), 521–543.

Sun, M. (2012). How does the variance of product ratings matter? *Management Science*, *58*(4), 696–707.

van der Maaten, L.J.P., Postma, E.O., & van den Herik, H.J. (2009). Dimensionality reduction: a comparative review. Tilburg University Technical Report, TiCC-TR 2009-005.

Wang, X., Mai, F., & Chiang, R. H. L. (2013). Market dynamics and user-generated content about tablet computers. *Marketing Science*, *33*(3), 449–458.

Zhao, Y., Yang, S., Narayan, V., & Zhao, Y. (2013). Modeling consumer learning from online product reviews. *Marketing Science*, *32*(1), 153–169.

Matthew J. Schneider is Assistant Professor in the Integrated Marketing Communications Department at the Medill School of Journalism, Media, Integrated Marketing Communications at Northwestern University. He holds a Ph.D. and M.S. in Statistics, a M.S. in Public Policy and Management, and a B.S. in Quantitative Economics.

Sachin Gupta is the Henrietta Johnson Louis Professor of Management and Marketing at the Samuel Curtis Johnson Graduate School of Management at Cornell University. He received a BA in Economics from St. Stephen's College, University of Delhi, and an MBA from the Indian Institute of Management at Ahmedabad, both in India. He received a Ph.D. in management at Cornell University in 1993 and began his teaching career at the Kellogg Graduate School of Management at Northwestern University. His research interests include manufacturer and retailer pricing, advertising and sales promotions, competitive strategies, models of consumer choice behavior, and sustainable innovations.