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 impracticably 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 pre-processing 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