The Operational Value of Social Media Information

Article *in* Production and Operations Management · April 2017

DOI: 10.1111/poms.12707

Ruomeng Cui
Emory University
36 PUBLICATIONS **1,251** CITATIONS

Santiago Gallino
University of Pennsylvania
38 PUBLICATIONS **2,335** CITATIONS

Antonio Moreno
Harvard University
31 PUBLICATIONS **2,083** CITATIONS

Dennis J. Zhang
Washington University in St. Louis
43 PUBLICATIONS **1,458** CITATIONS

POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT POMS

Vol. 0, No. 0, xxxx–xxxx 2017, pp. 1–21
ISSN 1059-1478 | EISSN 1937-5956 | 17 | 00 | 0001

DOI 10.1111/poms.12707

The Operational Value of Social Media Information

Ruomeng Cui
Kelley School of Business, Indiana University, Bloomington, Indiana 47405, USA, cuir@indiana.edu

Santiago Gallino
Tuck School of Business, Dartmouth College, Santiago, New Hampshire 03755, USA, gallino@tuck.dartmouth.edu

Antonio Moreno
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208, USA, a-morenogarcia@kellogg.northwestern.edu

Dennis J. Zhang
John M. Olin Business School, Washington University in Saint Louis, Saint Louis, Missouri 63130, USA, denniszhang@wustl.edu

While the value of using social media information has been established in multiple business contexts, the field of operations and supply chain management have not yet explored the possibilities it offers in improving firms' operational decisions. This study attempts to do that by empirically studying whether using publicly available social media information can improve the accuracy of daily sales forecasts.We collaborated with an online apparel retailer to assemble a dataset that combines (1) detailed internal operational information, including data on sales, advertising, and promotions, as well as (2) publicly available social media information obtained from Facebook. We implement a variety of machine learning methods to forecast daily sales. We find that using social media information results in statistically significant improvements in the out-of-sample accuracy of the forecasts, with relative improvements ranging from 12.85% to 23.23% over different forecast horizons. We also demonstrate that nonlinear boosting models with feature selection, such as random forests, perform significantly better than traditional linear models. The best-performing method (random forest) yields an out-of-sample MAPE of 7.21% when not using social media information and 5.73% when using social media information is used. In both cases, this significantly improves the accuracy of the company's internal forecasts (a MAPE of 11.97%). Combining these empirical results, we provide recommendations for forecasting sales in general as well as with social media information.

Key words: social media; sales forecast; machine learning
History: Received: February 2016; Accepted: February 2017 by Ram Ganeshan, after 2 revisions.

1. Introduction

Social media is creating new channels for customers to interact and communicate with companies: more than 83% of the Inc. 500 companies use at least one of the main social media sites (Hameed 2011). Customers routinely use social media to engage in discussions about companies' products and services. As a consequence, customers' preferences, opinions, and emotions are embedded in their activities on social media.

More customers are using social networks such as Facebook and Twitter to express their preferences online, and their purchasing decisions increasingly are influenced by their friends' referrals. Recognizing this, many companies have started to use social media information to improve their marketing and advertising decisions. At the same time, a number of academic studies in marketing, finance, and information systems have demonstrated the value of social media information when making managerial decisions (e.g., see Aral 2011, Bollen et al. 2011, Goh et al. 2013).

Despite the attention that practitioners and academics from different fields have devoted recently to social media, the field of operations management has not yet studied the opportunities that social media information offers to improve operational decisions. In this study, we explore whether it is possible and recommend to use social media information to improve sales forecasts.

There are two challenges that may prevent researchers from exploring the operational value of social media information. The first is that researchers need to combine two types of data: internal *operational data* available to the firm and external *social media information* available on social media sites. To close this gap, we collaborate with an online apparel retailer to assemble a dataset that allows us to study the value of social media information in sales forecasting. We begin by obtaining advertising and promotional

Please Cite this article in press as: Cui, R., et al. The Operational Value of Social Media Information. *Production and Operations Management* (2017), doi 10.1111/poms.12707

data from the company, internal operational data—e.g., daily sales for new customers who have never transacted with the company before and repeat customers who previously purchased from the company—and the company's internal sales forecast. We then collect publicly available social media information from Facebook, such as company posts, user comments, and "shares" and "likes" data.[1] In addition, we analyze the informativeness and sentiment of user comments using natural language processing (NLP) techniques.

The second challenge is that forecasting sales with social media information is a high-dimensional problem (i.e., the number of variables is large relative to the number of observations) and traditional forecasting methods, such as linear regression and time-series models, often perform poorly with high-dimensional data. To overcome this challenge, we develop a machine-learning-based framework. Machine learning models are well equipped to handle this type of high-dimensional problem and prevent over-fitting. Consequently, in our analysis, we implement a wide variety of machine learning models with different characteristics and degrees of sophistication.[2] The models we use include simple linear regression, linear regression with forward selection, lasso, support vector machines with linear and radial kernels, gradient boosting method, and random forests.

We apply each of the machine learning models to construct two different aggregate daily sales forecasts: (1) a *baseline* forecast that uses only the internal operations data and (2) a *social-media-enhanced* forecast that incorporates publicly available social media information in addition to the internal operations data. In each case, we use cross-validation to select the model's hyper-parameters. The comparison between the out-of-sample forecast accuracy using these two types of forecasts allows us to quantify the value of social media information in improving sales forecasts. In addition, the apparel company maintains its own internal sales forecasts that use the operational data we have access to as well as other information that is unknown to us. Comparing the company's internal forecasts with our baseline forecasts is informative of the potential value of using advanced statistical learning models to build forecasts.

This study has two main contributions. First, we demonstrate that considering social media information improves daily sales forecast accuracy, measured by the out-of-sample mean absolute percentage error (MAPE). The accuracy improvement of our social-media-enhanced forecasts is large, ranging from 12.85% to 23.23% across different forecast horizons, and even larger relative to the company's internal forecasts, ranging from 39.35% to 52.13%. The positive and significant value of social media is robust with respect to a number of forecasting models, prediction horizons, testing periods, and sales coming from new or repeat customers. While our analysis focuses on forecasting aggregate daily sales, we discuss how our models can be used in product-category or SKU-based forecasting.

Second, since we implement a wide variety of machine learning models, we provide empirical evidence on what type of models can effectively benefit from social media information and perform the best in sales forecasting. Interestingly, the value of social media information increases with the machine learning technique's level of sophistication. We show that boosting and nonlinear models with feature selection benefit more from social media information and perform the best. In particular, our best performing model—random forest—achieves an out-of-sample MAPE of 5.73%, compared to a MAPE of 14.43% from the linear lasso model and a MAPE of 11.97% from the company's internal forecast. Naive models, such as linear regression without variable selection, can have deteriorated performance when incorporating social media information. We interpret this in light of the high-dimensional nature of the sales forecasting problem with social media information using the well-established bias variance trade-off framework. In order to fully unlock the potential of social media information in improving forecast accuracy, our results suggest companies should invest in more advanced forecasting techniques.

2. Literature Review and Theoretical Mechanism

2.1. Literature Review

Social media has become part of our lives over the last decade. As social media data have become accessible to researchers, various studies have attempted to demonstrate the value of this information in finance, marketing, and information systems. For example, Bollen et al. (2011) propose a predictive model for the stock market using microblog data and demonstrate that social media information improves forecast returns of major financial indexes; Chen and Xie (2008) study how sellers should adjust their marketing strategies dynamically to maximize their revenues based on online reviews and comments, and Luca (2011) quantifies the impact of local restaurant reviews posted on social media sites and points out the necessity to manage social media information.

Despite this growing social-media-related literature from different academic areas, to our knowledge limited research in the operations management community has attempted to empirically assess the value of social media information. Practitioners often have expressed interest in unlocking the potential of social

Please Cite this article in press as: Cui, R., et al. The Operational Value of Social Media Information. *Production and Operations Management* (2017), doi 10.1111/poms.12707

media information to improve their daily operations (see, e.g., Supply Chain Nation Blog),[3] but there is no rigorous evidence regarding the impact of incorporating social media information within the forecasting process. Our work addresses these shortcomings by quantifying the operational value of social media information.

There are three streams of literature in operations management that are relevant to our work: the literature on sales forecasting, the theoretical literature that studies operations in the presence of social networks, and the emerging literature that uses machine learning techniques to study operational issues.

Sales forecasting has a long tradition of research in operations management. The operations management literature has been concerned with the issue of sales forecasting and improving sales forecasts. Accurate sales forecasts have been shown to convey information on future earnings and returns to investors (Nichols and Tsay 1979, Penman 1980) and also can facilitate online and in-store inventory management for companies. For example, Gaur et al. (2007) use dispersion among experts' forecasts as a measure of demand uncertainty for new products; Bassamboo et al. (2015) study how crowds can obtain better forecasts than can individuals; Kesavan et al. (2010) show how using historical inventory and gross margin data can improve forecasting; Osadchiy et al. (2013) show that a model combining analysts' forecasts and financial market returns improves forecast accuracy; and Kremer et al. (2011) study the behavioral bias of forecasting and propose an intervention to improve the forecast accuracy. Downstream information, such as point-of-sales, has also been shown to add value to upstream order forecasting (see Gaur et al. 2005 and Cui et al. 2015). These papers use financial market index data, accounting variables, or external operations information to improve sales forecasts. As some of the aforementioned studies discuss, even a small improvement in sales accuracy can lead to large impacts on stock prices and inventory cost reductions. Our work relates closely to this stream of research but focuses on the role of a novel indicator: social media information. Our results complement this stream of literature by pointing to the value of social media information, which is available publicly and can be obtained easily, to help predict sales.

A recent literature stream has studied various issues in the operations management space when considering the presence of social network/learning effects. Most of the work in this space is based on theoretical models. Candogan et al. (2012) study a revenue management problem when customers' consumption depends on their friends' consumption in a social network. Zhang et al. (2015) and Allon and Zhang (2015) study how to manage services in the presence of social interactions. Jing (2011) propose dynamic pricing strategies for new products when customers strategically delay their purchases to learn more about products. Ifrach et al. (2011) provide insights into dynamic pricing strategies when the strategies affect not only the revenue but also the information flow to subsequent customers and their friends. Ye et al. (2014) study retailers' dynamic bidding strategy in use of sponsored search. While the aforementioned literature mainly focuses on the theoretical value of social network information in pricing decisions, we study the empirical value of publicly available social media information in forecasting.

In terms of methodology, our work uses machine learning methods. Recent papers in operations have used machine learning tools to analyze such problems as demand estimation and price optimization in online retail Ye et al. (2014), predicting emergency department waiting times (Ang et al. 2015), and developing a feature-based algorithm to solve nursing staff problems (Rudin and Vahn 2015). We contribute to this emerging stream of work by implementing machine learning methods to construct sales forecasts that incorporate social media information.

2.2. Theoretical Mechanism

Social media information affects customers' purchasing behaviour through two mechanisms: the "attention" effect, which captures a person's awareness of products, and the "endorsement" effect, which informs users of product quality based on their friends' online comments. The social media literature has identified the attention and endorsement effects as providing different types of informational signals that capture user behaviors (see Banerjee et al. 2012, Chen et al. 2011, Li and Wu 2014). In particular, Banerjee et al. (2012) differentiate the effect of information passing (i.e., people should be aware of the product before purchasing it) and endorsement (i.e., individuals' purchasing decisions are affected by their friends' decisions) as sales drivers. We next elaborate on the definition of these two social media effects.

Attention. Attention plays a critical role in affecting buyers' behavior and this potentially can be leveraged when forecasting the future. For example, the finance literature (Barber and Odean 2008, Da et al. 2011) finds both theoretical support and empirical evidence that investor attention (e.g., indicated by a high Google search volume) leads to a stock price increase. The economic literature (Bikhchandani et al. 1992, Cai et al. 2007, Moretti 2011) explains attention-driven sales when people engage in social learning. Vahl

(2013) points out that attention is particularly important in the online industry, where "a wealth of information creates a poverty of attention."

Endorsement. Users can disseminate their sentiments via social media channels. Sharing and giving positive comments signals users' positive emotions to their friends and followers. The literature has documented the endorsement effect. For example, Li and Wu (2014) show customers' sentiments on a deal's web page provide an information signal that helps update customers' beliefs about product quality and thus has a significant impact on sales.

In the next section, we use the theory described above as a guidance to select social media features as predictive variables.

3. Empirical Setting and Data

We partner with an online retailer that sells men's clothing. We use data from this company to document the value of social media in forecasting aggregate daily sales. In section 8, Discussion and Conclusions, we discuss how our approach can be used in other contexts, such as product-category or SKU-based forecasting.

The company sells exclusively through its online website and was listed among the top 500 largest e-retailers in America in 2014. To study the value of social media information, we collect two datasets: internal operational data from the apparel company and publicly available social media data from the company's Facebook page.

3.1. Operational Data

The internal operational data we obtain from the apparel company includes sales, the advertising and promotions schedule, and the company's internal sales forecasts. Sales data includes daily total sales, daily sales by new customers (who have not purchase previously from the company), and daily sales from repeat customers (who have a buying history with the company). We obtain this data for the period between January 2013 and July 2013.

The company frequently advertises its new product launches and announces promotional activities to customers. These actions are likely to affect sales and potentially correlate to social media activity since advertising can attract comments and "likes" on the company's Facebook page. For this reason, it is crucial to collect this internal information and use it in our baseline forecasts. If we do not do this, the improvements in forecast accuracy that we obtain when incorporating social media information could in fact arise from its ability to capture information present in other internal variables that usually are available to firms through other, more immediate channels.

In order to capture the value of social media information on top of the information related to marketing efforts, which is internally available to the firm, we collect detailed information about the company's marketing activities. In particular, we obtain the complete history of the company's email campaigns because email is the company's primary marketing and communications tool (since its main presence is online). The number of advertising and promotional activities using other channels is small and highly correlated with the email campaigns. Consequently, including information extracted from the email campaigns in our baseline forecasts allows us to consider the effect of marketing interventions independently from social media information. We construct three important daily variables related to the company's marketing effort, specifically, whether the company (1) sends out advertisement emails, (2) sends out an email announcing a promotion,[4] and (3) is running a promotion on that day.[5]

Note that there can be other variables available to the firm but not to us that have important predictive power when constructing forecasts. It is not feasible for us to access all of the company's internal information that could be incorporated into our baseline forecasts. However, we are able to obtain the company's internal sales forecasts, which provide a very useful benchmark. The company internally predicts daily sales to facilitate its operational decisions. We do not have details about the specific algorithm used, but based on our conversations with company officials, we know the firm uses a variety of available structured and unstructured information; we learn this does not include social media information.

Table 1 presents summary statistics for the operations data (a full set of the operational features can be found in Table A1 in Appendix A). On average, there are 646.9 total transactions per day. Sales from repeat customers account for around 65% of the total sales. The company sends out marketing emails 69% of the days in our sample, and 20% of these emails are related to promotions. The company runs promotions 27% of the days in our sample. Note that our period of analysis covers several major holidays, such as New Year's Day and Memorial Day, which usually have high volatility in sales. Therefore, we also include holiday dummies to capture this information.

How is new years day covered if the data starts from Jan?

3.2. Social Media Information

In addition to the company's operational data, we collect detailed social media information for the company, specifically focusing on Facebook. Facebook is the main social network used by the retailer, accounting for more than 90% of visits to the

Please Cite this article in press as: Cui, R., et al. The Operational Value of Social Media Information. *Production and Operations Management* (2017), doi 10.1111/poms.12707

company's website that come from social media sites. An additional advantage of using Facebook data is that unlike other social media sites, such as Twitter or Pinterest, Facebook offers a public application programming interface (API) to access the complete dataset of social media activities on any company's Facebook page. Thus, we document the operational value of using publicly available social media information.

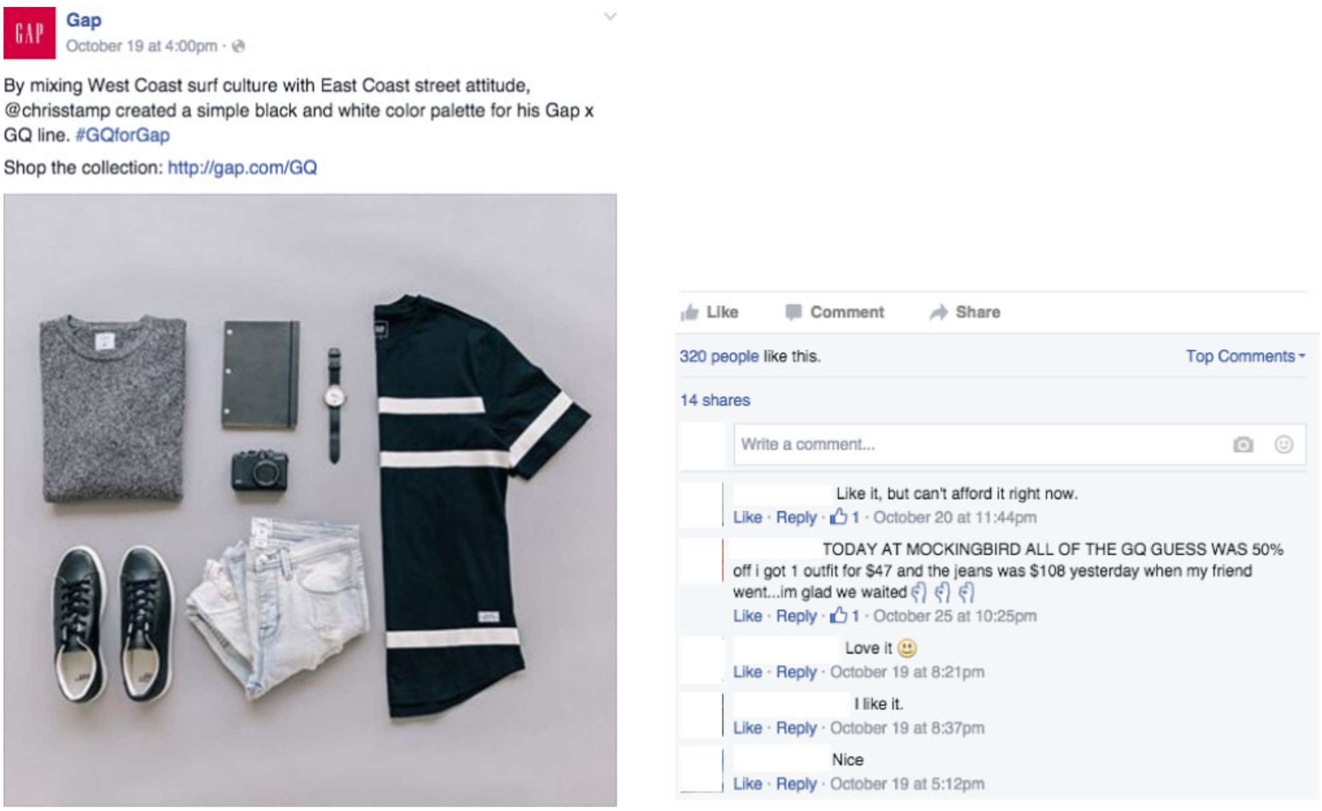
Our partner company has an official Facebook page that more than 300,000 users follow. Figure 1 shows an example of what a standard post by a similar company looks like.[6] In general, users can interact with a company's Facebook page in three ways. First, they can write remarks using the "comment" option under each post. Second, a "like" button allows users to endorse or approve the post. Third, a "share" button allows users to publish the company's post on their own Facebook page. If a user shares, likes, or comments on a post, her friends may see the same post in their news feed. Moreover, a user's comments and likes can be seen by her friends if they visit her main Facebook page. As shown in Figure 1, there are 320 Facebook users who liked this post, 5 users who commented, and 14 people who shared it.

Data Collection. We record all social activities that users made on the company's Facebook page during our period of analysis. To do so, we develop a Python program to extract all posts on the company's Facebook page through the API. Since each post is identified by a universal post ID in the Facebook database,

Table 1 Summary Statistics of Operations Data

Category	Variable	Mean	Median	SD	Maximum	Minimum
Sales data	Total sales	646.93	617.00	254.62	2003.00	225.00
	New sales	225.93	224.00	60.08	454.00	94.00
	Repeated sales	421.00	386.00	209.92	1596.00	108.00
Marketing data	Ad	0.69	1.0	0.46	1.0	0.0
	Promotion	0.14	0.0	0.35	1.0	0.0
	In promotion	0.27	0.0	0.45	1.0	0.0
Forecast data	Internal sales forecasts	634.00	644.50	129.25	873.00	362.00

Figure 1 A Facebook Post Example [Color figure can be viewed at wileyonlinelibrary.com]



Gap
October 19 at 4:00pm ·

By mixing West Coast surf culture with East Coast street attitude, @chrisstamp created a simple black and white color palette for his Gap x GQ line. #GQforGap

Shop the collection: http://gap.com/GQ

320 people like this.

14 shares

[redacted] Like it, but can't afford it right now.
Like · Reply · 1 · October 20 at 11:44pm

[redacted] TODAY AT MOCKINGBIRD ALL OF THE GQ GUESS WAS 50% off i got 1 outfit for $47 and the jeans was $108 yesterday when my friend went...im glad we waited
Like · Reply · 1 · October 25 at 10:25pm

[redacted] Love it 🙂
Like · Reply · October 19 at 8:21pm

[redacted] I like it.
Like · Reply · October 19 at 8:37pm

[redacted] Nice
Like · Reply · October 19 at 5:12pm

we follow each post to find the identifiers of each comment, like, and share made by users. Finally, following the unique identifiers we extract information related to each comment, like, and share, such as the unique users, each user's number of comments, the comment's content, and the timestamps of these events when available. In total, we have 171,279 unique users who interacted with at least 1 of the 1,943 company's posts. Among these interactions, there are 25,730 comments and 266,534 likes. Note that Facebook does not allow developers to get the timestamps of likes and shares. We focus post and comment information, for which we know the timestamps, to map the social media information with the company's operations information on a specific day.

Feature Extraction. We use two attributes to measure the "attention" and "endorsement" effects presented in section 2.2: volume and valence respectively. Following the literature, the volume features consist of the daily number of posts (representing the company's social media activity), the daily number of comments (representing the users' activity), the daily number of unique users who commented, the number of comments by the company, etc. The valence feature for a comment includes its informativeness and its sentiment. Following Resnik et al. (1999), we measure informativeness by the number of words, sentences, and unique words in a comment. Informativeness is a widely adopted measure that has been proven to be useful in many other prediction settings, such as forecasting the stock market and brand equity (Chen et al. 2014).

We also adopt state-of-the-art natural language processing techniques to obtain the sentiment of comments. Following a seminal paper in natural language processing (Socher et al. 2012), we use a recursive neural tensor network (RNTN) on top of the *Stanford Sentiment Treebank* corpus to extract a compositional vector representation of each phrase in the comment.[7] Each comment in this case is classified into positive, negative, or neutral. We measure the daily sentiment by aggregating the sentiments across all comments in the same day. Note that simpler approaches to sentiment analysis—such as bag-of-words classifiers, widely adopted in finance, accounting, and computer science (Pang and Lee 2008 and Devitt and Ahmad 2007)—cannot classify our social media information to a satisfactory precision. Bag-of-words classifiers work well in longer documents by relying on a few words with strong sentiment like "agony." However, unlike financial documents with hundreds of words and written by a relatively homogeneous set of users, comments in social media are very short and written by heterogeneous users. In this case, the classic method does not work. In fact, past literature has shown that bag-of-words classifiers with three levels of sentiment (i.e. positive, negative and neutral) often are below 60% accurate (Wang et al. 2012).

Table 2 summarizes the statistics for the social media data (a full set of social features can be found in Table A2 in Appendix A). On average, the company has 1.11 posts per day during our study period and each post has on average 14.34 words. Each post generates an average of 27 comments by users who respond. An average Facebook user in the US has about 350 friends,[8] indicating that these comments can affect thousands of users per day. Figure 2 presents a time series plot of sales, number of posts, and number of comments, where the data has been normalized for sales and comments. Visual inspection of this figure indicates there appears to be a correlation between social media information and lagged sales.

4. Forecasting Framework and Machine Learning Models

For each model, we implement a variety of machine learning techniques and we compare the accuracy of these models with and without social media information. Section 4.1 discusses the forecasting setting and the overall framework we use to construct, train, and test the forecasting models. Section 4.2 describes the various machine learning methods that we implement.

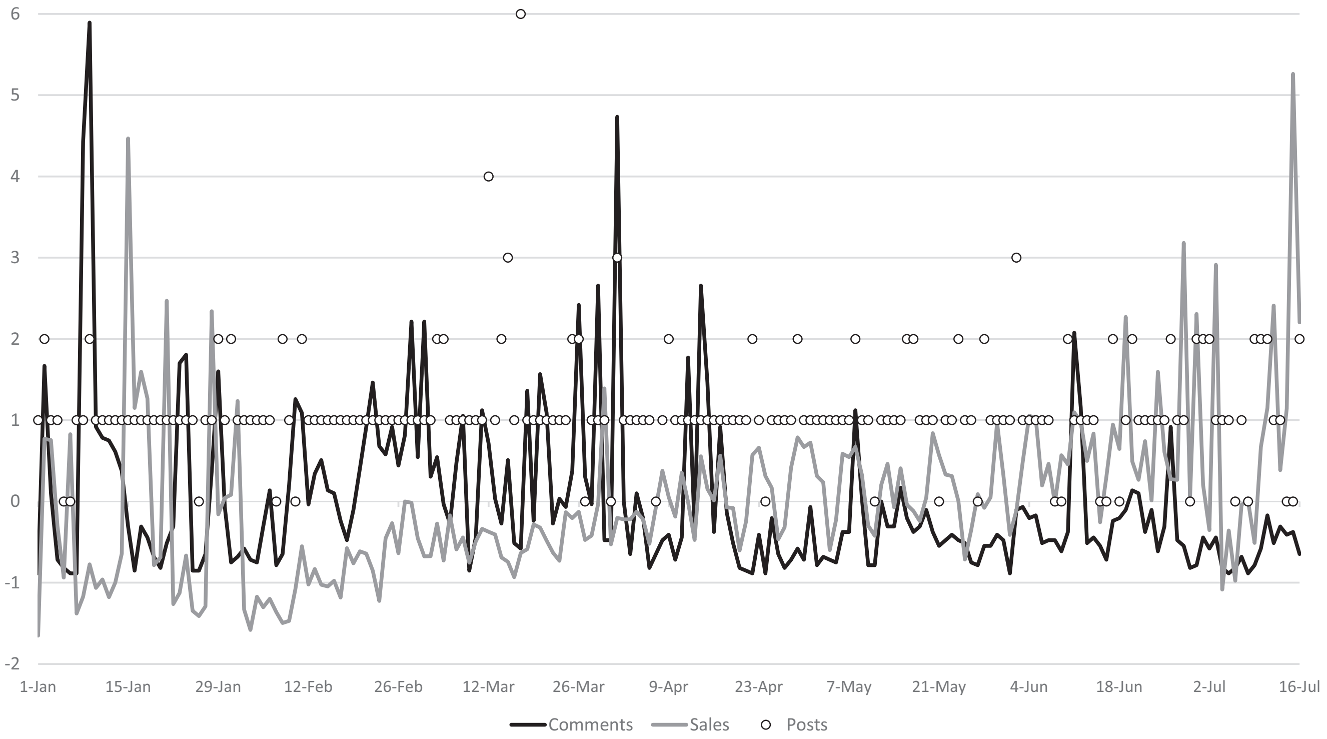
4.1. Forecasting Framework

We construct two forecasts: a "baseline forecast" that includes only operations features as input variables, and a richer forecast, which we call "social media forecast," that includes both operations information and social media information. We fit the same machine learning models for the two forecasts so we have pairs of models where the only difference between them is whether social media information is included. Thus, the difference in forecast accuracy captures the value of social media information.

Table 2 Summary Statistics of Four Social Media Features

Category	Variable	Mean	Median	SD	Minimum	Maximum
Volume data	No. of comments	27.11	18.00	29.40	1.00	200.00
	No. of posts	1.11	1.00	0.70	0.00	6.00
Valence data	Average length of comments	10.74	10.00	5.01	1.00	46.00
	Average sentiment of comments	1.92	1.92	0.25	1.00	3.00

Figure 2 Daily Time Series of Number of Posts, Standardized Sales and Number of Comments



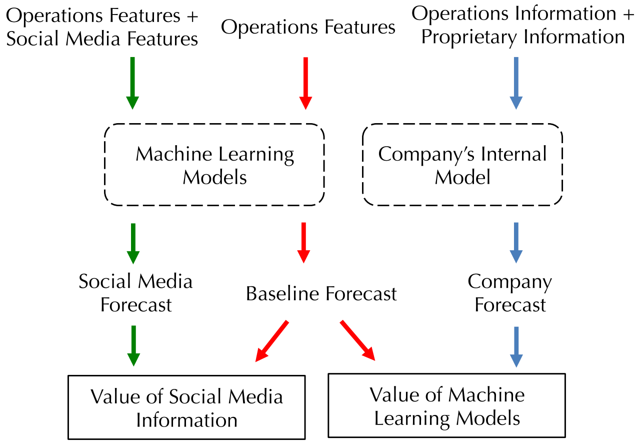
Besides these two daily forecasts, recall that we obtain the company's internal daily forecasts. The company does not consider social media information in their forecasts, as is the case with our baseline forecasts, and they do not use state-of-the-art machine-learning techniques to generate their forecasting models. Therefore, an improvement of our baseline model over the company's benchmark would point to the value of utilizing advanced statistical tools in forecasting. This can be thought of as a lower bound, given that the company has access to more proprietary information than we do. Figure 3 summarizes the differences in the information and tools used in the company's internal model, our baseline model, and our social media model.

Baseline Forecast Model. In the baseline model without social media information, we assume that sales on day t are a function of the sales, promotions, and advertisements in the past week (i.e., past seven days) and the characteristics associated with that t day,

$$S_t = f^{baseline}(S_{t-1}, \ldots, S_{t-7}, A_t, A_{t-1}, \ldots, A_{t-7}, X_t) + \epsilon_t, \quad (1)$$

where X_t are the characteristics specific to day t, such as day of the week; A_t represents advertising variables, such as whether the company runs advertisement or promotions on day t; and ϵ_t is the idiosyncratic demand shock. We include advertising and promotional variables because they are correlated with social media information and these controls help us rule out potential confounding effects. Note that different machine learning models specify different functional forms, so function $f(\cdot)$ can take many forms, such as additive linear, nonlinear, continuous, or discrete.

Figure 3 Forecasting Framework [Color figure can be viewed at wileyonlinelibrary.com]


Operations Features + Social Media Features
Operations Features
Operations Information + Proprietary Information
Machine Learning Models
Company's Internal Model
Social Media Forecast
Baseline Forecast
Company Forecast
Value of Social Media Information
Value of Machine Learning Models

Social Media Forecast Model. In the models that incorporate social media information, we assume the same structure as the baseline model, with the addition of social media variables. That is, current sales are a function of past sales, promotions, advertisements, and social media information over the past seven days, and the characteristics associated with that day X_t,

$$S_t = f^{socialmedia}(S_{t-1}, \ldots, S_{t-7}, A_t, A_{t-1}, \ldots, A_{t-7}, M_{t-1}, \ldots, M_{t-7}, X_t) + \epsilon_t, \quad (2)$$

where M_t represents the social media features in day t, such as the number of comments, the number of words in comments, and the sentiment of average comments. Depending on the specification, S_t can refer to total, repeated, or new sales.

Training and Cross-Validation. To evaluate out-of-sample prediction accuracy, we divide the data into two parts: in-sample training set and out-of-sample testing set. The training set is used to train the model and the testing set is used to evaluate it. We denote the training period as {1, 2, ..., T} and the out-of-sample testing period as {T + 1, T + 2, ..., T + N}, where N is the number of days in the out-of-sample period. In our study, the in-sample training set has 90 days, T = 90, and the out-of-sample testing set has 45 days, N = 45.

We use a cross-validation approach to choose the hyperparameters of our models (such as the number of features to include in the random forest model or the magnitude of λ in the lasso model). For each potential value of the hyperparameter, we evaluate its performance, using the 10-fold cross-validation with three repeats (Kohavi et al. 1995, Stone 1977). The training set is randomly partitioned into 10 equal-size subsets. Of the 10 subsets, a single subset is retained as the validation data for testing the model's performance and the remaining 9 subsets are used as training data. This process is repeated 10 times (the folds), with each of the 10 subsets used exactly once as the validation data. We then average the performance of models in each of 10 subsets. We repeat the process three times to further average out errors. After doing this, we have a measure of the performance for each potential value of the hyperparameter. We choose the value of the hyperparameter that gives the best performance, using grid search. Once we set the best value for the hyperparameter, we use the entire training set to estimate the parameters of the model and the features we retain.

Out-of-Sample Evaluation. Let $\hat{f}^{baseline}_{[1,T]}$ and $\hat{f}^{socialmedia}_{[1,T]}$ denote the baseline and social-media trained models, using the training period (periods 1 to T) data, for which we use the information set $\{S_{1,T}, A_{1,T}, X_{1,T}, M_{1,T}\}$. When we construct forecasts for the out-of-sample testing period, we retrain (re-estimate) parameters and hyperparameters of the model every day. Using the trained model, we make predictions using the most recent information up to the time of the forecast. More specifically, when we forecast sales for day T + 1, we use all past information up to day T, plus the advertising and promotional information on day T + 1[9] to retrain the model. Based on the updated model, we use the past seven days' information and future advertisement plans as input to our model to obtain the forecast. Therefore, to predict the sale at day T + 1, we use the model that is estimated using data from day 1 to day T, that is, $\hat{f}^{baseline}_{[1,T]}$ and $\hat{f}^{socialmedia}_{[1,T]}$. When we predict the sale at day T + 2, we re-estimate the model using data from day 1 to day $T + 1$ and fit the updated model with updated variables by incorporating the data on day T + 1. This rolling updating mechanism in the out-of-sample forecast has been extensively used in the literature (see Osadchiy et al. 2013).

We measure the forecast accuracy using the mean absolute percentage error (MAPE). (See Gaur et al. 2007 and Kesavan et al. 2010 for examples of papers that use these measurements.) In addition, we test whether the differences in forecasting accuracy are statistically significant by performing a t-test of the differences in MAPE, $\frac{1}{N}\sum_{i=1}^{N}\left|S_{t+i} - \hat{S}_{t+i-1,t+i}\right|/S_{t+i}$. Note that while using more information would necessarily result in a smaller or equal error for an in-sample measure of error, this is not the case when we use an out-of-sample measure of error. If the additional data contains irrelevant information, including more information would result in a worse out-of-sample performance due to over-fitting.

L-Day-Ahead Forecasts. In the discussion above, we are constructing sales forecasts for the following day. In practice, companies need lead time to adjust their operational decisions and often want to forecast farther ahead. Consequently, we also construct the L-day-ahead forecast. We estimate how historical information up to t impacts the sale on day t + L,

$$S_{t+L} = f^{baseline}_{L}(S_{t-1}, \ldots, S_{t-7}, A_{t+L}, \ldots, A_{t-7}, X_{t+L}) + \epsilon_t,$$

$$S_{t+L} = f^{socialmedia}_{L}(S_{t-1}, \ldots, S_{t-7}, A_{t+L}, \ldots, A_{t-7}, M_{t-1}, \ldots, M_{t-7}, X_{t+L}) + \epsilon_t.$$

We also verify whether improvements in the accuracy of next-day forecasts are robust to using longer lead times, ranging from one to seven days.

4.2. Machine Learning Models

The approach we describe in section 4.1 is generic and can be used with any forecasting model. In this

Please Cite this article in press as: Cui, R., et al. The Operational Value of Social Media Information. *Production and Operations Management* (2017), doi 10.1111/poms.12707

section, we present the machine learning models considered in our study. We use machine learning models because of the problem's high dimensionality. Our training dataset has 90 days, whereas our social media model needs to estimate more than 280 parameters (40 features per day times the past seven days plus today's characteristics). The number of independent variables is much larger than the sample size, leading to a high-dimensional challenge. Simple linear models, such as moving-average and autoregressive models, cannot produce an identifiable estimate. Advanced variable selection and classification methodologies, such as machine learning models, can solve such problems. Another critical reason to apply machine learning tools is that social media information, advertising, promotions, and even past sales may not affect current sales in a linear fashion. Advanced forecasting models, like random forests, offer a much more flexible structure for the analysis.

In the machine learning literature, forecasting/supervised learning algorithms are classified into four families: (1) linear regression models with regularization, (2) ensemble models based on trees, (3) support vector machines (SVM) with different kernel methods, and (4) neural networks (Friedman et al. 2001). Each algorithm has its own strengths and weaknesses. For instance, generalized linear models tend to demand lower computation costs. Hence, they are often used in large-scale prediction tasks, such as click-through-rate prediction in Google (McMahan et al. 2013). Similarly, boosting models are designed to deal with complex and high-dimensional data. For example, Netflix uses boosting models to improve its movie recommendation system (Bell and Koren 2007).

We implement seven widely adopted statistical models: simple linear regression, lasso, and forward-selection models from category (1), random-forest and gradient-boosting models from category (2), and linear-and-radial-SVM models from category (3). We do not consider neural network models because they are known to perform poorly in small samples with high-dimensional features since these models are very flexible and in turn easy to over-fit. We classify these implemented models along two dimensions: whether the model is linear or nonlinear and whether there is a variable selection procedure to drop insignificant variables, which is summarized in Table 3. We next briefly summarize each model in order of sophistication.

Table 3 Machine Learning Models

	Without variable selection	With variable selection
Linear models	Linear regression	Lasso regression Forward selection
Nonlinear models	SVM with linear kernel	Gradient boosting model (GBM)
	SVM with radial kernel	Random forests

Linear Regression. The simple linear regression model is the simplest model we adopt. It imposes a strict linear structure and does not incorporate variable selection. It uses all variables, even the insignificant ones, to predict the future. The simple linear regression model captures the autoregressive nature of the sales process.

Linear Regression with Forward Selection. Forward selection regression is built on linear regressions, with an additional step to select only important variables in the prediction process. When estimating models $\hat{f}^{baseline}$ and $\hat{f}^{socialmedia}$, we obtain the best estimator by adding the variables that most improve the model, as measured by the Bayesian information criterion (BIC). This maximizes the likelihood function while penalizing the number of parameters included in the model. For details, see Friedman et al. (2001, p. 68).

Lasso Regression. The least absolute shrinkage and selection operator (lasso) model, proposed by Tibshirani (1996), is another type of variable selection regression. It selects variables by penalizing the absolute size of the regression coefficients so that the weak parameter estimates are shrunk towards zero, effectively dropping them. To be specific, for a linear regression $y = \beta_0 + \sum_j \beta_j x_j$, the objective is to minimize $\sum_i (y_i - \beta_0 + \sum_j \beta_j x_i j)^2 + \lambda \sum_j |\beta_j|$, where λ is the hyperparameter that determines how much the estimates are penalized. The value of λ is estimated in the cross-validation step. For details, see Friedman et al. (2001, p. 68).

Support Vector Machines (SVM). This model is used for a variety of classification and regression analyses, such as solving pattern-recognition problems. It maps data into a higher dimensional space and constructs an optimal separating hyperplane (see Cortes and Vapnik 1995). In our study, we use both linear and nonlinear radial kernels. For details, see Friedman et al. (2001, p. 434).

Gradient Boosting Model (GBM). The idea of boosting models is to start with several simple models and combine them in an adaptive way that leads to a stronger predictor. The method starts with a set of classifiers $h_1, \ldots, h_k$ like possible decision trees (called boosting with trees) or regression models (called statistical boosting based on additive logistic regression). Then it creates a classifier that combines those functions $f(x) = sgn(\sum_i a_i h_i(x))$ that can minimize the

Please Cite this article in press as: Cui, R., et al. The Operational Value of Social Media Information. *Production and Operations Management* (2017), doi 10.1111/poms.12707

error. The advantage of a boosting algorithm is theoretically guaranteed good performance for any set of reasonable weak learners. The disadvantage is that the best combination of weak learners is not obvious. In our study, we choose the most widely used boosting model: boosting with trees. For details, see Friedman et al. (2001, p. 339).

Random Forest (RF). Random forest is built on regression trees. Regression trees use a tree-like structure to map observations of an item in order to reach conclusions about the item's target value. The random forest model first uses cross-validation to determine the best number of variables, *m*, to include in a tree. It then draws bootstrap subsamples from the training dataset and grows a regression tree on each of those bootstraps. This is done by bootstrapping variables at each split—randomly selecting *m* variables at each node and picking the best one as a split-point, which generates a diverse set of trees. In this way, we grow a vast number of trees and average those trees in order to yield a prediction. Random forest's biggest advantage is that it is capable of generating highly accurate results, making it one of the most popular statistical learning methods in practice. Figure A1 in Appendix A provides a specific example of regression tree and random forest. For details, see Breiman (2001) and Friedman et al. (2001, p. 588).

5. The Value of Social Media Information in Sales Forecasting

We implement seven machine learning models, described above, to predict sales with and without social media information. Now we can compare the performance between pairs of forecasts—with and without social media information—to evaluate the value of social media information. Table 4 summarizes the out-of-sample MAPE for each of the machine learning models.

Among the studied statistical learning models, random forest performs the best in terms of its forecast accuracy for both the baseline and social media models. In this section we focus on this particular model to discuss the value of social media information and study the robustness of the results with respect to the forecast horizon, testing periods, and forecasting variable (i.e., new and repeat sales). In section 6 we discuss the importance of the social media features, and in section 7 we compare the performance of different machine-learning methods.

Table 4 Comparison of Out-of-Sample Error for Statistical Learning Models

	RF	GBM	SVM radial	SVM linear	Lasso	Forward	Linear
Social media forecast	5.73	9.45	8.31	11.44	14.43	19.62	23.49
Baseline forecast	7.21	11.32	10.59	13.34	19.84	22.27	18.98

Forecasting Result of the Random Forest Model. As discussed in section 4, we conduct the analysis by fitting a random forest algorithm to both our baseline and social media models to generate the baseline and social media forecasts respectively for total sales. We first present our results when we implement the analysis to forecast total sales one day ahead to seven days ahead.

The daily-level out-of-sample forecast accuracy and accuracy improvements for total sales are presented in Figures 4 and 5. Figure 4 portrays how the MAPE of the two forecasts we develop changes with respect to the forecast lead time. Ideally, we would like to compare our *L*-day-ahead forecast with the baseline and social media models to the company's corresponding *L*-day-ahead internal forecast. But such variety of forecasts does not exist. We do not have variation in the number of days ahead with which the company produces its internal forecast. We include the MAPE of the company's forecasts simply as a reference benchmark.[10]

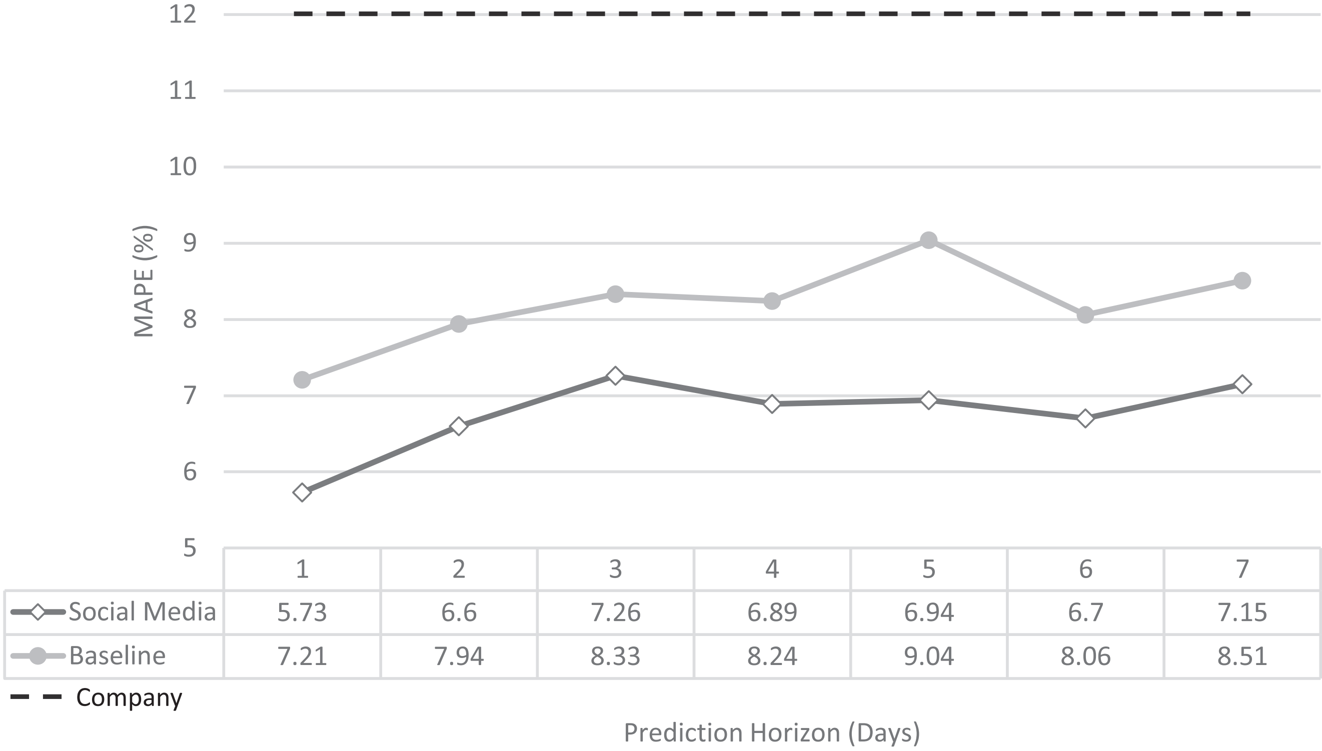
Figure 4 shows that for both the baseline and social media forecasts, the forecast error increases as the forecast lead time increases.[11] This is what we would expect, since future sales are less likely to depend on historical observations and the farther ahead we make forecasts, the harder it is to resolve future uncertainty. When the lead time is five days, the forecast with social media information generates a MAPE of 6.94%, whereas the baseline forecast gives a higher MAPE of 9.04%, with a relative forecast accuracy improvement of 23.23% calculated as:

$$\frac{\textit{Baseline MAPE} - \textit{Social Media MAPE}}{\textit{Baseline MAPE}}$$

The relative improvements of using statistical learning tools and incorporating social media information are summarized in Figure 5. These improvements are statistically significant for MAPE and MSE metrics. This result is consistent across different forecast lead times.

Comparing the difference between the baseline and company forecasts, we obtain an estimate of the value of applying statistical learning tools. The company may use proprietary information we do not observe to aid prediction, so our statistical value evaluated here is likely to be underestimated. The baseline forecasts are all more accurate than company's internal

Figure 4 MAPE (%) Over Prediction Horizon



	1	2	3	4	5	6	7
Social Media	5.73	6.6	7.26	6.89	6.94	6.7	7.15
Baseline	7.21	7.94	8.33	8.24	9.04	8.06	8.51

forecasts, with a relative MAPE improvement ranging from 24.48% to 39.77%, calculated as:

$$\frac{\mathit{Company's\ MAPE} - \mathit{Baseline\ MAPE}}{\mathit{Company's\ MAPE}}$$

If we compare the performance of the social media and company forecasts, we can evaluate the combined benefit of using more advanced methodology and social media information to predict sales. The relative MAPE improvement ranges from 39.35% to 52.13%. This result further strengthens and completes our key message: social media information, which is publicly available and easy to obtain, adds value in predicting sales.

Robustness with Respect to New and Repeat Sales. We conduct the same analysis for sales coming from repeat and new customers and present the results in Table A3 in Appendix A. We observe a result similar to the one presented above. There is significant value of social media information across all the different lead times, with a MAPE improvement ranging from 17.51% to 25.30% for repeat customers and from 14.87 to 28.62 for new customers.

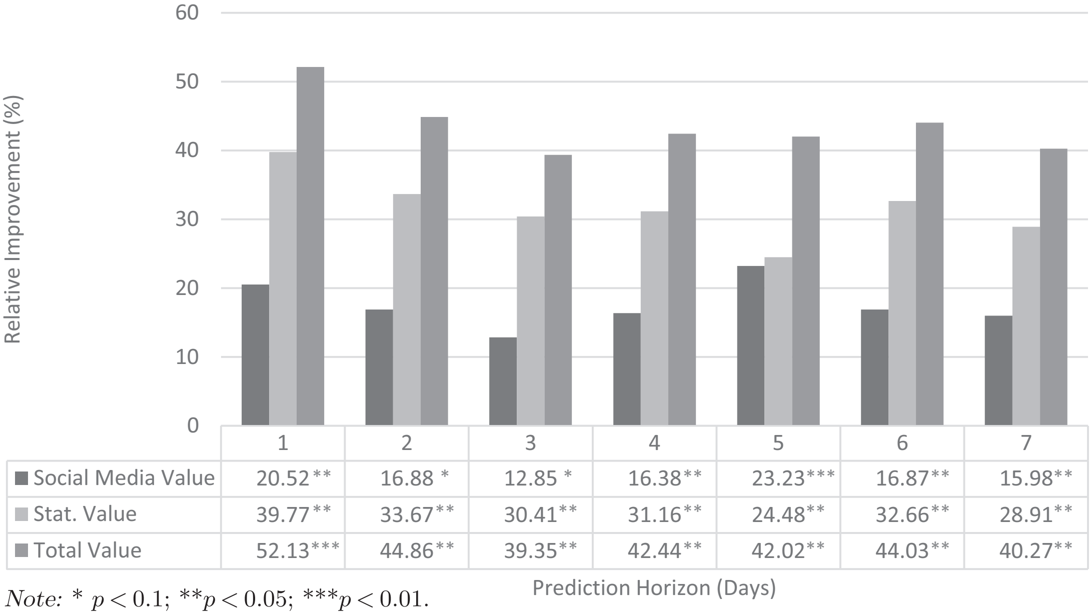
Robustness across Different Time Periods. The analysis obtained so far is based on the training set from January 1 to March 31, 2013, and the testing set from April 1 to May 16, 2013. To validate the robustness of the findings across different time periods, we vary the starting time of the training set and repeat the analysis. We display the one-day, three-day, five-day, and seven-day-ahead forecasts in Table A4 in Appendix A. The result indicates social media information consistently leads to significant forecast improvements, proving the robustness of the results regardless of the time period we study.

To summarize, we have shown that social media information brings statistically significant improvements to sales forecasting. The result is robust across many factors.

6. The Importance of Social Media Features

Social media information can play a critical role in affecting buyers' behavior by attracting their attention or signaling product quality to them. In our setting, the company's posts could be photos of new arrivals and questions to encourage engagement from users. When a Facebook user comments on a company's post, this action will appear in her friends' news feeds, which helps broadcast product information and increases product awareness. In this process, the user also disseminates her sentiments to her friends. By providing a positive and informative comment on

Figure 5 Relative Forecast Improvement Over Prediction Horizon



	1	2	3	4	5	6	7
■ Social Media Value	20.52**	16.88 *	12.85 *	16.38**	23.23***	16.87**	15.98**
■ Stat. Value	39.77**	33.67**	30.41**	31.16**	24.48**	32.66**	28.91**
■ Total Value	52.13***	44.86**	39.35**	42.44**	42.02**	44.03**	40.27**

Note: * $p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Facebook, users send a favorable signal or convey positive emotions about the product to their friends and followers. The users' friends can see these activities on their news feeds and follow the link to visit the company's Facebook page or official website. Consequently, this increased traffic might translate into increased sales. We next study how important/significant the social media variables are and which social media features are most relevant in the model.

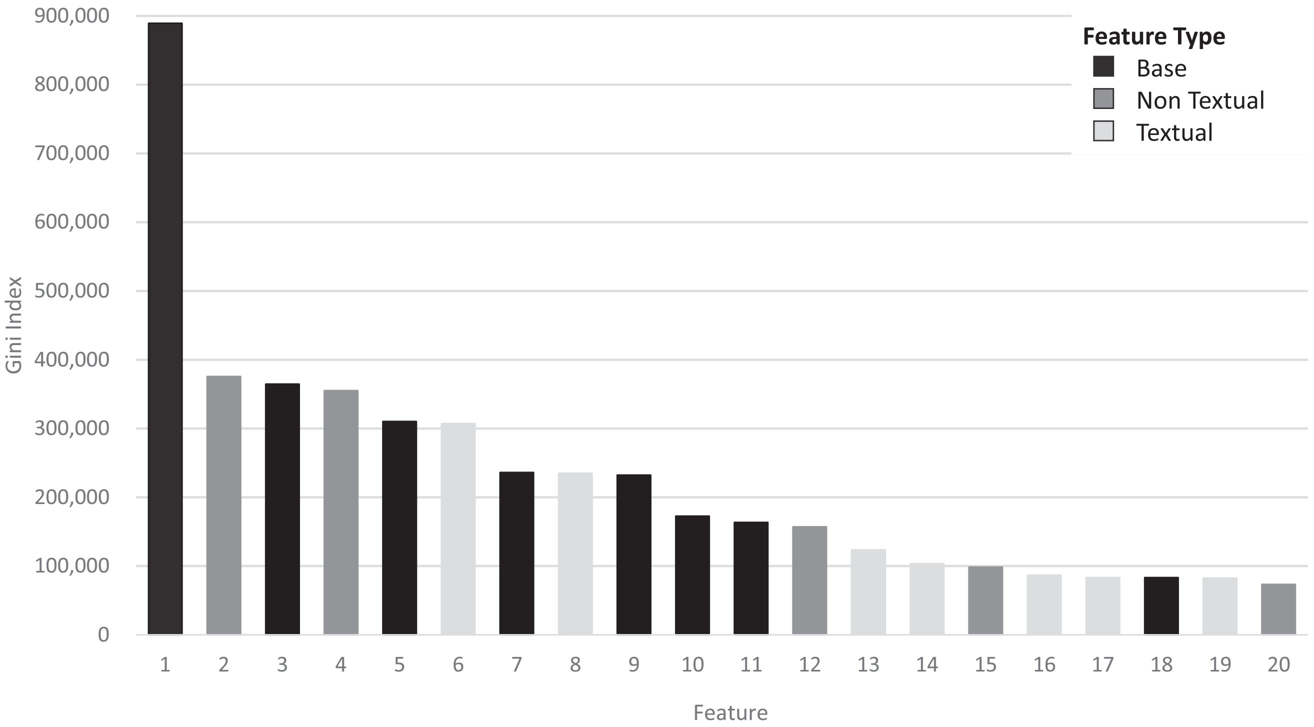
We do so by exploring all variable's rank of importance in the random forest model. The random forest model selects important features from unimportant ones by placing the former higher in the regression tree. For each tree, a feature's importance is defined based on a metric called the Gini impurity index. Based on the value of a feature in a tree, the tree will split into different subsets of other features or the outcome. The Gini importance of a feature split in the regression tree measures how often an element from the set—before the split—would be misclassified if it is classified randomly according to the distribution of labels in the whole set. A natural way to measure a feature's importance in random forests, denoted as Gini importance, is the weighted average Gini index of that feature across all regression trees.[12]

In Figure 6, we rank the Gini importance of basic operational features, volume (i.e., attention) social media features and valence (i.e., endorsement), and social media features for the one-day-ahead random forest model. (The detailed index values are in Table A5 in Appendix A.) The top five relevant variables are last day's sale, number of comments six days ago, whether there is a promotion tomorrow, number of unique users commenting six days ago, and sales seven days ago. The result is intuitive: historical sales predict the trend of future sales, promotional activities capture the potential spike, and, last but not least, social interactions on the company's Facebook page are an important indicator of the firm's future sales. We can observe that among the top 20 features, 12 of them are social media features, further confirming the importance of their impact on sales.[13] Moreover, there are more valence than volume features among the top 20 features. This argues for the importance of using natural language processing techniques to extract valence features from the observed social media information.

It is also interesting to see the performance of the algorithm with only volume and only valence features to see how endorsement and attention mechanisms affect the forecasting outcomes. Table A6 in Appendix A compares the performance of our random forest model with all social media features, only valence (endorsement) features, only volume (attention) features, and no social media features (baseline). For all forecast lead times, including either attention or endorsement features can significantly improve performance. Moreover, the magnitude of performance improvements through either set of features is comparable, suggesting that both mechanisms are important in demand forecasting. Including both features will result in significant performance improvement, which suggests that interactions among these two sets of features are also important.

Please Cite this article in press as: Cui, R., et al. The Operational Value of Social Media Information. *Production and Operations Management* (2017), doi 10.1111/poms.12707

Figure 6 Top 20 Features in Random Forest with the Highest Gini Importance



We check the most relevant variables for both new and repeat sales. They are quite similar, with only one difference: promotion does not have a high Gini importance when predicting sales from new customers, whereas it is important in forecasting sales from repeat customers. This is because the major channel the company uses for promotions is email, sent to customers who subscribe to it. New customers do not receive promotional notifications from the company and therefore their purchases do not depend much on promotional information. This side finding further confirms the validity of the studied statistical learning model—only the most relevant variables are selected.

7. Comparison across Statistical Learning Models

We have explicitly shown the forecasting outcomes for the random forest model. We now compare across all seven statistical learning models and provide guidance on which are best for the problem we are studying.

7.1. Bias-Variance Trade-Off

We adopt a well-established bias-variance trade-off framework in machine learning and statistics literature (Friedman et al. 2001, p. 223). The framework assumes the data generation process of future sales is $Y = f(X) + \epsilon$, where X is the set of observed features and ϵ is a random variable with mean 0 and variance σ_ϵ^2. A forecasting model based on a series of past observations $q \in Q$ can be expressed as $\hat{f}_q(\cdot) : X \rightarrow R$. Given any input $x_0 \in X$, the expected mean squared error for estimating $f(x_0)$ is

$$\begin{aligned}\mathrm{MSE}(x_0) &= E[(Y - \hat{f}_q(x_0)^2 | X = x_0)]\\ &= \epsilon_\sigma^2 + E[\hat{f}_q(x_0) - f(x_0)]^2 + E[(\hat{f}_q(x_0) - E_Q[\hat{f}_q(x_0)])^2]\\ &= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}.\end{aligned}$$

The above equation tells us that in order to minimize a model's prediction error, we need to minimize the model's bias and variance. *Bias* represents the error introduced by approximating an unknown data generation process $f(\cdot)$ with the best fit $\hat{f}_q(\cdot)$ based on a model and a series of training data. For example, if the underlying data generation process is nonlinear, a model that uses linear approximation would have very large biases.

In general, less flexible models have higher biases. On the other hand, *variance* refers to the amount by which $\hat{f}_q(\cdot)$ changes in regard to a different training set q. It represents the error introduced by estimating a model based on a realized finite training set. If a model varies greatly with the training data, it has high variance and in turn will not perform well in prediction. In general, more flexible models have higher variances.

Please Cite this article in press as: Cui, R., et al. The Operational Value of Social Media Information. *Production and Operations Management* (2017), doi 10.1111/poms.12707

Table 5 Comparison of In-Sample and Out-of-Sample Error for Statistical Learning Models

	Out-of-sample MAPE (%)		In-sample MAPE (%)	
Model	Social media out-of-sample error	Baseline out-of-sample error	Social media In-sample error	Baseline in-sample error
RF	5.73	7.21	5.28	7.59
GBM	9.45	11.32	10.32	10.79
SVM radial	8.31	10.59	10.17	13.61
SVM linear	11.44	13.34	10.18	14.04
Lasso	14.43	19.84	14.33	24.91
Forward	19.62	22.27	23.48	26.92
Linear	23.49	18.98	7.04	23.78

A lower variance in a model usually leads to a higher bias, and vice versa. For instance, random guessing the future sales equal to 500 is a model that has 0 variance since it does not depend on the past training data; however, it has high bias (i.e., the bias is equal to the mean squared distance of the realized sales toward 500). Similarly, if a model is degenerate in training data (i.e., $\hat{f}_q(x_0) = y_0 \ \forall (y_0, x_0) \in q$), the model has 0 bias for any set of training data. However, the model has high variance since a small perturbation in the training data can result in a large variation in the estimated model.

To analyze the model performance empirically, researchers often use in-sample and out-of-sample errors as proxies for biases and variances (Friedman et al. 2001 and James et al. 2013). Models, based on their errors, are classified into three categories: (1) under-fitting: in-sample and out-of-sample errors are both high, representing the model has a large bias; (2) over-fitting: in-sample error is low while out-of-sample error is large, representing the model has a large variance; and (3) good fitting: out-of-sample error is only slightly higher than in-sample error and both are low. Table 5 presents the in-sample and out-of-sample error comparison across different machine learning models. Random forest is the model that best fits both in-sample and out-of-sample[14] and linear regression has the most serious over-fitting problem.

7.2. Forecasting Results of other Statistical Learning Models

Figure 7 plots the out-of-sample MAPE performance for the baseline and social media forecasts across the seven machine learning models discussed in section 4. For each model, the lighter column represents the base line forecast and darker line represents the social media forecast. in Figure 8 present detailed statistics on the forecasts' improvement.

Figure 7 shows that the social media forecast outperforms the baseline forecast for all statistical models except linear regression, which further validates our main finding. That is, social media information is

Figure 7 MAPE (%) for Different Statistical Learning Models — Prediction Horizon = 1

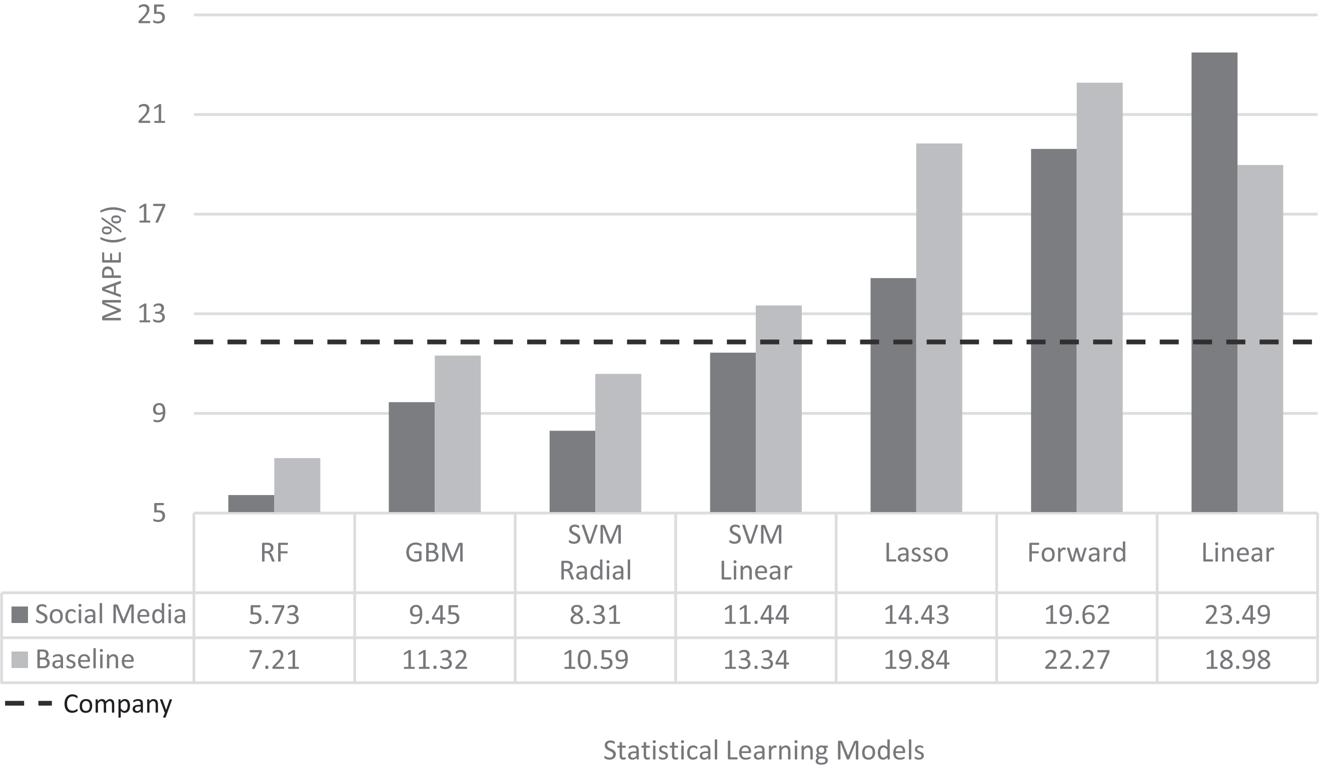
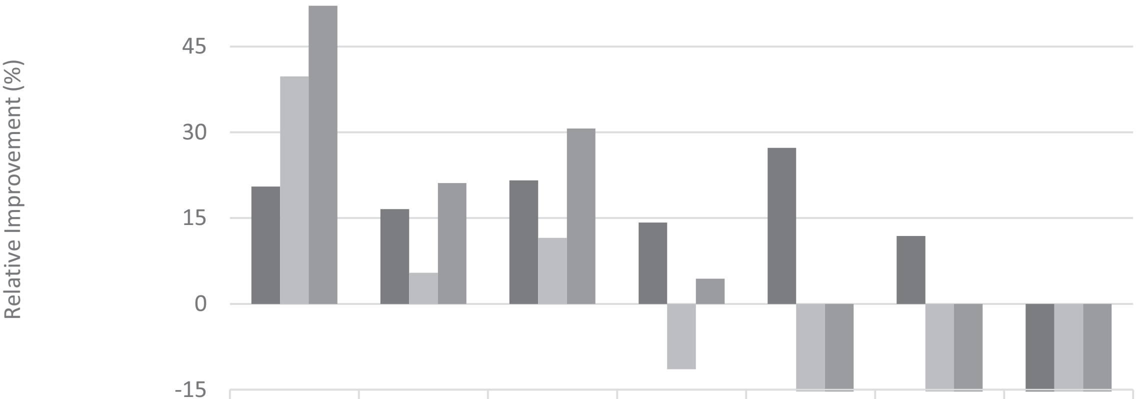


Figure 8 Relative Forecast Improvement for Different Statistical Learning Models — Prediction Horizon = 1



	RF	GBM	SVM Radial	SVM Linear	Lasso	Forward	Linear
■ Social Media Value	20.52**	16.58 *	21.58 *	14.21	27.27**	11.87 *	-23.75***
■ Stat. Value	39.77**	5.43	11.53 *	-11.45 *	-65.75***	-63.91***	-58.56***
■ Total Value	52.13***	21.12 **	30.63 *	4.42	-20.51**	-63.85***	-96.14***

Statistical Learning Models

Note: * $p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

shown to be valuable as long as it is incorporated in a forecasting framework that is appropriate to handle this information. The random forest model beats all other statistical learning tools in forecasting sales, which is consistent with its reputation of high accuracy. This is the model we would recommend companies use in their forecasts, particularly if they consider incorporating social media information. Although its estimation time might be on the longer side, most companies do not need to update (retrain) the model every period. Instead, they could retrain the model every several periods, which also provides additional robustness to abnormal input data.

All the nonlinear models (i.e., support vector machine models, gradient boosting models, and random forest) outperform the linear models (i.e., linear regressions, forward selection, and lasso model) for both the baseline forecast without social media and the social media forecast.[15] The linear regression model without variable selection performs worse when social media information is included. Our finding aligns with that of Ferreira et al. (2015), where the authors find non-parametric regression technique outperforms many of the parametric demand prediction models.

The main reason for the poor performance of the linear models is that forecasting problems with social media information are high-dimensional problems. It is well known that linear models do not approximate the underlying data generation process very well in high-dimensional problems since linear models tend to over-fit the data. Indeed, Table 5 shows that the in-sample error of the linear regression model is half its out-of-sample error, while the in-sample and out-of-sample errors of the random forest model are similar. This evidence of over-fitting indicates that social media information and some operations information, such as promotions, may influence current sales in a nonlinear fashion. Nonlinear models allow for a more flexible structure, are better able to capture the relationship between those variables, and generate more accurate out-of-sample sales predictions.

Figure 7 shows that linear regression models have a negative value for social media information, while other linear models with feature selection present positive values. This is not surprising since many social media features included in our models are weakly correlated, if not uncorrelated, with future sales. However, as a model designer it is difficult to know a priori which features are more important in predicting future sales. Therefore, a model should actively select the important social media features that make the model a better approximation of the underlying data-generating process and thus less likely to over-fit. Table 5 shows that linear models with feature selection, e.g., lasso and forward selection models, have in-sample and out-of-sample error in the same range, while the linear regression model without feature selection is over-fitted.

We conclude that the most effective models to use when incorporating social media information in sales forecasting are methods based on *nonlinear* models *with feature selection*.

8. Discussion and Conclusions

We have studied the value of using social media information to improve a firm's sales forecasts and have shown how incorporating social media information with sales forecasting results in statistically significant improvements. These results apply to out-of-sample forecast tests for both the full dataset and different subsamples of the data, such as new and repeat customers, and different training periods. The results are also robust to a number of different statistical models.

Improving sales forecasts accuracy can lead to substantial operational benefits in a variety of situations. A good forecast can be translated into a reduced safety inventory cost,[16] smoother scheduling of staff, and more consistent delivery and production planning, especially for fast-moving goods. Accurate forecasts can be extremely valuable for retail "fast fashion" companies, such as H&M, Zara, or Peacocks. These companies place heavy emphasis on quick reaction to demand changes and have developed supply chains that are ready to deliver products to their stores twice a week. The value of a better sales forecast to these companies can be substantial.

Having better forecasts can be valuable even when the company cannot influence its supply within the forecast horizon, since it can often work with the demand side of the business. A more accurate forecast can help the company react to customers' information and adjust its promotional or communications efforts, for example, to better match its inventory position with actual demand. Improving accuracy in forecasting can be extremely valuable in a dynamic pricing environment like today's online retailing. When thinking about dynamic pricing in online retailing, a one-to-seven-days-ahead forecast horizon is actually a long period, especially given evidence of online retailers changing prices multiple times a day. In addition, most of these types of retailers offer the option of buying online. The retailer can start collecting information, in terms of demand and social media content before the products actually arrive at the stores, making the forecasting task even more meaningful.[17]

Conversations with the partner retailer highlight the important practical implications of our results. In their context, short-term forecasts are used to, for example, adjust their pricing and promotion decisions, decide on advertising investments, and adjust online store layout. Being a pure-play retailer, they are constantly focusing their communications effort on Facebook and other media platforms. They were happily surprised and very interested in the improvements that can be obtained from incorporating social media in their forecasting process. After we shared our results, they put some processes in place to systematically track social media information and feed their daily forecasting model with it. The main operational benefits they would obtain from improved short-term forecasting are in pricing, promotions, advertising, store layout, and decisions about reallocating inventory across stores decisions.

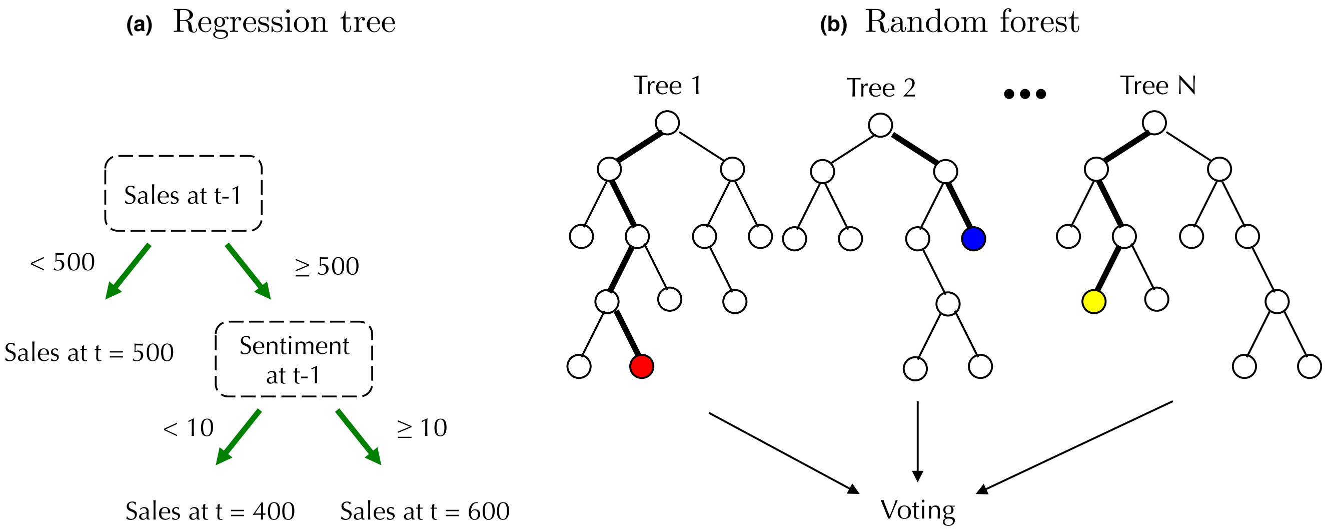
Our study could be generalized in different ways. It mainly deals with short-term forecasts, but similar methodology could be adopted and used for longer-term forecasts. Also, while our study focuses on the total sales forecast (top-level forecast), similar models could be used to generate more granular forecasts. As different parts of the company may require different levels of detail—e.g., regional sales, sales by category, or SKU-level sales—one could apply the widely adopted *top-down* approach to break down the top-level forecasts to bottom-level forecasts[18] (see Ord and Fildes 2012, p. 406). The significant forecast improvements at the total sales level can translate into increased forecast accuracy at the detailed sales level, which may facilitate managerial decisions for different departments in a company. Alternatively, a company could also adapt the approach we have used in our aggregate daily forecasts to directly build forecasts at different product or time aggregation levels. For example, a company may be interested in creating product-category-level or SKU-level forecasts for inventory planning purposes. This would require collecting operational and social media data at the desired level of aggregation. The sales and advertising data are usually available to the company at the product level. Obtaining social media data at the product or category level can be more challenging, as much of the social media data is not linked to a specific product. To generate product-level social media data, the company could create product-level content that allows assigning social media interactions to a specific product. For example, the company could create posts or show ads featuring different products on its Facebook page and track the interaction with each product to obtain product-level social media data. Once the data is available at the desired level, the same forecasting approach we used for aggregate sales could be applied with the more granular data.

Dealing with social media data comes with challenges. An important challenge for both practitioners and academics is that a significant amount of social media information has an unstructured, textual nature, which is not the type of data that academics and practitioners traditionally have dealt with. In our case, for example, we used natural language processing techniques to encode the sentiment of comments users make in social media. Once these initial challenges are overcome, there is a huge potential to use this data in novel ways.

This new data can make firms more efficient while providing a better experience to their customers. For example, in a novel application of social media information, Nordstrom has started to feature in its brick-and-mortar stores the products that are receiving the most "Pins" in Pinterest by users from nearby locations.[19] It is only a matter of time until more firms start to use this type of information to routinely improve their assortment decisions and enhance their customers' experience. With the development of these novel applications, we expect the operations management community to pay increasing attention to the value of social media information. We hope our project inspires others to conduct further studies, using structured and unstructured data from social media information.

Appendix A. Tables and Figures

Figure A1 Illustration of Regression Tree and Random Forecast [Color figure can be viewed at wileyonlinelibrary.com]



Notes. In (a), the root starts with sales on day t−1. Depending on whether it is less or more than 500, we move on to different subgroups of features. If the last day's sales are on the left branch, the tree reaches one end leaf with the current day's sale being 500. Otherwise, we have to further split the subgroup based on the node of sentiment on day t − 1. An actual tree can be deeper: the subgroups are split until the groups are sufficiently homogeneous or pure—that is, a single class of features predominates a group. In (b), each tree is based on a random bootstrap sample. Given a new observation, we run this observation through tree 1 and get a particular prediction outcome. The same observation is run through tree 2, which directs a different path, leading to a different prediction. We continue this process until tree N, where N = 500 in our study. We then average those predictions together as the final forecast.

Table A1 All Daily Basic Features

Feature name	Explanation
Sales	Daily number of sales
Ad	A dummy variable equal to 1 if there is an advertisement email
Promotion	A dummy variable equal to 1 if there is a promotion email
PromotionLast	A dummy variable equal to 1 if there is a promotion
WeekdayMonday	A dummy variable equal to 1 if it's Monday
WeekdayTuesday	A dummy variable equal to 1 if it's Tuesday
WeekdayWednesday	A dummy variable equal to 1 if it's Wednesday
WeekdayThursday	A dummy variable equal to 1 if it's Thursday
WeekdayFriday	A dummy variable equal to 1 if it's Friday
WeekdaySaturday	A dummy variable equal to 1 if it's Saturday
WeekdaySunday	A dummy variable equal to 1 if it's Sunday
Holiday	A dummy variable equal to 1 if it's a national holiday

Table A2 All Daily Social Media Features

Category	Feature name	Explanation
Volume	c_comments_count	Number of comments from the company
	c_unique_comments_posts	Number of unique posts that the company comments on
	u_comments_count	Number of comments from users
	u_unique_comments_posts	Number of unique posts that users comment on
	unique_comments_users	Number of unique users who commented
	posts_counts	Number of unique posts
	photo_posts	Number of unique posts with photos
	status_posts	Number of unique status posts
	offer_posts	Number of unique posts with offers
	note_posts	Number of unique posts that are notes
	event_posts	Number of unique posts with events
	link_posts	Number of unique posts with links
	video_posts	Number of unique posts with videos
Valence	c_avg_num_words	Average number of words of the company's comments
	c_avg_comments_size	Average byte size of the company's comments
	u_avg_num_words	Average number of words per comments by users
	u_avg_comments_size	Average byte size of users' comments
	avg_post_num_words	Average number of words in the post
	avg_post_size	Average byte size of posts
	c_avg_sentiment	Average sentiment of the company's comments
	u_avg_sentiment	Average sentiment of users' comments

Table A3 Forecast Accuracy Improvements for New and Repeat Customers

	Repeat customers			New customers		
Forecast lead time	Social media	Baseline	Relative improvement (%)	Social media	Baseline	Relative improvement (%)
$L = 1$	7.00	9.37	25.30***	4.44	6.22	28.72***
$L = 2$	9.13	11.17	18.30**	4.83	6.67	27.58***
$L = 3$	9.20	11.38	19.19**	5.88	6.69	12.04*
$L = 4$	9.01	11.11	18.86**	5.92	7.04	15.97*
$L = 5$	9.21	11.92	22.79**	6.07	7.13	14.87**
$L = 6$	9.08	11.62	21.82***	6.03	7.14	14.95**
$L = 7$	8.81	10.68	17.51**	5.93	7.06	16.06**

Notes. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. The result shows the robustness relative to the type of customers.

Table A4 Sales Forecast Improvements with Different Starting Dates of the Training Set

Forecast lead time		$L = 1$			$L = 3$		
Training end	Testing end	Social media	Baseline	Relative improvement (%)	Social media	Baseline	Relative improvement (%)
04-01	05-16	5.73	7.21	20.57***	7.26	8.33	12.80*
04-16	05-31	6.39	7.40	13.72*	6.98	7.37	5.22
05-01	06-15	5.42	7.31	25.80***	6.12	7.81	21.63***
05-16	06-30	6.07	7.52	19.25**	5.79	7.52	22.90***
06-01	07-15	10.36	12.73	18.61**	10.06	12.69	20.71***

Forecast Lead Time		$L = 5$			$L = 7$		
Training End	Testing End	Social Media	Baseline	Relative Improvement (%)	Social Media	Baseline	Relative Improvement (%)
04-01	05-16	6.94	9.04	23.26***	7.16	8.51	15.90**
04-16	05-31	7.22	8.06	10.47*	7.13	7.67	7.15
05-01	06-15	6.74	8.36	19.43**	5.80	6.86	15.49*
05-16	06-30	6.09	7.85	22.48***	5.84	8.28	29.46***
06-01	07-15	9.52	12.53	24.01***	9.73	12.63	23.00***

$^{*}p < 0.1$; $p < 0.05$; $p < 0.01$. The result shows the robustness relative to time.

Table A5 Top 20 Features in Random Forest with the Highest Gini Importance

Rank	Feature name	Gini importance	Rank	Feature name	Gini importance
1	salesLag1	889,058	11	salesLag6	163,499
2	comments_countLag6	375,627	12	unique_commentsLag7	157,133
3	promotion	364,691	13	post_wordsLag2	123,300
4	unique_commentsLag6	354,859	14	num_wordsLag1	103,422
5	salesLag7	310,088	15	comments_countLag7	98,226
6	post_wordsLag3	307,044	16	avg_sentimentLag1	86,421
7	promotionLast	235,860	17	b_num_wordsLag2	83,326
8	post_wordsLag5	234,949	18	salesLag5	83,021
9	salesLag2	232,365	19	b_num_wordsLag5	82,321
10	salesLag3	172,468	20	comments_countLag4	73,427

Table A6 Comparison of Out-of-Sample MAPE (%) for Different Feature Sets

Lag	Social media (%)	Social media with attention features (%)	Social media with endorsement features (%)	Baseline (%)
1	5.73	6.20	6.15	7.21
2	6.6	6.90	6.73	7.94
3	7.26	7.66	7.60	8.33
4	6.89	7.33	7.23	8.24
5	6.94	7.43	7.33	9.04
6	6.7	7.34	7.27	8.06
7	7.15	7.32	7.23	8.51

Notes

[1] We obtain this information from the company's official Facebook page using the Facebook public application programming interface (API). Facebook is the largest social media channel for the focal company: over 90% of the company's customers who visit its official website from social media websites come from Facebook.

[2] The statistics and computer science communities use slightly different terminology. In statistics, these methods usually are referred to as *statistical learning* methods; in computer science, they usually called *machine learning*. We use both terms interchangeably.

[3] http://blog.jda.com/the-impact-of-social-media-on-the-supply- chain-is-there-one/

[4] These emails notify customers of promotions the company is running or going to run in the near future.

[5] We focus on whether there was a promotion on that day because promotion depth is relatively homogeneous over time (between 15% and 20%).

[6] In this particular example, we use a post from Gap to protect the anonymity of our company partner .

[7] We train the RNTN classifier with 215,154 phrases in the *Stanford Sentiment Treebank* dataset. See http://nlp.stanford.edu/sentiment/treebank.html

[8] http://www.fool.com/investing/general/2015/03/28/the-average- american-has-this-many-facebook-friend.aspx

[9] Since advertising schedules and promotional events are planned in advance, the future information is available on day *T*.

[10] The company generates daily forecasts by making a weekly forecast in the week before and then decomposing the weekly forecast into daily forecasts. Thus, we can think of the company's forecasts as having been produced with an average lead time of about L = 4.

[11] Note the sixth step-ahead forecast error drops slightly. However, the decrease is not statistically significant.

[12] This metric of Gini importance is widely used in the random forest literature in assessing feature importance (Breiman 2001). Louppe et al. (2013) further discuss the theoretical performances of Gini importance.

[13] We also check the variables selected by the lasso model and find all the top 20 features with highest Gini importance are kept by the lasso model. This suggests that these features are important for predicting future sales regardless of the underlying machine learning model.

[14] Note that for some models, the in-sample error is larger than the out-of-sample error. This may be because the loss function used to train the models is RMSE instead of MAPE.

[15] In addition to the models reported here, we have explored an autoregressive integrated moving average with exogenous covariates (ARIMAX), which assumes a linear relation. Its performance is also worse than the nonlinear models reported in Table 5.

[16] For example, using the generalized demand model Martingale Model of Forecast Evolution (MMFE) and the general order-up-to inventory policy (proposed by Chen and Lee 2009, Graves 1986, Graves et al. 1986, Heath and Jackson 1994, and developed by Chen and Lee 2009), it is easy to show that improvement in sales forecasting translates to a safety stock reduction. The detailed theoretical analysis is available from authors upon request.

[17] For related settings, see Moe and Fader (2002) and Huang and Van Mieghem (2014).

[18] The relative market shares across product categories or SKUs usually are used as the basis for the decomposition.

[19] https://business.pinterest.com/en/success-stories/nordstrom

References

Allon, G., D. J. Zhang. 2015. Managing service systems in the presence of social networks. Available at SSRN, 2673137.

Ang, E., S. Kwasnick, M. Bayati, E. L. Plambeck, M. Aratow. 2015. Accurate emergency department wait time prediction. *Manuf. Serv. Oper. Manag.* **18**(1): 141–156.

Aral, S. 2011. Commentary-identifying social influence: A comment on opinion leadership and social contagion in new product diffusion. *Market. Sci.* **30**(2): 217–223.

Banerjee, A., A. G. Chandrasekhar, E. Duo, M. O. Jackson. 2012. The diffusion of microfinance. Technical report, National Bureau of Economic Research.

Barber, B. M., T. Odean. 2008. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Rev. Finan. Stud.* **21**(2): 785–818.

Bassamboo, A., R. Cui, A. Moreno. 2015. The wisdom of crowds in operations: Forecasting using prediction markets. Available at SSRN, 2679663.

Bell R. M., Y. Koren. 2007. Lessons from the net ix prize challenge. *ACM SIGKDD Explor. Newsl.* **9**(2): 75–79.

Bikhchandani, S., D. Hirshleifer, I. Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *J. Polit. Econ.* **100**(5): 992–1026.

Bollen, J., H. Mao, X. Zeng. 2011. Twitter mood predicts the stock market. *J. Comput. Sci.* **2**(1): 1–8.

Breiman, L. 2001. Random forests. *Mach. Learn.* **45**(1): 5–32.

Cai, H., Y. Chen, H. Fang. 2007. Observational learning: Evidence from a randomized natural field experiment. *National Bureau of Economic Research*, w13516.

Candogan, O., K. Bimpikis, A. Ozdaglar. 2012. Optimal pricing in networks with externalities. *Oper. Res.* **60**(4): 883–905.

Chen, L., H. L. Lee. 2009. Information sharing and order variability control under a generalized demand model. *Management Sci.* **55**(5): 781–797.

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

Chen, Y., Q. Wang, J. Xie. 2011. Online social interactions: A natural experiment on word of mouth versus observational learning. *J. Market. Res.* **48**(2): 238–254.

Chen, H., P. De, Y. J. Hu, B-H. Hwang. 2014. Wisdom of crowds: The value of stock opinions transmitted through social media. *Rev. Financ. Stud.* **27**(5): 1367–1403.

Cortes, C., V. Vapnik. 1995. Support-vector networks. *Mach. Learn.* **20**(3): 273–297.

Cui, R., G. Allon, A. Bassamboo, J. A. Van Mieghem. 2015. Information sharing in supply chains: An empirical and theoretical valuation. *Management Sci.* **61**(11): 2803–2824.

Da, Z., J. Engelberg, P. Gao. 2011. In search of attention. *J. Finance* **66**(5): 1461–1499.

Devitt, A., K. Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *ACL*.

Ferreira, K. J., B. H. A. Lee, D. Simchi-Levi. 2015. Analytics for an online retailer: Demand forecasting and price optimization. *Manuf. Serv. Oper. Manag.* **18**(1): 69–88.

Friedman, J., T. Hastie, R. Tibshirani. 2001. *The Elements of Statistical Learning*, volume **1**. Springer series in statistics Springer, Berlin.

Gaur, V., A. Giloni, S. Seshadri. 2005. Information sharing in a supply chain under arma demand. *Management Sci.* **51**(6): 961–969.

Gaur, V., S. Kesavan, A. Raman, M. L. Fisher. 2007. Estimating demand uncertainty using judgmental forecasts. *Manuf. Serv. Oper. Manag.* **9**(4): 480–491.

Goh, K.-Y., C.-S. Heng, Z. Lin. 2013. Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Inf. Syst. Res.* **24**(1): 88–107.

Graves, S. C. 1986. A tactical planning model for a job shop. *Oper. Res.* **34**(4): 522–533.

Graves, S. H. C. Meal, S. Dasu, Y. Qiu, S. Axsäter, C. Schneeweiss, E. Silver. 1986. Two-stage production planning in a dynamic environment. *Multi-Stage Prod. Plann. Inventory Control* **266**: 9–43.

Hameed, B. 2011. Social media usage exploding amongst fortune 500 companies. Available at http://socialtimes.com (accessed date November 24, 2013).

Heath, D. C., P. L. Jackson. 1994. Modeling the evolution of demand forecasts ith application to safety stock analysis in production/distribution systems. *IIE Trans.* **26**(3): 17–30.

Huang, T., J. A. Van Mieghem. 2014. Clickstream data and inventory management: Model and empirical analysis. *Prod. Oper. Manag.* **23**(3): 333–347.

Ifrach, B., C. Maglaras, M. Scarsini. 2011. Monopoly pricing in the presence of social learning.

James, G., D. Witten, T. Hastie, R. Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer, New York.

Jing, B. 2011. Social learning and dynamic pricing of durable goods. *Market. Sci.* **30**(5): 851–865.

Kesavan, S., V. Gaur, A. Raman. 2010. Do inventory and gross margin data improve sales forecasts for us public retailers? *Management Sci.* **56**(9): 1519–1533.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, August 20–25, 1995, Montreal, Quebec, Canada, pp. 137–1143.

Kremer, M., B. Moritz, E. Siemsen. 2011. Demand forecasting behavior: System neglect and change detection. *Management Sci.* **57**(10): 1827–1843.

Li, X., L. Wu. 2014. Herding and social media word-of-mouth: Evidence from groupon. Working paper, Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2264411

Louppe, G., L. Wehenkel, A. Sutera, P. Geurts. 2013. Understanding variable importances in forests of randomized trees. In *Adv. Neural Inf. Process. Syst.* 431–439.

Luca, M. 2011. Reviews, reputation, and revenue: The case of yelp. com. Technical report, Harvard Business School.

McMahan, H. B., G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, Chikkerur, S., D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, J. Kubica. 2013. Ad click prediction: a view from the trenches. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 11–14, 2013, Chicago, Illinois, USA, 1222–1230.

Moe, W. W, P. S. Fader. 2002. Fast-track: Article using advance purchase orders to forecast new product sales. *Market. Sci.* **21**(3): 347–364.

Moretti, E. 2011. Social learning and peer effects in consumption: Evidence from movie sales. *Rev. Econ. Stud.* **78**(1): 356–393.

Nichols, D. R., J. J. Tsay. 1979. Security price reactions to long-range executive earnings forecasts. *J. Acc. Res.* **17**(1): 140–155.

Ord, K., R. Fildes. 2012. *Principles of Business Forecasting*. Cengage Learning, Boston, MA.

Osadchiy, N., V. Gaur, S. Seshadri. 2013. Sales forecasting with financial indicators and experts' input. *Prod. Oper. Manag.* **22**(5): 1056–1076.

Pang, B., L. Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retrieval* **2**(1–2): 1–135.

Penman, S. H. 1980. An empirical investigation of the voluntary disclosure of corporate earnings forecasts. *J. Acc. Res.* **18**(1): 132–160.

Resnik, P., et al. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res. (JAIR)* **11**: 95–130.

Rudin, C., G.-Y. Vahn. 2015. The big data newsvendor: Practical insights from machine learning. Available at SSRN 2559116.

Socher, R., B. Huval, C. D. Manning, A. Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, pp. 1201–1211.

Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *J. Roy. Stat. Soc. Ser. B (Methodol.)* **44**(1): 44–47.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. Ser. B (Methodol.)* **58**(1): 267–288.

Vahl, A. 2013. Facebook news feed updates: How marketers should respond to story bump. Available at http://www.socialmediaexaminer.com/story-bump (accessed date December 25, 2013).

Wang, H., D. Can, A. Kazemzadeh, F. Bar, S. Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. *Proceedings of the ACL 2012 System Demonstrations, Association for Computational Linguistics*, pp. 115–120.

Ye, S., G. Aydin, S. Hu. 2014. Sponsored search marketing: Dynamic pricing and advertising for an online retailer. *Management Sci*. **61**(6): 1255–1274.

Zhang, D. J., G. Allon, J. A. Van Mieghem. 2015. Does social interaction improve service quality? Field evidence from massive open online education.