DO CUSTOMERS SPEAK THEIR MINDS? USING FORUMS AND SEARCH FOR PREDICTING SALES

Completed Research Paper

Tomer Geva
Recanati Business School
Tel Aviv University
Tel Aviv 6997801
Israel
tgeva@tau.ac.il

Gal Oestreicher-Singer
Recanati Business School
Tel Aviv University
Tel Aviv 6997801
Israel
galos@tau.ac.il

Niv Efron
Google Inc. Israel
Yigal Alon 98
Tel Aviv 6789141
Israel
niv@google.com

Yair Shimshoni
Google Inc. Israel
Yigal Alon 98
Tel Aviv 6789141
Israel
shimsh@google.com

Abstract

A wide body of research uses data from social media websites to predict offline economic outcomes such as sales. However, in practice, such data are costly to collect and process. Additionally, sales forecasts based on social media data may be hampered by people's tendency to restrict the topics they publicly discuss. Recently, a new source of predictive information—search engine logs—has become available. Interestingly, the relationship between these two important data sources has not been studied. Specifically, do they contain complementary information? Or does the information conveyed by one source render the information conveyed by the other source redundant? This study uses Google's comprehensive index of internet discussion forums, in addition to Google search trend data. Predictive models based on search trend data are shown to outperform and complement forum-data-based models. Furthermore, the two sources display substantially different patterns of predictive capacity over time.

Keywords: Search trends, forums, word-of-mouth, consumers' interest, sales prediction, online data

Introduction

The availability, level of detail and scale of social media data have encouraged researchers as well as practitioners to explore means of using such data to explain and predict offline economic outcomes.

As a result, in the fields of information systems (IS) and marketing research, a dominant stream of research has emerged that focuses on abstracting data from social media websites such as blogs or internet discussion forums as a measure for the word of mouth (WOM) a product enjoys. Clearly, being visible to all, publicly broadcasted opinions and conversations may influence other consumers; it is this far reach that first made researchers expect that such publicly broadcasted opinions might be useful in predicting economic outcomes. Indeed, the number of mentions (Liu 2006), as well as the sentiment (Chintagunta et al., 2011) expressed in such publicly available data have been shown to predict offline sales.

Social media monitoring, whereby companies obtain information by sifting through data from social media websites, has become a widespread practice. In 2012 companies spend an estimated $840 million dollars on social media monitoring, nearly 1/3 of their social media marketing budget.[1] Social media monitoring is considered less expensive than traditional market research, which involves surveys or focus groups. In practice, however, it is often costly to collect and process social media data, especially when implementing more complex content processing procedures such as sentiment analysis.

Cost is not the only challenge to companies attempting to exploit online social media data to predict economic outcomes. In particular, it is not clear to what extent consumers indeed "speak their minds". Specifically, do we talk about every product we use? Do we talk about all products to the same extent or with the same level of excitement? It is possible that when consumers are aware that their opinions are visible to others, they limit the topics and volume of the opinions they express.

An alternative source of online data may offer companies a better glimpse into customers' true purchase intentions—those they pursue when their activities are not visible to all. Search engine logs, aggregating billions of individual search engine queries, have recently been made public through tools such as Google Trends. It has been argued that search data actually reflect the "true intentions" of consumers (Wu and Brynjolfsson 2009), suggesting that these data could serve as a proxy for consumers' interest in a product, which could be used to accurately predict true demand patterns. If search trend data can indeed replace WOM analysis or even simply enhance it, it has the potential to revolutionize sales predictions in terms of both cost and scope.

The two sources of data – publicly available WOM from social media websites such as internet blogs or discussion forums (hereafter referred to as "forum data"), and search data - are very different in nature. In effect, public forum data and private search are characterized by different tradeoffs between potential influence on others and the level of self-restraint they are likely to entail: Opinions posted publicly on social media sites may influence other consumers yet may not fully represent consumers' intentions, whereas privately-conducted search may better reflect actual interest in the product. This raises questions regarding the interplay between these two sources of data. Interestingly, the predictive power of the two data sources has not been compared, and the specific characteristics of the informativeness of each source have not been studied. This research seeks to close this gap.

While data from discussion forum websites and data from search engine logs (hereafter "search trend data") have separately been shown to be predictive of offline sales, usage of forum data for prediction purposes has been much more popular in marketing and IS research. Therefore, our first research question examines whether the use of search trend data can replace data taken from forums in the prediction of sales. In other words, we investigate whether predictive models using search trend data obtain similar or better accuracy than predictive models utilizing forum data. In this vein, we also examine whether predictive models based on a combination of search trend data and forum data are superior to those based on forum data alone, or whether the information conveyed in forums renders search trend information redundant. If the incorporation of search trend data into a model based on forum data improves the model's predictions, then this could facilitate building more accurate sales prediction models. On the other hand, if the information already contained in forum data renders search

[1] Business Week, http://www.businessweek.com/stories/2010-10-20/wanted-social-media-sifters

trend data redundant, then adding trend data to a forum-data-based predictive model can lead to overfitting the data and may ultimately harm out-of-sample predictive accuracy.

Our second research question explores the temporal aspects of the informativeness of each online data source. Specifically, we are interested in measuring how long an item of information from a given data source remains useful for the purpose of prediction. As forum data and search trend data are different in nature and have been reported to proxy different phenomena, it is reasonable to expect that the two sources will have different temporal patterns of informativeness. For example, one may expect forum data to be predictive for a longer period of time, as public WOM affects future consumers, whereas search usually ends after the purchase is made. On the other hand, much of the forum data is posted by consumers who have already purchased, and hence may be less accurate in predicting future sales, whereas search is often conducted by consumers during their decision process.

We also look for differences in informativeness across different types of brands. Different brands have diverse characteristics, which may influence the informativeness of search and forum data. It is reasonable to expect, for instance, that some brands are more likely to be mentioned in forums (Lovett et al., 2013), whereas others are more likely to be searched for. For example, in forums people may tend to discuss brands that are fashionable or luxurious (e.g., Porsche), as opposed to more common and less fashionable brands (e.g., Chevrolet). However, it is not clear how such differences affect the informativeness of forum data and search data for prediction. On the one hand, if a brand enjoys more WOM in forums, then forum data may be particularly successful in predicting sales of that brand. On the other hand, forum data may add noise when people discuss a brand with no purchase interest (for example, when discussing the appearance of a car in a James Bond movie). Similarly, some brands may be searched for with more purchase intent, while others may be searched for out of curiosity (following a news report, for example).

In this paper we focus on the automotive industry. Given that, for most consumers, the purchase of a car is a substantial financial expense and an information-intensive decision, we expect both WOM and search to be important. The automotive industry also provides us with the context to answer our last research question, as it includes both "luxurious" and "standard" brands. Finally, its large scale, multi-billion dollar marketing budget, as well as its importance to the economy, render automotive sales an interesting test-bed with important practical implications.[2]

The modeling methodology in this study is predictive, rather than explanatory (Shmueli and Koppius 2011). The substantial differences between predictive and explanatory methodologies, their use cases and the justifications for using each approach are thoroughly detailed by Shmueli (2010) and Shmueli and Koppius (2011). Specifically, Shmueli and Koppius (2011) state that predictive methodology is particularly useful for "assessment of the predictability of empirical phenomena". Indeed, assessment of inherent predictability of sales given different online data sources is at the core of this study's research questions. Once the predictive capabilities of the data have been assessed, the outcomes can motivate new theory generation by subsequent studies.

We find that predictive models based on search trend data can outperform forum-based predictive models. This suggests that predictive models based on publicly available search trend data could replace models based on forum data—eliminating the need to collect and process such data. Furthermore, forecasting models that incorporate both forum data and search trend data provide significantly more accurate sales predictions compared with models using forum-based data alone. We also find that the two sources display substantially different patterns of informativeness over time and that search trend data are informative even when utilizing deeper lags of data.

We also show that search trend-based models outperform the forum-based models for both "common" and "premium" brands. Nevertheless, for "common" brands, this difference is considerably larger, while for premium car brands the improvement obtained by using search trend data is more moderate. Last, we test our hypotheses using two different modeling approaches and show that predictive accuracy is strongly affected by the prediction model selection, with non-linear methods considerably outperforming linear methods.

[2] According to Nielsen, during 2008, marketing expenditures in the automotive industry were more than $2.5 billion per quarter. http://blog.nielsen.com/nielsenwire/consumer/ad-spending-in-u-s-down-11-5-percent-in-first-three-quarters-of-2009/

Related Literature

The prevalence of online platforms in which users can communicate product information to each other, such as discussion groups, forums and even product reviews on online sellers' websites, has led to an increase in publicly available WOM. This WOM is different from traditional person-to-person communication, which is often between familiar parties and limited in reach. Online public WOM has drawn much attention from both marketing and IS researchers, who have studied its effect on sales. For example, WOM in the form of posts on websites such as Yahoo! Movies has been shown to impact box office revenues (Duan et al. 2008a; Liu 2006); music blog buzz has been shown to impact music listening (Dewan and Ramprasad 2012) and sales (Dewan and Ramprasad 2009; Dhar and Chang 2009); book reviews published on a seller's own website were shown to impact the sales of reviewed books (Chevalier and Mayzlin 2006); and conversations on Usenet have been shown to affect TV ratings (Godes and Mayzlin 2004). Researchers have also studied the interplay between online WOM and critics' reviews (Chakravarty et al. 2010) and its usefulness for predicting movie revenues (Dellarocas et al. 2007); as well as the impact of internal and external WOM on sales (Gu et al. 2012). Other researchers evaluated the positive feedback of sales on WOM (Duan et al. 2008b); and the optimal response of firms to WOM (Dellarocas 2006; Chen and Xie 2008). In addition, various moderating and influential factors in relation to WOM influence on sales have been evaluated, such as the role of product and consumer characteristics (Zhu and Zhang 2010) as well as reviewer characteristics (Hu et al. 2008) and identity exposure (Forman et al. 2008).

More recently, the valence or sentiment of WOM has received increased attention. However, findings on this topic are somewhat varied. For instance, Liu (2006) and Duan et al. (2008b) have found that it is WOM volume, and not valence or user rating, that affects sales. In contrast, more recent studies such as those of Rui et al. (2012) and Chintagunta et al. (2011) report valence as an important factor in explaining sales. Rui et al. (2012) suggest that the difference between their outcomes and prior findings may have resulted from their use of an automated classifier, rather than reported user ratings, to measure valence. Chintagunta et al. (2011) attribute the difference in valence results to their improved modeling, which takes into account various complications of using a national-level data set that were not considered in previous studies.

Search engine logs have received some attention for their usefulness in explaining and predicting a variety of economic and social events. Although search is conducted privately, tools such as Google Trends have recently made search logs publicly available at the aggregate level. Choi and Varian (2009, 2011) used this type of data to demonstrate contemporaneous predictive capabilities in various fields, including sales of motor vehicle parts, initial claims for unemployment benefits, travel, consumer confidence index, and automotive sales. Wu and Brynjolfsson (2009) utilized Google search data to predict future house sales and price indices as well as home appliance sales. Vosen and Schimdt (2011) used Google search data to predict private consumption, while Ginsberg et al. (2008) used Google search query data to build an early detection system for influenza epidemics. Du and Kamakura (2012) developed a method for extracting latent dynamic factors in multiple time series and demonstrated their method by utilizing search trend data and predicting automotive sales. Seebach et al. (2011) also used Google data to predict automotive sales. Goel et al. (2010) used Yahoo!'s search engine data to predict various outcomes, including weekend box office revenues for feature films, video game sales and song ranks. They point to several factors that can affect predictions based on search data, including variability in the predictive power of search in different domains and possible difficulties in finding suitable query terms. They also discuss the need to utilize benchmark data when available. An explanation of why web search data are useful in predicting future sales is provided by Wu and Brynjolfsson (2009), who suggest that web search logs constitute "honest signals of decision-makers' intentions". That is, if buyers reveal their true intentions to purchase, future sales levels are expected to correspond to these intentions.

Data and Representation

This research uses monthly data for 20 car brands (encompassing about 90% of new car sales) sold in the US over the 4-year period between 2007 and 2010. We use three different sources of data, described below: sales, search, and forums. Note, that following common practice we use brand-level sales rather than specific car model sales (e.g., the *BMW* car brand, rather than the *528i* car model) (Choi and Varian

2009; Du and Kamakura 2012; Seebach et al. 2011).[3] See Appendix A for the list of brands and monthly sales.

Sales data

We utilize data on US unit sales of new cars and light trucks, obtained from the Automotive News website (www.autonews.com/). Automotive News provides sales data at a monthly level of aggregation. This is a well-known source for automotive sales information that has been used in various related studies such as Choi and Varian (2009) and Du and Kamakura (2012). In what follows, we use $Sales_{i,t}$ to denote the sales volume of brand *i* during month *t*.

Search data

We use Google search engine query logs. These are the same raw data that Google uses to display search engine query trends on the Google Trends website (http://www.google.com/trends/). Specifically, we collect the reported volume of monthly Google search queries for each of the car brands. We limit our data to searches originating from the US and to searches related to the automotive industry, by selecting the relevant category options in Google Trends. In what follows, we use $Search_{i,t}$ to denote the search volume of brand *i* during month *t*.

Forum data

To represent forum data we use Google's vast scan of the internet. To the best of our knowledge, this is the most comprehensive scan of forum data that has been made available for any academic research.[4] Specifically, we extracted data from all English-language forums indexed by Google's discussion forum search.[5]

Following recent literature on this topic, we extracted two aspects of forum data for each car brand: the number of times the brand was mentioned in forums ("forum mentions") and the overall sentiment (valence) of these forum mentions ("forum sentiment"). To represent brand *i*'s forum mentions in month *t* (denoted $forum\_mentions_{i,t}$), we used the number of new forum posts mentioning brand *i* during month *t*. To represent forum sentiment for brand *i* in month *t* (denoted $forum\_sentiment_{i,t}$), we used the ratio between the sums of "positive mentions" and "negative mentions" for brand *i* in month *t*. To label forum postings as "positive" or "negative", we used a dictionary-based sentiment analysis approach that is popular in the literature (see, for instance, Berger and Milkman 2012). Specifically, we utilized the extended positive and negative word dictionaries from the well known Harvard IV-4 psychological dictionary[6] and summed the number of new forum posts mentioning "positive words" and posts mentioning "negative words", alongside brand *i* during month *t*. The advantages of using this dictionary approach are its generalizability and reproducibility over proprietary or "black box" types of sentiment analysis solutions.[7]

Key words

In order to collect search data and forum data, it was necessary to specify key words that could be used to identify searches or forum mentions associated with each brand. This section elaborates on the design decisions we made regarding key word selection.

---

[3] The motivation for this practice is twofold. First, brand-level data are much more abundant than car model data. Second, key word identification is considerably more accurate at the brand level than at the car model level (see discussion below on key word selection).

[4] Related studies that used data from forums or blogs utilized data either from a specific website, from a domain-specific search engine for forums, or from a simple hit count from Google web searches.

[5] Results for Google's discussion forum search are currently presented when selecting the *"more"* option under Google's search box, and subsequently selecting *"Discussions"*.

[6] http://www.wjh.harvard.edu/~inquirer/

[7] Our findings reported in later stages show that despite its simplicity, this type of sentiment representation considerably improves predictive accuracy (see figures 1,2).

Let *K* denote a set of key words and *B* denote a given brand. We use the term "accuracy" to denote the ratio between the number of search queries (or forum posts) that specify (any word in) *K* and that actually relate to brand *B*, and the total number of search queries (or forum posts) specifying any word in *K*. We use the term "coverage" to denote the ratio between the number of searches (forum posts) using any word in *K*, and the hypothetical, full number of searches (forum posts) referring to brand *B* (using any key word).

In general, when selecting a set of key words to identify a given brand, there is a tradeoff between accuracy and coverage. Clearly, inclusion of a larger number of key words can increase coverage, but it may introduce noise and decrease accuracy. On the other hand, if we choose a limited set of terms for a given car brand and obtain high accuracy, we may not fully capture the brand's "online presence". For example, if one wishes to capture search queries pertaining to the Chevrolet car brand, one will most likely use the term "Chevrolet". Next, one can increase coverage by adding car model names such as "Malibu" (capturing additional searches for "Chevrolet Malibu") or "Spark" (capturing additional searches for "Chevrolet Spark"). However, adding search terms such as "Malibu" or "Spark" may also introduce a large number of irrelevant queries, e.g., queries relating to the town of Malibu, California. Note that there is no point in adding the more specific, two-word term "Chevrolet Malibu" (or "Chevrolet Spark"), to the set of key words, as a search using this term is a subset of the searches using the key word "Chevrolet".

To the best of our knowledge, the literature does not offer a methodology for optimal selection of key words with the aim of achieving best predictive performance using both search and forum data, for different domains. Therefore, in this study we utilized brand-level key words (e.g., "Chevrolet" for the Chevrolet brand), similarly to Seebach et al. (2011).[8] (See Appendix A for a detailed list of the key words we used.)

While brand-level key words can naturally provide high accuracy in capturing brand-related search queries, we also adapted our modeling procedures to mitigate coverage concerns. First, as mentioned above, we expected that brand-level key words (e.g., "Chevrolet") would be considerably more commonplace than model-level keywords (e.g., Spark), for most car brands. Thus, the initial level of coverage was already expected to be relatively high. Second, we excluded from our data a few "atypical" brands with relatively high volume of searches at the car model level, compared to the quantity of searches at the car brand level. Specifically, we excluded car brands for which the search volume for the brand's most searched car model name was more than 0.3 times the search volume for the car brand name. This led to the exclusion of Volkswagen and Chrysler brands.[9]

Third, we note that when a separate prediction model is constructed for each car brand, as long as the key-word coverage is sufficiently representative of the brand, to the point that the ratio between the volume of searches (or forum mentions) captured by the brand-level key word and the hypothetical, unknown, "full volume of relevant searches" (or forum mentions) remains stable over time—there is actually no need to fully capture the hypothetical, unknown, "full coverage". Even simple models such as linear regression can overcome this problem by simply adjusting the coefficient values. As we are generally interested in predictive capability, rather than specific coefficient values, scaled coefficient values are not a concern.

Finally, to control for different levels of baseline coverage across multiple brands, we converted the dependent and independent variables into per-brand, normalized variables and utilized the distance, in term of standard deviations, from the brand's mean, rather than the original values.

[8] Seebach et al., reported that using brand level key words produced best results in a similar scenario of automotive sales prediction using search trends data.

[9] In selecting the 0.3 threshold we limit the number of excluded brands and exclude only two brands whose brand names are clearly not commonly used.

Modeling

Setup

Our dependent variable is $Sales_{i,t}$ - automotive sales for brand *i* in month *t*. To make a prediction for each brand's sales in month *t*, we use data that are available at month *t-1*.

To predict a brand's sales in month *t*, we used forum data and search trend data, as elaborated above. In addition, we incorporated important benchmark information into our model. This included data on previous months' sales, and seasonality—sales in the same month, in the previous year. Usage of such data to represent seasonality is common in autoregressive models (see, for example, Choi and Varian, 2009). Modeling was carried out on a monthly basis, beginning with one lag of historical data (month *t*-1) and gradually incorporating additional lags (up to five lags of data, months: *t-1*, …, *t-5*). In what follows we use the notation "lag *j*" to denote data from month *t-j*.

Following common practice in predictive research, we measured the model's performance "out-of-sample", i.e., we used one set of data to train the model and another set to measure its performance. Specifically, in line with other studies in this field (e.g., Dellarocas et al. 2007), we used the 10-fold cross-validation performance evaluation method (see also Efron and Tibshirani 1993). This method is effectively capable of measuring model performance over an entire set of observations as a form of an external "validation set", that is, without the need to "lose" observations for training purposes as in the case of a simple "training" and out-of-sample "validation" data split. As a result, the cross-validation method is well suited for this study, in which the number of observations for the aggregate-level variables (i.e., monthly data for sales, search trends or forums) is not large, essentially including a total of 720 observations[10] (36 observations per brand).

Table 1. Data Included in Each Model

	Basic Model - Previous Sales	Benchmark Model - Previous sales and Seasonality	Forum-Based Model	Extended Forum-Based Model	Search Trends-Based Model	Combined Search and Forum-Based Model
$Sales_{i,t-1}$,…, $Sales_{i,t-n}$	√	√	√	√	√	√
$Sales_{i,t-12}$		√	√	√	√	√
$Forum\_mentions_{i,t-1}$,…, $Forum\_mentions_{i,t-n}$			√	√		√
$Forum\_sentiment_{i,t-1}$,…, $Forum\_sentiment_{i,t-n}$				√		√
$Search_{i,t-1}$,…, $Search_{i,t-n}$					√	√

We define a "*basic model"* as a model that relies solely on previous sales data. We define a *"benchmark model*" as a model that utilizes both previous sales and seasonality ($Sales_{i,t-12}$). In order to gauge the relative informativeness of forum-based data and search-trend data in comparison to benchmark data, as well as the benefit of augmenting forum-based data with search trend data, we define several additional models incorporating different sets of data, as follows: The "*forum-based model*" utilizes the benchmark

[10] Although we have 48 months of sales data per brand we "lose" 12 months' worth of data when accounting for seasonality by using $Sales_{i,t-12}$.

model data in addition to forum mentions; the "*extended forum-based model*" adds forum sentiment data to the forum-based model; the "*search trends-based model"* utilizes both benchmark information and search trend data; and the "*combined search and forum-based model"* utilizes all the sets of information mentioned above. Table 1 summarizes the different sets of data utilized in each prediction model.

Forecasting Algorithms

To better examine how the different types of data affect prediction accuracy, we utilized two well-known algorithms to generate predictions using the data models detailed above.

The first algorithm is the popular linear regression (LR) algorithm. This method has been used in the vast majority of related studies seeking to predict economic outcomes on the basis of either forum data or search trend data.[11]

The second algorithm is the back-propagation neural network (NN) algorithm. This is a non-linear method that is estimated by the backprop algorithm (Werbos 1974). One of the strongest properties of the NN algorithm is that it inherently accounts for non-linear relationships and complex interactions between variables, making it especially applicable for modeling complex non-linear relations (Bishop 1995).

To implement the NN algorithm we used the "nnet" package in R software. This implementation involves one layer of hidden nodes, and the minimization of a sum of square errors criterion. In specifying the network architecture, one must choose the number of nodes in the hidden layer. While the literature does not offer clear rules about the optimal complexity of the network in terms of hidden nodes, it proposes general guidelines (Zhang et al. 1998). For instance, the number of hidden nodes should be proportional to the number of inputs. Specific architectures have been proposed, e.g., number of inputs multiplied by 0.5 or 1. We carried out our analysis using both levels of complexity, and both yielded similar findings. The results reported here are based on a network architecture in which the number of hidden nodes is equal to the number of inputs multiplied by 0.5.[12] Finally, while NNs are well-known for their ability to "learn" complex relations, in practice NN results may sometimes produce unstable predictions, overfit the data or converge to a local optimum. These problems can be even more pronounced in cases of relatively small datasets such as ours. As a safeguard against these problems, our specific implementation utilizes the median prediction of an ensemble of 100 NNs, each using a different random seed.

The purpose of evaluating the performance of models relying on two different algorithms is to discern whether a given data source enables better prediction regardless of the type of algorithm (linear or non-linear) or whether a certain type of algorithm is better suited to exploit the information from a given data source. While it is possible to use many other algorithms, our choice reflects two standard, linear and non-linear, model types that are commonly used in econometric and machine learning modeling.

Performance Measure

We used the mean absolute percentile error (MAPE) as our performance criterion. We made this choice for two main reasons: First, MAPE controls for volume differences across brands. For example, using MAPE, a 10% error in prediction for a large manufacturer is treated similarly to a 10% error in prediction for a small manufacturer. Second, MAPE is indifferent to the direction of the error (either over-estimation or underestimation). This is appropriate for our approach, as we are interested in evaluating the predictive capacity of the data, i.e., the extent to which reliance on the data can reduce prediction error, regardless of a brand's sales volume or the direction of the error.

While we present our results using MAPE, for robustness, we repeated the analysis using mean square error (MSE) criteria and reached similar findings in term of the relative performance of the different models.

Normalization

We transformed each of the variable representations into normalized values (for each brand). There are

[11] Using LR, coefficients for a linear combination of predictors are determined when minimizing a sum of squared errors cost function. For additional details see, for instance, (Greene, 2011).
[12] If the number of inputs is an odd number, we round up the number of hidden nodes.

two motivations for normalization on the brand level. First, this controls for differences in sales volume across different brands. Second, as discussed in the "Data and Representation" section, normalization is a key component in our key-word handling methodology.

We note that while our models use normalized variables, in order to provide interpretable results, we calculate the MAPE according to the actual, "de-normalized" numbers (therefore, in our results, a MAPE of X% would mean a difference of X% between actual and predicted sales). We further note that in deriving the results reported below, we used normalization/de-normalization procedures in which the entire set of observations were used in order to calculate the sample mean and standard deviation. For robustness, however, we carried out our analysis once again—this time, for each iteration of the 10-fold cross validation, we used as normalization factors the sample mean and standard deviation only of that iteration's training set. The results obtained with this procedure were similar to the main results.

Results

Figure 1 displays the results obtained with LR using different data representations and different numbers of monthly lags. Figure 2 displays the results with NN using different data representations and different numbers of monthly lags. Correspondingly, Tables 2 and 3 present significance testing for LR and NN models, respectively, using a bootstrap confidence interval for the performance differences between models utilizing different sets of data. Table 4 presents significance testing for the "best" results in terms of algorithm and number of lags for each model type.

Our first core finding is that models based on search trend data significantly outperform models based on forum data (the green vs. the blue lines in Figure 1 and Figure 2). In particular, search-trend-data based models outperform the more elaborate forum-data-based models, i.e., those that include sentiment analysis (the purple vs. the blue lines on Figure 1 and Figure 2). We also find that combining forum data and search trend data further improves the prediction accuracy (the orange lines in Figure 1 and Figure 2). This finding suggests that the two types of data sources hold complementary information.[13]

An additional interesting finding that arises from our analysis is that is that the non-linear NN models considerably outperform the corresponding LR models for all data sources and lags. This suggests that complex relations may exist within such data sets, and that linear models—which are prevalent in many related studies that examine either forum data or search trend data—may be underutilizing the available information. Due to the clear superiority of the NN models, in the rest of our discussion we focus on the NN model results.[14]

Our second core finding relates to the temporal informativeness of the different sources of data. As evident from Figure 2 and Table 3, forum data and search trend data exhibit substantially different patterns in terms of informativeness over time. In particular, we find that the predictive power of forum data (either forum mentions, or both forum mentions and sentiment), is concentrated in more recent lags (lags 1,2, i.e., data from up to two months preceding the prediction). Incorporating additional lags of forum data (lags 3,4) has only a marginal contribution to predictive accuracy, and incorporating the most remote lag (lag 5) actually reduces predictive accuracy, most likely due to the introduction of "noise" into the model. Interestingly, this is not the case when using search trend data. Adding deeper lags of search trend data consistently increases predictive accuracy, up to lag 4, and incorporation of lag 5 does not degrade the results. In other words, the information encompassed in the more distant lags of search trend data not only improves the predictive accuracy of the search trend-based models but also enables these models to significantly outperform forum-data-based models (see Figure 2, and Table 4).

[13] Results for LR are consistent with the results for NN when comparing the relative performance of the different data sources and/or different lags. However, the differences in predictive performance are attenuated in the case of LR.

[14] We also note that the NN algorithm is known for its capabilities in approximating complex functions based on relations within data. Therefore, this algorithm is also more suitable to accurately model "special" cases in which an increase in search volume is actually due to negative circumstances (e.g., a certain car model recall). In this case, the NN model can potentially "learn" from the sentiment data about the emergence of negative opinions concerning the product, and take into account the interaction between an increased search volume and lower brand related opinion.

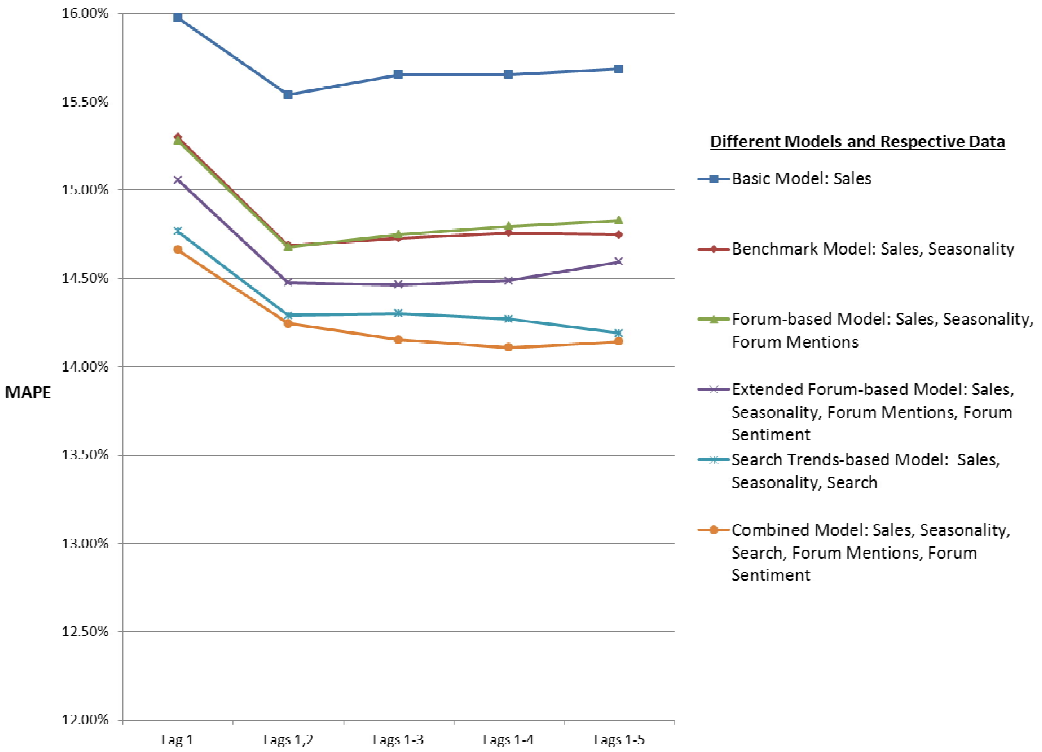

Figure 1. Prediction Results – LR

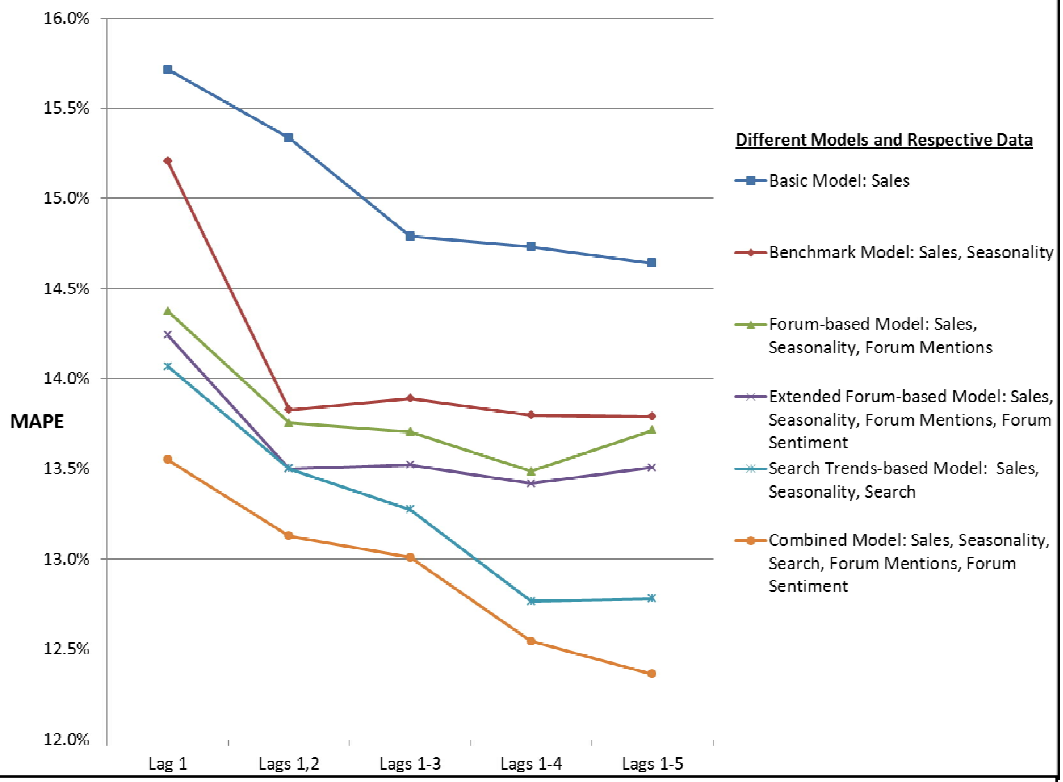

Figure 2. Prediction Results – NN

Table 2. One-Sided Confidence Intervals for the Difference in MAPE Values Using an LR Algorithm with Different Data Sources

Model A	**Model B**	*Lag 1*	*Lags 1,2*	*Lags 1-3*	*Lags 1-4*	*Lags 1-5*
Forums-Based Model	Benchmark Model					
Extended Forum-Based Model	Benchmark Model	**	**	**	**	
Search Trends-Based Model	Benchmark Model	***	***	***	***	***
Combined Model	Benchmark Model	***	***	***	***	***
Search Trends-Based Model	Forums-Based Model	***	***	***	***	***
Search Trends-Based Model	Extended Forum-Based Model	*				**
Combined Model	Forums-Based Model	***	***	***	***	***
Combined Model	Extended Forum-Based Model	***	**	**	**	***
Combined Model	Search Trends - Based Model					

diff = MAPE(model A) - MAPE(model B)
* 0.9 upper confidence bound for *diff* is negative;
** 0.95 upper confidence bound for *diff* is negative
*** 0.99 upper confidence bound for *diff* is negative
Confidence intervals for *diff* were calculated using 2000 iterations of the BCA Bootstrapping confidence interval calculation method implemented in the R software

Table 3. One-Sided Confidence Intervals for the Difference in MAPE Values Using an NN Algorithm with Different Data Sources

Model A	**Model B**	*Lag 1*	*Lags 1,2*	*Lags 1-3*	*Lags 1-4*	*Lags 1-5*
Forums-Based Model	Benchmark Model	***			*	
Extended Forum-Based Model	Benchmark Model	***	**	**	*	
Search Trends-Based Model	Benchmark Model	***	**	***	***	***
Combined Model	Benchmark Model	***	***	***	***	***
Search Trends-Based Model	Forums-Based Model			**	***	***
Search Trends-Based Model	Extended Forum-Based Model				**	***
Combined Model	Forums-Based Model	***	***	***	***	***
Combined Model	Extended Forum-Based Model	***	**	***	***	***
Combined Model	Search Trends-Based Model	***	**			*

diff = MAPE(model A) - MAPE(model B)
* 0.9 upper confidence bound for *diff* is negative
** 0.95 upper confidence bound for *diff* is negative
*** 0.99 upper confidence bound for *diff* is negative
Confidence intervals for *diff* were calculated using 2000 iterations of the BCA Bootstrapping confidence interval calculation method implemented in the R software

Table 4. One-Sided Confidence Intervals for the Difference in MAPE Values Using the Best Setup (Forecasting Algorithm and Lags) for Each Model

Model A		**Model B**		MAPE(model A) - MAPE(model B)
Model	**Best Setup**	**Model**	**Best Setup**	
Forum-Based Model	NN, lags 1-4	Benchmark Model	NN, lags 1-5	*
Extended Forum-Based Model	NN, lags 1-4	Benchmark Model	NN, lags 1-5	*
Search Trends-Based Model	NN, lags 1-5	Benchmark Model	NN, lags 1-5	***
Combined Model	NN, lags 1-5	Benchmark Model	NN, lags 1-5	***
Search Trends-Based Model	NN, lags 1-5	Forum-Based Model	NN, lags 1-4	***
Search Trends-Based Model	NN, lags 1-5	Extended Forum-Based Model	NN, lags 1-4	**
Combined Model	NN, lags 1-5	Forum-Based Model	NN, lags 1-4	***
Combined Model	NN, lags 1-5	Extended Forum-Based Model	NN, lags 1-4	***
Combined Model	NN, lags 1-5	Search Trends-Based Model	NN, lags 1-5	*

diff = MAPE(model A) - MAPE(model B)
* 0.9 upper confidence bound for *diff* is negative
** 0.95 upper confidence bound for *diff* is negative
*** 0.99 upper confidence bound for *diff* is negative
Confidence intervals for *diff* were calculated using 2000 iterations of the BCA Bootstrapping confidence interval calculation method implemented in the R software.

To gauge additional aspects of the informativeness of the different data sources, we explored the extent to which the informativeness of forum data and search trend data is dependent on the characteristics of the brand for which predictions are being made[15]. Specifically, we examined whether each data source produces more accurate predictions for "high end" car brands, i.e., brands with higher pricing or higher perceived quality, or for "lower end" or "standard" car brands. The intuition regarding the relationship between data informativeness and brand type is not straightforward: We may expect, for example, that low end, less fashionable cars will be "under-represented" in public forums, as it may not be considered "cool" to discuss them. On the other hand, if people tend to discuss prestigious and fashionable brands on forums, this may introduce noise into the prediction models, as people may discuss exciting cars even if they have no intention of buying them.[16] Similarly, in terms of search data, we might expect high end, or prestigious cars to be associated with a substantial quantity of searches due to mere curiosity rather than specific purchasing intentions (e.g., people looking for a photo of a new BMW sports car, even if they cannot afford to buy it). The ability of different sources of data to predict the sales of specific brands may also depend on factors such as the demographic characteristics of the potential buyers—e.g., the amount of time potential buyers have to perform searches or to post to forums, their ability to access the internet, tendency to write in forums, etc.

Figure 3 presents prediction results for the models based on different sets of information, splitting the car brands into two subsets: car brands for which the list price for the least expensive car model of each brand was more than $20,000, referred to as "premium", and the rest of the car brands, referred to as

[15] The analyses reported here were conducted using the best-performing models, in terms of the number of lags, while using the different sets of information: For the full forum-based model—i.e., the model using data on previous sales, seasonality, forum mentions and sentiment—we used the model with 4 lags of data. For the other models detailed in this section we used 5 lags of data.
[16] We find that the number of forum mentions per unit sales for high end brands is 2.1 times higher than the respective number for low end brands.

"standard".[17] The figure clearly shows that predictions for premium car brands are generally more accurate than for standard brands across all prediction models. A possible explanation for this is the difference in demographics, which could affect factors such as computer/internet access for potential buyers or tendency to perform searches or to participate in forums.

For both standard and premium brands, search trend models outperform the forum-based as well as the extended forum-based models. Interestingly, for standard brands, this difference is considerably larger, while for premium car brands the improvement obtained by using search trend data is more moderate. This seems to support the conjecture that for less expensive cars, forum data may be less informative regarding true purchase intentions, compared with data on searches that are conducted in private.

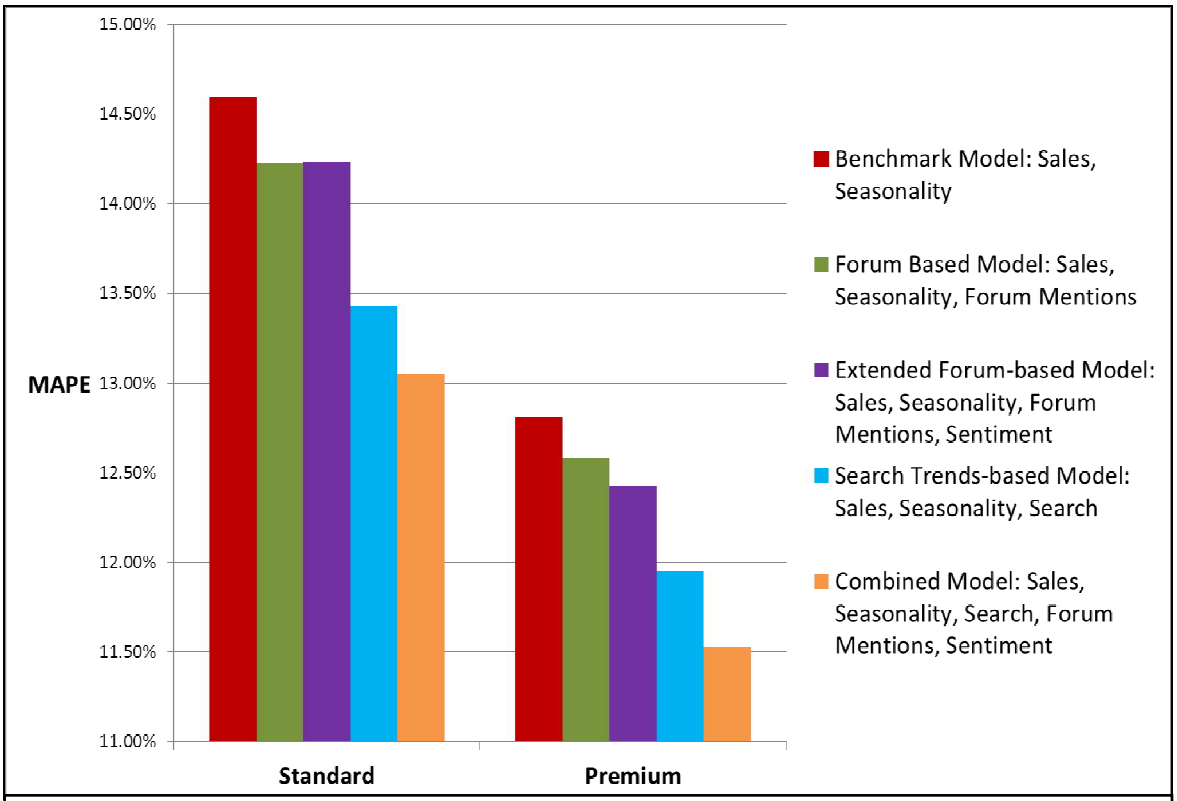

Figure 3. Prediction Results for "Standard" and "Premium" Brands

Conclusions

In this work, we used a predictive methodology to analyze the informativeness of forum data and search trend data for the purpose of predicting future sales. We provide first evidence that search trend data facilitate more accurate sales predictions compared with well-accepted forum-based data. Furthermore, the difference in performance can be attributed to our finding that search trend data and forum data display substantially different patterns of informativeness over time. Specifically, while forum data are informative only for recent lags of data, search trend data are informative even when using deeper lags of data. Additionally, we find that the difference in predictive accuracy, between forum-based and search-

[17] Prices for 2007. Source: Automotive News website.

based models is considerably larger for "standard" brands.

We also provide first evidence that augmenting forum-based models with search trend data significantly improves predictive accuracy. This evidence indicates that customer interest—as proxied by search trends—provides external, and non-overlapping information, to forum data, for the purpose of sales forecasting. This finding provides reassurance that companies that have already invested in collecting forum-based data for modeling purposes have not "wasted" their money—while further suggesting that these companies can considerably improve forecasting accuracy with a relatively small additional investment in collecting search trend data. Another finding is that non-linear prediction methods considerably outperform linear methods when using either source of data, suggesting that complex relations exist within these types of data.

Our finding that search trend data continue to be informative even after five months may be due to the fact that automotive purchase decisions are substantial economic decisions that involve an extensive information-gathering process. In that sense, our conclusions in the context of car purchase decisions may be generalizable to a wide array of substantial decisions such as housing purchases, financial investments and travel planning. However, our findings may be less applicable to other domains, such as entertainment consumption, where decisions are made more lightheartedly.

In terms of business implications, more accurate sales prediction models using either search trend data or both search trend data and forum data can, in turn, drive better decision making in various domains such as marketing expenditure, competitive analysis, inventory management and supply chain optimization. For the specific case of automotive sales, these decisions involve the allocation of extremely large funds; therefore, even small improvements would have a considerable effect.

As one of the goals of predictive research is to stimulate theory development, we are currently working on an explanatory type of research looking at the interplay between sales, search and forum data. In parallel we are working on devising improved prediction models for car sales. Additional directions for future research include evaluating the informativeness of search trend data and forum data, using additional data in the domain presented in this study or in other domains, such as in the entertainment industry.

Acknowledgments

Tomer Geva has been a post-doctoral research scientist at Google Inc. Israel, during most stages of this research.

Appendix A – List of Brands

Table A1. List of Brands

Brand	Average Monthly Unit Sales	Key Word(s)	Grouping by Standard/Premium
Acura	10K-20K	acura	Premium
Audi	5K-10K	audi	Premium
Bmw	10K-20K	bmw	Premium
Buick	10K-20K	buick	Premium
Cadillac	10K-20K	cadillac	Premium
Chevrolet	100K-200K	chevrolet, chevy	Standard
Dodge	20K-50K	dodge	Standard
Ford	100K-200K	ford	Standard
Honda	50K-100K	honda	Standard
Hyundai	20K-50K	hyundai	Standard
Infiniti	5K-10K	infiniti	Premium
Jeep	20K-50K	jeep	Standard
Kia	20K-50K	kia	Standard
Lexus	10K-20K	lexus	Premium
Lincoln	5K-10K	lincoln	Premium
Mazda	10K-20K	mazda	Standard
Mercedes Benz	10K-20K	mercedes	Premium
Nissan	50K-100K	nissan	Standard
Subaru	10K-20K	subaru	Standard
Toyota	100K-200K	toyota	Standard

"Average monthly unit sales" relates to monthly unit sales in the US during the time period used for modeling (2008-2010). List of brands includes all brands with average sales above 5,000 cars per month. We excluded from this list the following brands due to data issues: Volkswagen, GMC, Chrysler. (See also Data and Representation section).

Appendix B – Search and Forum Mentions Volume

Table B1. Relative Search and Forum Mention Volume

Brand	Search Volume	Forum Mentions
Acura	16	19
Audi	25	48
Bmw	44	83
Buick	9	11
Cadillac	13	11
Chevrolet	75	69
Dodge	46	46
Ford	100	100
Honda	93	94
Hyundai	19	16
Infiniti	9	8
Jeep	33	30
Kia	14	9
Lexus	20	18
Lincoln	7	6
Mazda	21	29
Mercedes Benz	25	40
Nissan	49	47
Subaru	18	23
Toyota	80	66

"Search Volume" relates to the total Google search volume of the brand specific key word(s). "Forum Mentions" relate to the number of forums mentions which include the brand key words. The data pertains to the period 2008-2010. To preserve data confidentiality, data were scaled from 0-100 in relation to search/forum mentions of the brand with the highest volume.

References

Berger, J., and Milkman, K. L. 2012. "What Makes Online Content Viral?," *Journal of Marketing Research* (49:2), pp. 192–205.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*, Oxford, UK: Clarendon Press.

Chen, Y., and Xie, J. 2008. "Online Consumer Reviews: A New Element of Marketing Communications Mix," *Management Science* (54:3), pp. 477–491.

Chevalier, J., and Mayzlin, D. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research* (43:3), 345-354.

Chintagunta, P. K., Gopinath, S., and Venkataraman, S. 2011. "The Effect of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets," *Marketing Science* (29:5), pp. 944-957.

Chakravarty, A., Yong, L., and Mazumdar, T. 2010. "The Differential Effects of Online Word-of-Mouth and Critics' Reviews on Pre-Release Movie Evaluation," *Journal of Interactive Marketing* (24: 3), pp. 185-197.

Choi, H., and Varian, H. 2009. "Predicting the Present with Google Trends," working paper.

Choi, H., and Varian, H. 2011. "Predicting the Present with Google Trends," working paper.
Dellarocas, C. 2006. "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms," *Management Science* (52:10), pp. 1577–1593.
Dellarocas, C., Awad, N., and Zhang, X. 2007. "Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures," *Journal of Interactive Marketing* (21:4), pp. 23–45.
Dewan, S., and Ramprasad, J. 2009. "Chicken and Egg? Interplay between Music Blog Buzz and Album Sales," in *PACIS 2009 Proceedings*.
Dewan, S., and Ramprasad, J. 2012. "Music Blogging, Online Sampling, and the Long Tail," *Information Systems Research* (23:3), pp. 1056–1067.
Dhar, V., and Chang, E. A. 2009. "Does Chatter Matter? The Impact of User-Generated Content on Music Sales," *Journal of Interactive Marketing* (23:4), pp. 300-307.
Du, R. Y., and Kamakura, W. A. 2012. "Quantitative Trendspotting," *Journal of Marketing Research* (49: 4), pp. 514-536.
Duan, W., Gu, B., and Whinston, A. B. 2008a. "The Dynamics of Online Word-of-Mouth and Product Sales—An Empirical Investigation of the Movie Industry," *Journal of Retailing* (84:2), pp. 233–242.
Duan, W., Gu, B., and Whinston, A. B. 2008b. "Online Reviews Matter? An Empirical Investigation of Panel Data," *Decision Support Systems* (45:4), pp. 1007–1016.
Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman and Hall.
Forman, C., Ghose, A., and Wiesenfeld, B. 2008. "Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets," *Information Systems Research* (19:3), pp. 291–313.
Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. 2008. "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature* (457:7232), pp. 1012-1014.
Godes, D., and Mayzlin, D. 2004. "Using Online Conversations to Study Word-of-Mouth Communication," Marketing Science, (23:4), pp. 545-560.
Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J. 2010. "Predicting Consumer Behavior with Web Search," *Proceedings of the National Academy of Sciences* (107:41), pp. 17486–17490.
Greene, W.H. 2011. *Econometric Analysis,* 7th Edition, Prentice Hall.
Gu, B., Park, J., and Konana, P. C. 2012. "The Impact of External Word-Of-Mouth Sources on Retailer Sales for High Involvement Products," *Information Systems Research* (23:1), pp. 182-196.
Hu, N., Liu, L., and Zhang, J. 2008. "Do Online Reviews Affect Product Sales? The Role of Reviewer Characteristics and Temporal Effects," Information Technology Management (9:3), pp. 201–214.
Liu, Y. 2006. "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," Journal of Marketing (70:3), pp. 74–89.
Lovett, M., Peres, R., and Shachar, R. 2013. "On Brands and Word-of-Mouth," forthcoming J*ournal of Marketing Research*.
Rui, H., Liu, T., and Whinston, A. 2012. "Whose and What Chatter Matters? The Impact of Tweets on Movie Sales," working paper.
Seebach, C., Pahlke, I., and Beck, R. 2011. "Tracking the Digital Footprints of Customers: How Firms Can Improve Their Sensing Abilities to Achieve Business Agility," in *ECIS 2011 Proceedings*.
Shmueli, G. 2010. "To Explain or to Predict?," *Statistical Science* (25:3), pp. 289–310.
Shmueli, G., and Koppius, O. 2011. "Predictive Analytics in Information Systems Research," *MIS Quarterly* (35:3), pp. 553-572.
Vosen, S., and Schmidt, T. 2011. "Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends," *Journal of Forecasting* (30:6), pp. 565-578.
Werbos, P. J. 1974. "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences," PhD dissertation, Harvard University.
Wu, L., and Brynjolfsson, E. 2009. "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities," in *Proceedings of the 2009 International Conference on Information Systems*.
Zhang, G., Patuwo, B. E., and Hu, M. Y. 1998. "Forecasting with Artificial Neural Networks: The State of the Art," *International Journal of Forecasting* (14:1), pp. 35–62.
Zhu, F., and Zhang, X. M. 2010. "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics," *Journal of Marketing* (74:2), pp. 113-148.