

USING FORUMS AND SEARCH FOR SALES
PREDICTION OF HIGH-INVOLVEMENT
PRODUCTS

by

T. Geva*
G. Oestreicher-Singer**
N. Efron***
Y. Shimshoni****

Working Paper No 3/2015

June 2015

Research no.: 05050100

* The Leon Recanati Graduate School of Business Administration, Tel Aviv University, Ramat Aviv, Tel Aviv, 69978, Israel.

** The Leon Recanati Graduate School of Business Administration, Tel Aviv University, Ramat Aviv, Tel Aviv, 69978, Israel. Email: galos@tau.ac.il

*** Google Inc.

**** Google Inc.

This paper was partially financed by the Henry Crown Institute of Business Research in Israel.

The Institute's working papers are intended for preliminary circulation of tentative research results. Comments are welcome and should be addressed directly to the authors.

The opinions and conclusions of the authors of this study do not necessarily state or reflect those of The Faculty of Management, Tel Aviv University, or the Henry Crown Institute of Business Research in Israel.

USING FORUMS AND SEARCH FOR SALES PREDICTION OF HIGH-INVOLVEMENT PRODUCTS

Tomer Geva¹, Gal Oestreicher-Singer¹, Niv Efron², Yair Shimshoni²

A large body of research uses data from social media websites to predict offline economic outcomes such as sales. However, recent research also points out that such data may be subject to various limitations and biases that may hurt predictive accuracy. At the same time, a growing body of research shows that a new source of online information—search engine logs—has the potential to predict offline outcomes. We study the relationship between these two important data sources in the context of sales predictions. Focusing on the automotive industry—a classic example of a domain of high-involvement products—we use Google’s comprehensive index of internet discussion forums, in addition to Google search trend data. We find that adding search trend data to models based on the more commonly-used social media data significantly improves predictive accuracy. We also find that predictive models based on inexpensive search trend data provide predictive accuracy that is comparable, at least, to that of social media-based predictive models. Last, we show that the improvement in accuracy is considerably larger for “value” car brands, while for “premium” car brands the improvement obtained is more moderate.

Keywords: search trends, forums, social media, word-of-mouth, consumers' interest, sales prediction, online data, high involvement products

¹ Recanati Business School, Tel-Aviv University.

² Google Inc.,

Tomer Geva and Gal Oestreicher-Singer acknowledges financial support from a Marketing Science Institute research grant and a grant from the Henry Crown Institute of Business Research. Tomer Geva acknowledges financial support by the Israel Science Foundation (grant number 1443/14) and a post-doctoral research scientist fellowship at Google Inc (Tel Aviv Office).

USING FORUMS AND SEARCH FOR SALES PREDICTION OF HIGH- INVOLVEMENT PRODUCTS

INTRODUCTION

The availability, scale, and richness of detail 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, 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. Previous work in this stream has repeatedly provided evidence that the number of mentions that a product receives in social media, as well as the sentiment expressed in such data, can predict offline outcomes. Social media monitoring, whereby companies obtain information by sifting through data from social media websites, has also become a widespread practice in industry. In 2012 companies spent 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.²

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

² For example, simple monitoring and measurement tools (such as the Oracle solution on <http://www.oracle.com/us/solutions/social/overview/index.html>) cost a few thousand dollars per year, and more elaborate tools offered by marketing companies may cost tens of thousands of dollars (see, for instance, the Nielsen solution on <http://www.nielsensocial.com/>)

Cost is not the only challenge to companies attempting to exploit online social media data to predict economic outcomes. In fact, industry reports suggest that very few consumers actually contribute actively to social media discussions: Reports show that, on typical social media websites, only about 10% of website users are active participants, and that most of the content is contributed by only 1% of users.³ Thus, most users remain quiet observers or "lurkers", a phenomenon known as "participation inequality". While it is not uncommon to use small sample sizes (e.g., in surveys) to capture consumers' intentions for predictive purposes, such samples are typically selected with care, with the aim of ensuring representation of the opinions of the general consumer base. However, recent work has pointed out that social media data has significant biases and limitations, including: insufficient and skewed representativeness of the consumer population; biased representation of the true underlying product/brand evaluation; biases due to differences across brand and products as well as temporal variations; and potential intentional manipulation. Moreover, many of these biases have complex dynamics which are difficult to mitigate using modeling procedures. (See literature review for a more detailed discussion of those social media limitations for predictive purposes).

Interestingly, a common factor that underlies, or contributes to, many of the reported biases associated with social media data is the visibility of social media posts. Previous research has shown that the fact that social media mentions are visible to others affects users' willingness to participate in online discussions, causes users to express views that differ from their true opinions, and can also lead firms and even customers to attempt to manipulate online content.

A common practice in the context of predictive research is to mitigate flaws in imperfect

³ See <http://www.nngroup.com/articles/participation-inequality/> and blog.elatable.com/2006/02/creators-synthesizers-and-consumers.html

data by enriching the data with additional, meaningful, information. Nevertheless, it is important that the additional data will not be subject to the specific biases and flaws that the researcher is trying to alleviate. Herein, we focus on search trend data as an online data source that can potentially enrich social media data. Search engine logs, aggregating billions of individual search engine queries, have been made publicly available through tools such as Google Trends. It has been argued that search data actually reflect the “true intentions” of consumers (Wu and Brynjolfsson 2009) and can serve as a proxy for consumers’ interest in a product (Hu et al. 2014). Most importantly for our context, search trend data are very different in nature from social media data. Unlike social media postings, search is done in private. As consumers search in private, they have no self-presentation concerns, have no (or very limited) awareness of other searchers' activity, and no potential to influence sales. Additionally, search is conducted by a much larger sample of the population compared with active participation in social media discussions. While notably search trend data cannot mitigate all the known biases and limitations of social media, its private nature and widespread use render this data as a promising candidate to be added to social media-based prediction models.

Interestingly, the interplay between these two online data sources and its relationship to sales prediction has received very little research attention thus far. This raises an important question that lies at the core of the current paper: Can search trend data successfully augment social media data for sales prediction? Specifically, **our first research question** examines whether predictive models based on a combination of search trend data and social media data are superior to those based on social media data alone. If the incorporation of search trend data into a model based on social media data improves the model’s predictions, this could yield more accurate sales prediction models. Nevertheless, the risk of adding a second source of data is that

if the additional data does not contain, additional, valuable information, it can potentially lead to over-fitting and may ultimately harm out-of-sample predictive accuracy. Therefore, superiority of a model combining the two data sources would show that the two data sources contain non-overlapping, useful, information.

Although search trend information constitutes a potentially useful data source, which can be accessed at low cost and might alleviate some of the biases associated with social media data (as discussed above), it is important to note that it also has some downsides of its own. In particular, search trend information is not as rich as social media data; for example, it has very little ability to reflect user sentiment. Thus, while search volume indicates users' level of interest in a product, it may not be clear whether an increase in interest is due to positive or negative circumstances (e.g., launching of a new car model or a car model recall). See the "Related Literature" section for more details on search trend data.

The fact that search trend data and social media data have different advantages and limitations provides the setting for **our second research question**. This question examines whether sales prediction models using search trend data can obtain similar or better accuracy compared with more commonly-used sales prediction models that are based on the sentiment and volume of social media data. If sales prediction models using inexpensive search trend data can be shown to yield predictive accuracy that is comparable to or better than the predictive accuracy obtained using social media data, then managers might be encouraged to consider replacing social media data with search trend data in applied prediction settings.⁴

Finally, previous research on WOM has provided evidence that consumers' motivation to

⁴ In many practical settings, in addition to using social media data, sales prediction models utilize proprietary or private data as well as proprietary modeling procedures. Therefore, it is not possible to determine to what extent our findings would generalize to such cases.

engage in online WOM differs across brands, and that consumers may be particularly inclined to converse about highly-regarded or high-quality brands, about luxury goods, or about brands with a high degree of differentiation (Lovett et al. 2014). Such differences in online WOM activity may naturally affect the informativeness of social media data for prediction. Therefore, **our third research question** explores whether the benefits of augmenting social media data with search trend data vary across brands. Specifically, we compare between “premium” and “value” brands.

Notably, evidence suggests that the answers to our research questions depend on the specific product context in which data are being collected. In fact, information-search behaviors have been shown to vary for different levels of product involvement. Product involvement refers to consumers' interest in a product and their perceptions regarding its importance (Blackwell et al. 2001) and the perceived risk associated with it (Dholakia 2001). For instance, durable products such as cars or consumer electronics are typical examples of high-involvement products, whereas consumable products, such as groceries, movies and music, are typical examples of low-involvement products (Moorthy et al. 1997; Viswanathan et al. 2007). Simply stated, the more important the product is to a consumer, the more motivated the consumer is to search and be involved in the decision. Recently, Gu et al. (2012) observed that for a high-involvement product (digital cameras) consumers did not suffice with the WOM available on the retailer's website (Amazon.com) and actively searched for external sources of information online (consumer forums and opinion websites). The authors empirically showed that such external sources of WOM and their sentiment have a strong effect on sales of such products. This observation highlights two key premises of our work: consumers of high-involvement products actively search for information, which implies that they leave “footprints” in search logs; and at

the same time, they are strongly influenced by the WOM available on social media. Therefore, in this work we will limit our discussion to high-involvement products, for which both search and social media are expected to influence purchase decisions.

Specifically, in studying the interplay between search trend data and social media data, we will focus on the automotive industry, a classic example of a domain of high-involvement products. Given that, for most consumers, the purchase of a car is a substantial financial expense and an information-intensive decision, we expect both social media and search to have important roles in the purchase decision. Moreover, the automotive industry provides us with the opportunity to answer our third research question, as it includes both “premium” and “value” brands. Additionally, the automotive industry's vast marketing budget (estimated to spend \$15.2 billion on advertising during 2013), as well as its importance to the economy, render this industry an interesting test-bed with important practical implications.⁵

To represent search trend data, we use information from Google search query logs. To represent social media data, we rely on Google's comprehensive index of internet discussion forums (hereafter referred to as “forum data”). The latter is, to the best of our knowledge, the most comprehensive forum data set that has been made available for any academic research. 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

⁵ <http://www.adweek.com/news/advertising-branding/retail-automotive-biggest-spenders-2013-156524>

sources is at the core of this study's research questions. Once the predictive capabilities of the data have been assessed, subsequent studies can generate new theories on the basis of the obtained outcomes. Additionally, accurate sales forecasting has considerable practical implications and is a critical factor in a variety of key business processes, including inventory control, manufacturing decisions and marketing activities.

We find that 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—suggesting that the two data sources contain non-overlapping, useful, information. We also find that predictive models based on inexpensive search trend data provide predictive accuracy that is at least comparable to that of more commonly used forum-data-based predictive models. Last, we show that when adding search trend-based data to forum-based data, prediction results outperform the results of using forum-data, for both “value” and “premium” brands. Nevertheless, this difference is considerably larger for “value” car brands, while for “premium” car brands the improvement obtained by adding search trend data is more moderate.

RELATED LITERATURE

In this work we draw on and add to three main streams of research. The first studies the predictive and explanatory power of modeling using social media data; the second includes the smaller but growing body of work that documents the predictive power of search trend data; and the third studies the relationship between consumers' decision processes with regard to certain products and their involvement with those products. Finally, this work contributes to the specific field of automotive sales prediction and modeling.

Modeling Using Social Media Data

The prevalence of social media platforms in which users can publicly communicate product information to one another—platforms 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. Marketing and information systems researchers have devoted substantial attention to the effects of social media data on sales. For example, posts on websites such as Yahoo! Movies have been linked to box office revenues (Duan et al. 2008a; Liu 2006); music blog buzz has been shown to impact music listening (Dewan and Ramaprasad 2012) and sales (Dewan and Ramaprasad 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 explain TV ratings (Godes and Mayzlin 2004). Researchers have also studied the interplay between online social media mentions 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 (the firm's own website) and external mentions (other websites) on sales (Gu et al. 2012). Other researchers evaluated the positive feedback effect of sales on social media mentions (Duan et al. 2008b) and the optimal response of firms to social media data (Chen and Xie 2008; Dellarocas 2006). In addition, several studies have evaluated various moderating factors that affect the influence of social media-based mentions on sales; these factors include product and consumer characteristics (Zhu and Zhang 2010) as well as reviewer characteristics (Hu et al. 2008) and identity exposure (Forman et al. 2008). Hill et al. (2012) study the real time social media response to TV advertisements aired during the Super Bowl. In the context of the automotive industry, social

media mentions of car brands have been used to study the market structure and competitive landscape of the industry (Netzer et. al 2012).

The valence, or sentiment, of social media mentions has been shown to carry importance for predictions. However, findings on this topic are somewhat varied. For instance, Liu (2006) and Duan et al. (2008b) have found that a product's sales are affected by the volume of its social media mentions but not by the valence of these mentions or by user ratings. 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.

Biases and Limitations Associated with the Use of Social Media Data for Prediction

An essential requirement of predictive modeling is that the available data represent the underlying phenomena. In what follows we review recent literature that discuss various aspects of social media data that result in a partial or a biased representation of the underlying phenomena of interest.

Representativeness of the Consumer Population: A first potential source of bias relates to the question of *who* chooses to participate in social media discussions, and whether these individuals represent the general population. As mentioned above, industry reports suggest that relatively few consumers of social media content (approximately 10% of all social media website users) actually contribute actively to online social media discussions, and that the majority of content is produced by only 1% of the users. While the large absolute volume of social media users may

mitigate the sample size problem, another concern is whether the sample of active participants constitutes a good representation of the entire population.

Indeed, there is evidence that the minority of consumers who do participate in social media discussions are not a random representation of the general consumer population or even of the population of social media users. For example, Dellarocas and Narayan (2006) show that individuals with extreme opinions, both positive and negative, are more likely to post opinions online compared with individuals with more moderate opinions. Moreover, this biased representation may involve complex social dynamics: Moe and Schweidel (2012) report that a user's decision regarding whether to participate in an online discussion is strongly dependent on opinions that were previously posted in that discussion, and that different users are affected in different ways by the types of opinions posted. Specifically, the authors show that pre-existing discussions characterized by consensus encourage participation from more positive and less-involved consumers, whereas disagreement encourages participation from more-involved consumers.

Biased Representation of Product Evaluation: Recent research shows that even when users choose to participate in online discussions, their stated preferences and opinions may differ from their true underlying product evaluations. That is, the online social interaction may bias *what* they write. For instance, Schlosser (2005) shows in an experimental setting that users who post opinions online are influenced by the negative opinions of others because of self-presentational concerns. Moreover, in addition to showing, as discussed above, that different types of pre-existing discussions encourage subsequent participation from different types of users, Moe and Schweidel (2012) suggest that users tailor their posts in accordance with the content of those pre-existing discussions. In particular, customers who join a discussion characterized by consensus

are likely to exhibit bandwagon behavior. In contrast, users who join discussions characterized by disagreement exhibit more differentiating behavior. Similarly, Moe and Trusov (2011) show that consumers' online product ratings is influenced by previously posted ratings and quantify the sales impact of observed social dynamics.

Temporal Shifts in Online Reviews: Social media information has also been shown to be subjected to shifts in review content over time. Li and Hitt (2008) show that product reviews decline over time and suggest that this trend can be attributed to the product life-cycle. Godes and Silva (2012) observe a similar downward trend in product ratings over time and associate it with the number of reviews that a product accumulates. In particular, they suggest an *order bias*, according to which a product's average rating decreases as it obtains additional reviews. Hong et al. (2014) show that, in the case of pure search products, the variance of product ratings decreases with the number of ratings; for experience products, however, the variance of ratings may remain constant or increase, depending on the importance of the experience attributes in determining consumer utility.

Intentional Manipulation: Several recent works (Dellarocas 2006; Mayzlin et al. 2014) have pointed to the potential of firms and individuals intentionally manipulating online reviews. Luca and Zervas (2015) show that a large percentage of reviews on the Yelp website (16%) are suspected to be manipulated, and that the likelihood of review manipulation may vary substantially in accordance with various factors such as level of competition and initial reputation. Furthermore, Anderson and Simester (2014) find that 5% of reviews are posted by customers who, on the one hand, have no financial incentive to influence product ratings, yet, on the other hand, show no record of ever purchasing the product they reviewed. (These reviews are

also significantly more negative than other reviews.) The authors' findings suggest that the phenomena of product rating manipulation may be far more prevalent than expected.

Effects of Product Characteristics: Recent work has shown that different products and brands have substantially different representation in social media. These differences in online representation involve complex dynamics that go beyond the simple differences in mean and variance that have typically been handled in studies thus far. For instance, Lovett et al. (2014) relate brand characteristics to online and offline WOM patterns. They find that contribution to the online discussion is mostly influenced by social drivers (such as the level of differentiation, esteem, and relevance) as well as by functional drivers (such as complexity and familiarity). For example, they find that in the online environment people tend to talk about less complex brands, and also less about brands with lower level of differentiation. Hong et al. (2014) discuss how rating dynamics vary across product types—specifically distinguishing between experience and search products. Furthermore, they show that online reviews can help infer product type (experience or search product). Berger and Schwartz (2011) show that, compared with less-interesting products, interesting products stimulate higher levels of immediate WOM but, contrary to intuition, do not receive more ongoing WOM over multiple months or overall. Also, products that are more publicly visible receive more WOM both immediately after being experienced and over time.

Additionally, several recent studies have compared the online mentions associated with hedonic versus utilitarian products. Kronrod and Danziger (2013) show that the language that consumers use when sharing experiences about hedonic consumption differs from the language used to describe utilitarian consumption. Berger and Milkman (2012) show that high or low emotional involvement—often linked to hedonic and utilitarian features, respectively—shapes

consumers' sharing behaviors. Schulze et al. (2014) also show that hedonic and utilitarian products are associated with different viral marketing mechanisms.⁶

Handling Social Media-Driven Data Biases: We note that modeling and data handling techniques can be used to mitigate some of the social media-driven data biases. For instance, de-trending techniques may adjust for certain time-dependent changes in the volume and sentiment of social media mentions. The use of per-product dummy variables, or per-product normalization, may account for some of the differences across different products. However, most of the biases discussed above involve complex dynamics that need to be addressed through complex modeling processes that are difficult to come by (e.g., adjusting for factors such as previous postings and user personality, which influence user participation levels and distort users' stated product evaluations). Moreover, in some cases, the current level of understanding of process dynamics is limited (e.g., why do individual users with no clear vested interest post fake reviews?), thereby hindering researchers' capacity to account for these dynamics in their modeling. In other cases, process dynamics are time-dependent, and it may not be clear to what extent they continue to influence economic outcomes after the model training period is over. In sum, the complex and sometimes unclear characteristics of social media-driven bias may limit corrective modeling efforts, especially when there is a need to adjust for multiple biases or data limitations simultaneously. Indeed, social media data is used in most cases without any modeling adjustments or only with minor adjustments.

Alongside modeling choices, another common means of alleviating the effect of flaws in imperfect data, is to enrich the data with additional, meaningful, information. We suggest that

⁶ While not in the context of forum mentions, Pöyry et al (2013) find that consumers claiming a primarily utilitarian (rather than hedonic) motivation to visit a travel agency's Facebook profile are more likely to recommend that company to their friends.

search trend data can supplement social media data in this manner, as elaborated below.

THE PREDICTIVE POWER OF SEARCH TRENDS

The second stream of research we draw from focuses on the use of search engine logs, commonly known as search trends, for predicting a variety of economic and social events. In addition to its widespread use, most important for our context, is the fact that online search is conducted in private. As consumers' individual search behavior is not revealed to others, many of the factors that might bias the behavior of social media users do not apply to search behavior. When performing online search, users have no self-presentation concerns and have very limited awareness of other searchers' activity, and individual search behavior cannot normally be used to manipulate or influence sales.

Although search is conducted privately, tools such as Google Trends have recently made search logs publicly available at the aggregate level. These logs have been used for prediction by various studies in different contexts. For instance, Choi and Varian (2009, 2011) used search trend 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 Schmidt (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. Hu et al. (2014) constructed a model that uses search trend data and automotive sales data to

decompose the impact of advertising into two components: generating consumer interest in information search and converting the interest into sales. They showed that search trend data is not merely a predictor but can also represent the level of interest in different products. 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. The latter study points 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. It also discusses 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.

Although search trend information has successfully been used for prediction, it is important to note that this data source also has limitations and downsides. First, as mentioned above, compared with social media, search trend information is relatively poor in terms of content and does not reflect user sentiment. Thus, search trend data may not reflect whether interest in a product is due to positive or negative circumstances. Second, search trend information, such as that provided by the Google Trends tool, is typically available only at an aggregate level, and raw searches are generally not made available to researchers. Therefore, search trend data are not as rich and granular as social media data, which commonly include individual-level postings.⁷ Moreover, the inability to access raw searches hinders researchers'

⁷ Another known limitation of search trend data, though less relevant in the context of this study, is that the Google Trends tool imposes various limitations in order to maintain user privacy. Therefore, Google avoids presenting

capacity to detect and analyze additional possible biases and limitations that may characterize search trend data.

To our knowledge, the interplay between publicly available WOM and search trend logs has not been previously studied in the context of sales prediction. In fact, only one paper has previously contrasted the two sources of data: Luo et al. (2013) combined social media data (using web blogs and consumer ratings), Google searches and web traffic data in modeling firm equity value. However, their paper focused on a substantially different domain, raised different research questions, and used a different methodology. For example, Luo et al. modeled the two data sources jointly, thus capturing the "marginal" informativeness of each source given that the other source exists. In contrast, our paper uses a predictive research framework and compares search trend data and forum data as sources of predictive information, while also seeking to identify the added value of search trend data to the predictive accuracy of models based on social media data.

Consumer Involvement

The third stream of literature we relate to discusses the effects of consumer involvement and perceived risk on search behavior patterns. Involvement levels range from low to high, and the degree of involvement associated with a product is determined by how important consumers perceive the product to be (Blackwell et al. 2001). Dholakia (2001) defines product involvement as "an internal state variable that indicates the amount of arousal, interest or drive evoked by a product class", and suggests that involvement is strongly linked to consumer risk perception.

search volume for low-frequency search terms. The outcome is that search trend data may not be available for small scale or local phenomena.

Involvement includes both enduring factors and situational factors (the level of interest evoked in a specific situation; Bloch and Richins 1983).

As indicated above, when a consumer feels more involved with a product, he or she is more likely to act with deliberation to minimize the risk and maximize the benefits gained from purchase and use. That is, the extent to which a product is "important" to a consumer influences the degree to which the consumer is motivated to be involved in the purchase decision, e.g., by seeking out information regarding the product.

Aspects of consumer search behavior that are influenced by product involvement include the volume of search conducted, the extent to which search is active or passive, and the quantity of information the consumer is able to process (Laurent and Kapferer 1985; Zaichkowsky 1985). Nevertheless, the effects of high involvement are not limited to search behavior; high involvement with a product has been shown to serve as a motivation for spreading WOM (Lovett et al. 2014) as well as for seeking out WOM and being influenced by it (Gu et al. 2012). Thus, in this work we focus on the context of high-involvement products, where we expect both publicly available social media mentions and search trend logs to be predictive of sales.

We integrate the three literature streams reviewed above to provide new insights as to the interplay between data from social media and from search engine logs in the context of predicting the sales of high-involvement products.

Automotive Sales Prediction

Last, we note that in addition to the three wide-ranging streams of research above, our

study also draws on and contributes to the specific field of automotive sales prediction and modeling, which has received extensive coverage in different contexts in previous literature. For instance, Hymans et al. (1970) focused on the context of automotive expenditures to demonstrate the importance of including baseline information such as consumer sentiment index in modeling durable goods sales. Carlson (1978) used Seemingly Unrelated Regressions to model the demand for different sizes of automobiles. Urban et al. (1990) developed a behavioral state model for pre-launch market prediction in which customers move between different behavioral states toward making an automotive purchase decision. This model was later extended by Urban et al. (1993), who added additional behavioral factors such as categorization and elimination of alternatives. Greenspan and Cohen (1999) developed a macroeconomic model for forecasting aggregate new car sales in the US; their model considered the stock of vehicles and vehicle scrappage. Recently, Wang et al. (2011) applied a non-linear method, using an adaptive network-based fuzzy inference system. Landwehr et al. (2011) adopted a somewhat different perspective on the automotive sales prediction problem, incorporating visual car design parameters such as design prototypicality and design complexity into the prediction model. Most relevant to our context are papers that use Google Trends data in the context of automotive sales prediction. For instance, Choi and Varian (2009, 2011) predicted automotive sales and automotive parts sales in the US; Seebach et al. (2011) predicted automotive sales for two car manufacturers in Germany; and Du and Kamakura (2012) demonstrated their modeling method in conjunction with US automotive sales.

DATA AND REPRESENTATION

This research uses monthly data for 23 car brands (all brands with average sales above 5,000 cars per month) 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 (for example, Choi and Varian 2009; Du and Kamakura 2012; Seebach et al. 2011) we focus on brand-level sales rather than specific car model sales (e.g., the *BMW* car brand, rather than the *528i* car model).⁸ See Appendix B for the list of brands. Nevertheless, for robustness, we also report on predicting car-model-level sales (see online appendix – section 5), reaching similar findings.

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

⁸ The motivation for this practice is twofold. First, brand-level data are much more abundant than car model data. Second, keyword identification is considerably more accurate at the brand level than at the car model level (see discussion below on keyword selection).

i during month t .⁹

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. Specifically, we extracted data from all English-language forums indexed by Google’s discussion forum search.¹⁰ This index includes dedicated websites in addition to websites that include sections in which users can publicly post opinions and reviews, as well as relate to previous content. (Examples include townhall-talk.edmunds.com, forums.motortrend.com, answers.yahoo.com, etc.)

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 dictionary¹¹ and summed the number of new forum posts mentioning “positive words” and posts mentioning “negative words”

⁹ As detailed in Choi and Varian (2009), Google Trends data are computed by a sampling method and therefore may contain some noise. To reduce the noise we use a similar procedure as in Preis et al. (2013) of averaging the value of multiple draws from Google Trends.

¹⁰ Results for Google’s discussion forum search are currently presented when selecting the “more” option under Google’s search box, and subsequently selecting “Discussions”.

¹¹ <http://www.wjh.harvard.edu/~inquirer/>

alongside brand i during month t . The advantages of using this dictionary approach are its generalizability and reproducibility (in contrast to the case of proprietary or “black box” types of sentiment analysis solutions).¹²

Keywords

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

Let K denote a set of keywords 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 keyword).

In general, when selecting a set of keywords to identify a given brand, there is a tradeoff between accuracy and coverage. Clearly, inclusion of a larger number of keywords 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,

¹² Our findings reported in later stages show that despite its simplicity, this type of sentiment representation considerably improves predictive accuracy (see Figure 1).

adding search terms such as “Malibu” or “Spark” may also introduce a large number of irrelevant queries, e.g., queries relating to the city 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 keywords, as a search using this term is a subset of the searches using the keyword “Chevrolet”.

To the best of our knowledge, the literature does not offer a methodology for optimal selection of keywords 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 keywords (e.g., “Chevrolet” for the Chevrolet brand), similarly to Seebach et al. (2011).¹³ (See Appendix B for a detailed list of the keywords we used.)

While brand-level keywords can naturally provide high accuracy in capturing brand-related search queries (or forum mentions), we also adapted our modeling procedures to mitigate coverage concerns. First, as mentioned above, we expected that brand-level keywords (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 note that when a prediction model is constructed for each car brand, as long as the keyword 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 keyword 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

¹³ Seebach et al. (2011) reported that using brand-level keywords produced the best results in a similar scenario of automotive sales prediction using search trend data.

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.

Third, 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.

For the reasons mentioned above, in what follows, in the main body of this paper we report on predictions at the brand level, using brand-level keywords. Nevertheless, in the online appendix we also report about car-model-level sales predictions (see online appendix section 5). Additionally, for robustness we examined a different keyword methodology involving a combination of brand-level and car-model-level keywords for predicting brand-level sales. This analysis is also reported in the online appendix (section 6). In all robustness checks we obtained similar findings.

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 $t - 1$. Predictors include sales in previous months, forum data, and search trend data (as elaborated above), as well as benchmark data. 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: $t - 1, \dots, t - 5$).¹⁴

Following previous research in this domain, we utilized the following benchmark data:

- Seasonality: sales in the same month, in the previous year (i.e., $Sales_{i,t-12}$). Usage of such data to represent seasonality is common in autoregressive models in this domain (see, for example, Choi and Varian 2009, 2011) due to the cyclic variation in customer automotive purchase patterns.
- Consumer sentiment index (see, for example, Hu et al. 2014; Hymans et al. 1970): This is a US national economic indicator based on a survey of a representative sample of the US population. It is designed to depict how consumers view their own financial situations and short/long-term economic conditions; thus, it has high relevance for consumer car purchase decisions. Consumer sentiment index is reported by the University of Michigan and Thompson Reuters.
- Gasoline prices (see, for example, Hu et al. 2014): This national average price information is collected and reported by the US Energy Information Administration and is based on the retail prices provided by a sample of approximately 800 gasoline stations. As gasoline prices influence the total cost of vehicle ownership, this economic indicator is expected to be associated with consumer car purchase decisions.

Having collected the data, we defined a “*benchmark model*” as a model that utilizes consumer sentiment, gasoline price, seasonality ($Sales_{i,t-12}$), and previous sales data.

Subsequently, in order to gauge the informativeness of forum-based data and search trend data,

¹⁴ The importance of using lagged data in the prediction model setup was recently discussed in Goel et al. (2010) and Lazer et al (2014).

as well as the benefit of augmenting forum-based data with search trend data, we defined several additional models incorporating different sets of data, as follows: The “*forum-based model*” utilizes the benchmark 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.

Table 1. Data Included in Each Model

	<i>Benchmark Model</i>	<i>Forum-Based Model</i>	<i>Extended Forum-Based Model</i>	<i>Search Trends-Based Model</i>	<i>Combined Search and Forum-Based Model</i>
Sales _{i,t-1} , ..., Sales _{i,t-n}	√	√	√	√	√
Consumer Sentiment _{t-1}	√	√	√	√	√
Gasoline price _{t-1}	√	√	√	√	√
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}				√	√

Prediction and Evaluation

We utilized the popular least-squares 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. For robustness, we repeated our analysis with nonlinear methods such as Neural Networks (NN), Support Vector Machines (SVM), and Random Forest, and obtained similar findings. We report the NN results in the online appendix (section 1).

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, we used the Moving Window approach. Implementing this method, we followed common practice and used 1/3 of our data as an independent validation set. We report performance on the basis of the entire out-of-sample validation period (months $t = 25, \dots, 36$). That is, for each validation month t , we measure performance while applying the model trained during the 24 preceding months (months $t - 24$ to $t - 1$). We note that month $t = 1$ is January 2008 and month $t = 25$ is January 2010.¹⁵ For robustness, we carried out a similar analysis using the Expanding Window approach, and obtained similar findings. These results are provided in the online appendix (section 2).

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 overestimation or underestimation). This is appropriate for our

¹⁵ Our data are for January 2007 - December 2010. We "lose" 12 months' worth of data when accounting for seasonality.

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. For robustness, we repeated the analysis using mean square error (MSE) criteria and reached similar findings. These results are provided in the online appendix (section 6).

Finally, we note that we transformed each of the variable representations into normalized values (for each brand). There are two motivations for normalization at the brand level. First, normalization 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 keyword handling methodology. Nevertheless, we note that while our models use normalized variables, in order to provide interpretable results, we calculated the MAPE according to the actual, "de-normalized" numbers. We further note that in deriving the results reported below, in order to avoid "information leakage" from the validation set data, we used normalization/de-normalization procedures in which the normalization factors (sample mean and standard deviation) were calculated solely on the basis of the specific training set data used in each "moving window" iteration.

RESULTS

Figure 1 displays the results obtained using the different data representations (i.e., the different model types defined in Table 1). Table 2 presents the differences in MAPE values between models utilizing different sets of data and the corresponding significance values, using a bootstrap confidence interval (in this table, a positive difference between the MAPE values indicates that Model A's predictive accuracy is better than Model B's predictive accuracy). The

results reported include models with one lag of data and with two lags of data. Notably, although we tested prediction models using up to five lags of data, we found that adding data from lag 3 or higher actually degraded predictive accuracy for all the models.

Recall that our first research question examines whether predictive models based on a combination of search trend data and forum data can obtain more accurate sales predictions compared with models based on forum data alone. Our first core finding is that the combined model is superior to models based on forum data alone. Specifically, we find that augmenting forum data with search trend data significantly improves prediction accuracy: When using one lag of data, the combined model yields an improvement in MAPE of 0.55% compared with the forum-based model, and when using two lags of data, the combined model yields an improvement in MAPE of 0.58%. Similarly, when comparing the results of the combined model versus the more elaborate extended forum-based model (which uses forum sentiment scores in addition to forum mentions), we observe that with one lag of data the combined model yields an improvement in MAPE of 0.42%, and with two lags of data it yields an improvement in MAPE of 0.37%. These findings are significant and suggest that search trend data contain additional valuable information not available in forum data. We obtain similar significant findings when using a nonlinear NN algorithm (see online appendix section 1).

Our second core finding relates to the second research question, which examines whether predictive models using search trend data obtain similar or better accuracy compared to predictive models utilizing forum-based data (either forum mention volume or forum mention volume combined with forum sentiment). We find that the performance of search trend-based models is, at least, comparable to that of forum-data-based models. Specifically, with one lag of data, the search trend-based model obtained an improvement in MAPE of 0.21% compared with

the forum-based model, and with two lags of data it yielded a smaller improvement in MAPE of 0.17%. Compared with the more elaborate extended forum-based model (i.e., the model that incorporates both forum mentions and sentiment scores), the search trend-based model obtained a small improvement in MAPE of 0.08% with one lag of data, and with two lags of data it yielded a minor deterioration in MAPE of -0.03%. The observation that the predictive performance of models based on search trend data is, at least, comparable to that of models based on forum data is also consistent with our robustness check using the nonlinear NN algorithm. In fact, when using NN, we find that models based on search data not only obtain comparable results to models based on forum and extended forum data, but also significantly outperform them (see online appendix 1).

Figure 1. Prediction Results (LR)

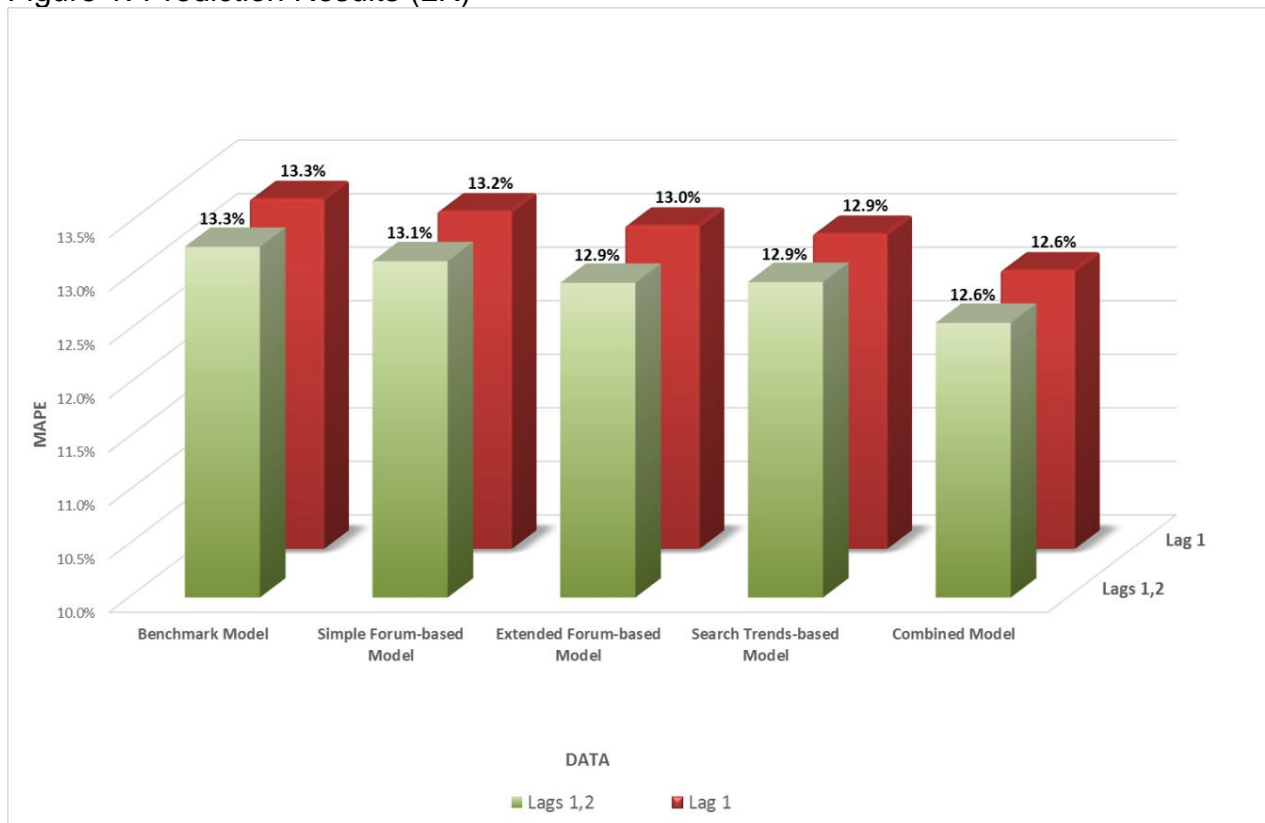


Table 2. MAPE Differences and One-Sided Confidence Intervals for the Difference in MAPE Values Using LR Method

Model A	Model B	MAPE(Model B) - MAPE(Model A)	
		LR - Lag 1	LR - Lag 1,2
Forum-Based Model	Benchmark Model	0.11%**	0.13%**
Extended Forum-Based Model	Benchmark Model	0.25%**	0.34%**
Search Trends-Based Model	Benchmark Model	0.32%***	0.31%**
Combined Model	Benchmark Model	0.67%***	0.71%***
Search Trends-Based Model	Forum-Based Model	0.21%*	0.17%
Search Trends-Based Model	Extended Forum-Based Model	0.08%	-0.03%
Combined Model	Forum-Based Model	0.55%***	0.58%***
Combined Model	Extended Forum-Based Model	0.42%***	0.37%***
Combined Model	Search Trends-Based Model	0.34%**	0.40%**

Table 2 reports the difference in MAPE using two models (*Model A* and *Model B* - each based on different data inputs) while considering 1 or 2 lags with the LR algorithm.

Specifically, the table reports the difference: $\text{diff} = \text{MAPE}(\text{Model B}) - \text{MAPE}(\text{Model A})$. Therefore, a positive value associated with the comparison between Model A and Model B indicates better predictive accuracy of Model A over Model B.

Lower confidence interval bounds for *diff* were calculated using 2000 iterations of the BCA bootstrapping confidence interval calculation method implemented in R software. A lower confidence interval bound for *diff*, with a positive value, provides confidence that $\text{MAPE}(\text{Model A})$ is indeed better than $\text{MAPE}(\text{Model B})$.

We report the following lower confidence bounds:

- * 0.9 lower confidence bound for *diff* is positive
- ** 0.95 lower confidence bound for *diff* is positive
- *** 0.99 lower confidence bound for *diff* is positive

Prediction Results According to Brand Characteristics

To gauge additional aspects of adding search trend data to forum data, we explored the extent to which predictive accuracy is dependent on the characteristics of the brand for which predictions are being made. Specifically, we focus on “premium” (luxury) car brands (i.e., brands with higher pricing or higher perceived quality or that elicit higher willingness to recommend) versus “value” car brands. As mentioned earlier, one of the fundamental differences between our two data sources—social media and search trends—is in their visibility: While search is conducted in

private, social media mentions are publicly visible. Specifically, previous work suggests that, because of their visibility, forum mentions may trigger biases (as discussed above) and may also be associated with different outcomes in the cases of premium versus value brands. In effect, research in the context of online as well as offline WOM has provided much evidence that consumers' motivation to engage in WOM is impacted by social drivers such as the need for self-enhancement, and that WOM behavior is used for social signaling (Sundaram et al. 1998, Henning-Thurau et al, 2004). Hence, consumers might be more inclined to converse about highly-regarded or high-quality brands (Amblee and Bui 2008), about luxury goods that signal high social status, or about brands with a high degree of differentiation, to express uniqueness (Lovett et al. 2014). In the context of forum mentions, this suggests that premium brands will be better represented in the data. (Indeed, our data show that, relative to sales volume, premium brands have on average 2.1 times more forum mentions compared with value brands.) We therefore expect that when predicting sales volume, adding search trend data to forum-based data will have a stronger impact on improving prediction accuracy for value brands rather than for premium brands (which are already better represented in forum data).

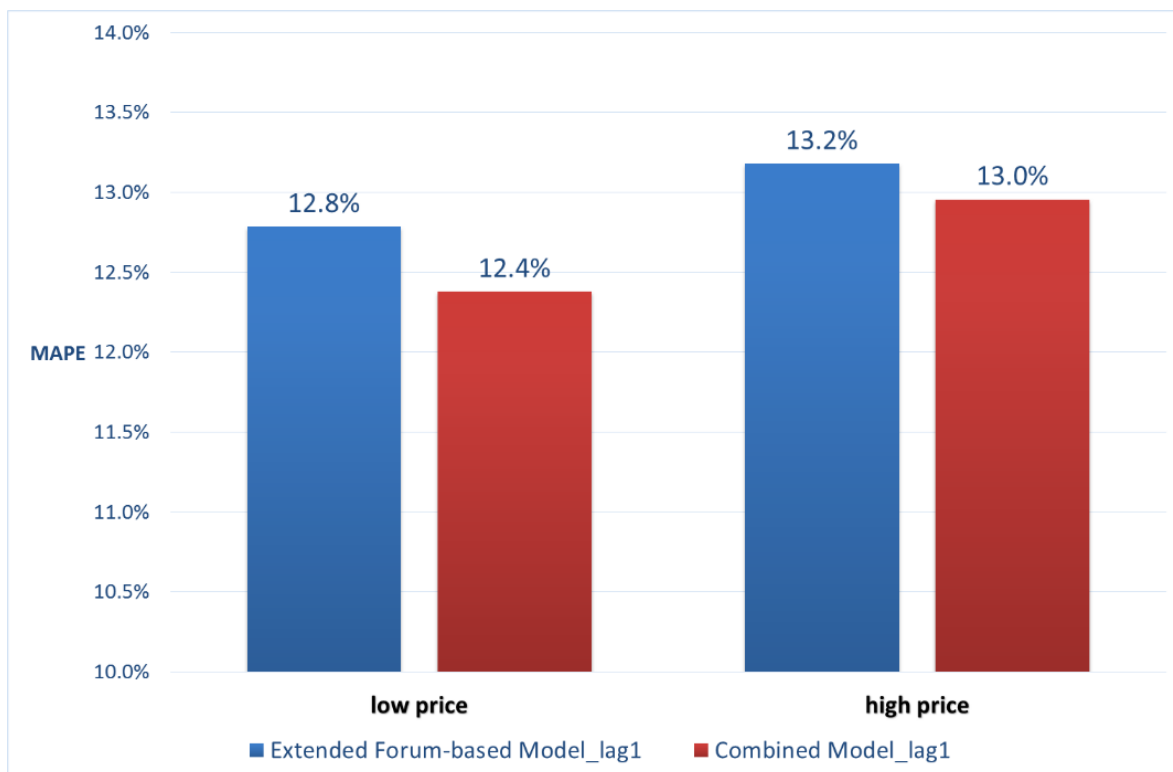
In what follows, the car brands are split into two subsets to represent the value and premium segments. Specifically, we split the car brands based on price, perceived quality or willingness to recommend, as follows. Car brands for which the list price for the least expensive car model of each brand was more than \$20,000 were referred to as "high price", and the rest of the car brands were referred to as "low price".¹⁶ Next, we obtained survey data about perceived quality and willingness to recommend each brand. These data were obtained from YouGov plc, a market research firm that monitors a panel of 5,000 people in the US, on a daily basis, and

¹⁶ Prices for 2007. Source: Automotive News website.

reports on brand-related perceptions. See online appendix (section 4) for more details about these survey data. For this analysis, we refer to the 12 brands with lower perceived quality as brands with “low perceived quality” and the remaining 11 brands as brands with “high perceived quality”. Similarly, we refer to the 12 brands associated with lower willingness to recommend as brands with “low willingness to recommend” and the remaining 11 brands as brands with “high willingness to recommend”.¹⁷

Figure 2 shows a comparison between the performance of the combined model and that of the extended-forum-based model in predicting the sales of “high price” (premium) versus “low price” (value) brands, with one lag of data.

Figure 2. Prediction Results for Premium vs. Value Brands (Split by Price)



¹⁷ Since our data set includes 23 brands, it cannot be precisely split into 2 groups with equal sizes. Nevertheless, the results are robust to including the median brand in either group.

The results show that for both premium and value brands, adding search trend data to the extended forum data, yielded better (lower) MAPE values compared with the extended forum-based model. Interestingly, for value brands, the combined model significantly outperformed the extended forum-based model, yielding an improvement in MAPE of 0.41%. In contrast, for premium brands, the difference was smaller (the combined model yielded an improvement in MAPE of only 0.2% as compared with the extended forum-based model) and was not significant. These results support the conjecture that, when attempting to predict car sales, adding search trend data to forum-based data will have a stronger impact on improving prediction accuracy for value brands rather than for premium brands.

For robustness, we repeated the analysis splitting the brands according to perceived quality or willingness to recommend metrics (see online appendix section 3), using a model with two lags of data (see online appendix section 3), and with car model level data (see online appendix section 5). In all analyses we obtained similar findings, namely, that adding search trend data to forum data improves predictive accuracy to a greater extent in the case of value brands than in the case of premium brands.

CONCLUSIONS

In this paper we empirically studied the interplay between search trend data and publicly available social media mentions from social media websites in the context of sales prediction. While social media data have been used extensively for sales predictions, previous literature has pointed to various limitations and biases associated with this source of data. One possible method of improving the predictive accuracy of this data source is to overlay it with search trend

data. However, thus far, the two data sources have largely been investigated in disparate literature streams and for separate purposes.

Prior literature suggests that, when attempting to obtain evidence related to customers' decision making processes, it is necessary to distinguish between different types of products. Specifically, consumers are more likely to engage in extensive and active information search in the case of high-involvement products than in the case of low-involvement products. Therefore, in this paper we focused on high-involvement products.

Using data from the automotive industry, we have provided first evidence that augmenting forum-based models with search trend data significantly improves predictive accuracy. This evidence indicates that the sales-relevant information embedded in search data is external and non-overlapping to the corresponding information embedded in forum data. From a practical perspective, this finding suggests that companies that have already invested in collecting forum-based data for modeling purposes can significantly improve forecasting accuracy with a relatively small additional investment in collecting search trend data. Our findings further suggest that the prediction accuracy that can be achieved using search trend data alone is comparable to that associated with the more commonly used forum-based data. This finding could encourage managers to evaluate using search trend data as a low-cost replacement for social media data. Finally, we find that adding search trend data to forum-based data has a stronger impact on improving prediction accuracy for value brands rather than for premium brands.

Our work carries managerial implications for car manufactures and, more broadly, for manufacturers of high-involvement products. Moreover, the advantage of our method is that it does not require proprietary data available only to the manufacturer. Hence, it can be used by

upstream and downstream players, as well as by stock market investors. Furthermore, car manufacturers can use this approach to evaluate the expected sales of their competitors. More accurate sales prediction models 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 and, therefore, even small improvements in forecasting accuracy are expected to have a considerable effect.

We expect that our findings may be generalizable to a wide array of purchase decisions regarding high-involvement products, such as housing purchases and travel planning. In the case of low-involvement products, such as music, mobile applications, and movie tickets, consumers do not conduct extensive and active search, and decisions are made more lightheartedly. The predictive power of search trends in this context is therefore unclear. This raises an interesting direction for future research.

Other possible avenues of future work include incorporating information about the popularity of different discussion forum websites in the prediction models. Additionally, it is possible to analyze the predictive value of search trend and forum data according to additional brand characteristics and model characteristics, such as vintage. It would also be interesting to study how the incorporation of predictions based on the investigated data sources into the managerial decision making process interacts with these data and affects their predictive capacity over time. Additionally, from a methodological perspective, it would be interesting to compare predictions based on publicly available information to industry experts' forecasts using proprietary data. Finally, this study demonstrated that search trend data can successfully complement forum data in the context of sales prediction. Nevertheless, forum data have been

used for other purposes such as detecting changes in consumer interest in products. Therefore, it would be interesting to evaluate the effectiveness of using search trend data for augmenting forum-based data in this context.

APPENDIX A – LIST OF MAIN ROBUSTNESS CHECKS

Robustness Check	Section in Online Appendix
Predictive capacity using NN algorithm	Section 1
Predictive capacity using Expanding Window approach	Section 2
Predictive capacity according to “premium” and “value” car brand characteristics based on price, perceived quality, and willingness to recommend metrics. Using one or two lags of data.	Section 3
Car model-level Analysis	Section 5
Predictive capacity using extended keyword selection	Section 6
Predictive capacity using MSE criteria	Section 6

APPENDIX B – LIST OF BRANDS AND GROUPING BY PRICE, PERCEIVED QUALITY, AND WILLINGNESS TO RECOMMEND

Table B1 provides details about the brands included in this study. The list of brands includes all car brands with average US sales exceeding 5,000 cars per month during 2007-2010 (source: Automotive News).¹⁸ Quality and Willingness to Recommend rankings are based on the YouGov BrandIndex product. (See additional details in the online appendix section 4.)

¹⁸ Combined search volume of multiple keywords can be obtained from Google Trends by utilizing the “+” sign between different terms.

Table B1. List of Brands and Grouping by Price, Perceived Quality, and Willingness to Recommend

Brand	Keyword(s)	Price	Average Quality Ranking	Average Recommend Ranking
Acura	acura	high price	high	high
Audi	audi	high price	high	low
Bmw	bmw	high price	high	high
Buick	buick	high price	low	low
Cadillac	cadillac	high price	high	low
Chevrolet	chevrolet, chevy	low price	low	high
Chrysler	chrysler	low price	low	low
Dodge	dodge	low price	low	low
Ford	ford	low price	low	high
Gmc	gmc	low price	low	low
Honda	honda	low price	high	high
Hyundai	hyundai	low price	low	low
Infiniti	infiniti	high price	high	low
Jeep	jeep	low price	low	low
Kia	kia	low price	low	low
Lexus	lexus	high price	high	high
Lincoln	lincoln	high price	low	low
Mazda	mazda	low price	low	low
Mercedes Benz	mercedes	high price	high	high
Nissan	nissan	low price	high	high
Subaru	subaru	low price	low	high
Toyota	toyota	low price	high	high
Volkswagen	volkswagen	low price	high	high

REFERENCES

- Amblee, N., and Bui, T. 2008. "Can Brand Reputation Improve the Odds of Being Reviewed On-Line?," *International Journal of Electronic Commerce* (12:3), pp. 11-28.
- Anderson, E. T., and Simester, D. I. 2014. "Reviews without a Purchase: Low Ratings, Loyal Customers, and Deception," *Journal of Marketing Research* (51:3), pp. 249-269.
- Berger, J., and Milkman, K. L. 2012. "What Makes Online Content Viral?," *Journal of Marketing Research* (49:2), pp. 192–205.

- Berger, J., and Schwartz, E. M. 2013. "What Drives Immediate and Ongoing Word of Mouth?," *Journal of Marketing Research* (48:5), pp. 869-880.
- Blackwell D., Miniard P. W., and Engel, J. F. 2001. *Consumer Behavior*, 9th ed. Orlando, FL: Harcourt.
- Bloch, P. H., and Richins, M. L. 1983. "A Theoretical Model for the Study of Product Importance Perceptions," *Journal of Consumer Research* (47), pp. 69-81.
- Carlson, R. L. 1978. "Seemingly Unrelated Regression and the Demand for Automobiles of Different Sizes, 1965-75: A Disaggregate Approach," *Journal of Business* (51:2), pp. 243-262.
- 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.
- 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), pp. 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.
- 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.
- Dellarocas, C., and Narayan, R. 2006. "A Statistical Measure of a Population's Propensity to Engage in Post-Purchase Online Word-of-Mouth," *Statistical Science* (21:2), pp. 277–285.
- Dewan, S., and Ramaprasad, J. 2009. "Chicken and Egg? Interplay between Music Blog Buzz and Album Sales," in *PACIS 2009 Proceedings*.
- Dewan, S., and Ramaprasad, 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.
- Dholakia, U. M. 2001. "A Motivational Process Model of Product Involvement and Consumer Risk Perception," *European Journal of Marketing* (35:11), pp. 1340–1360.
- 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.
- 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.
- Godes, D., and Silva, J. C. 2012. "Sequential and Temporal Dynamics of Online Opinion," *Marketing Science* (31:3), pp. 448-473.
- 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.
- Greenspan, A., and Cohen, D. 1999. "Motor Vehicle Stocks, Scrappage, and Sales," *The Review of Economics and Statistics* (81:3), pp. 369-383.
- 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.
- Hill, S., Nalavade, A., and Benton, A. 2012. "Social TV: Real-Time Social Media Response to TV Advertising," proceedings ADKDD 2012 (Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy). Article 4.
- Hong, Y., Chen, P.-Y., and Hitt, L. M. 2014. "Measuring Product Type with Dynamics of Online Product Review Variances: A Theoretical Model and the Empirical Applications," working paper.

- 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.
- Hu, Y, Du, R. Y., and Damangir, S. 2014. "Decomposing the Impact of Advertising: Augmenting Sales with Online Search Data," *Journal of Marketing Research* (51: 3), pp. 300-319.
- Hymans, S. H., Ackley, G., and Juster, F. T. 1970. "Consumer Durable Spending: Explanation and Prediction," *Brookings Papers on Economic Activity*, pp. 173-206.
- Kronrod, A., and Danziger, S. 2013. "'Wii Will Rock You!' The Use and Effect of Figurative Language in Consumer Reviews of Hedonic and Utilitarian Consumption," *Journal of Consumer Research* (40), pp. 726-739.
- Landwehr, J. R., Labroo, A. A., and Herrmann, A. 2011. "Gut Liking for the Ordinary: Incorporating Design Fluency Improves Automobile Sales Forecasts," *Marketing Science* (30:3), pp. 416-429.
- Laurent, G., and Kapferer, J. 1985. "Measuring Consumer Involvement Profiles," *Journal of Marketing Research* (22:1), pp. 41–53.
- Lazer , D., Kennedy, R., King, G., and Vespignani, A. 2014. "The Parable of Google Flu: Traps in Big Data Analysis," *Science* (343:6176), pp. 1203-1205.
- Li, X., and Hitt, L.M. 2008. "Self Selection and Information Role of Online Product Reviews," *Information Systems Research* (19:4), pp 456-474.
- 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. 2014. "On Brands and Word-of-Mouth," *Journal of Marketing Research*, forthcoming.
- Luca, M., and Zervas, G. 2015. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," Harvard Business School NOM Unit Working Paper No. 14-006. Available at SSRN:<http://ssrn.com/abstract=2293164>.
- Luo, X., Zhang, J., and Duan, W. 2013. "Social Media and Firm Equity Value," *Information Systems Research* (24:1), pp.146-163.
- Luo, X., Zhang, J., Gu, B., and Phang, C. W. 2014. "The Influence of Online Word-of-Mouth on Long Tail Formation," *Decision Support Systems*, forthcoming.
- Mayzlin, D., Dover, Y., and Chevalier, J. 2014. "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review* (104:8), pp. 2421-55.
- Moe, W. W., and Schweidel, D. A. 2012. "Online Product Opinions: Incidence, Evaluation and Evolution," *Marketing Science* (31:3), pp. 372-386.
- Moe, W. W., and Trusov, M. 2011. "The Value of Social Dynamics in Online Product Ratings Forums," *Journal of Marketing Research* (48:3), pp. 444-456.
- Moorthy, S., Ratchford, B. T., and Talukdar, D. 1997. "Consumer Information Search Revisited: Theory and Empirical Analysis," *Journal of Consumer Research* (23:4), pp. 263-277.
- Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M. 2012. "Mine Your Own Business: Market-Structure Surveillance Through Text Mining," *Marketing Science* (31:3), pp. 521-543.

- Pöyry, E., Parvinen, P., and Malmivaara, T. 2013. "Can We Get from Liking to Buying? Behavioral Differences in Hedonic and Utilitarian Facebook Usage," *Electronic Commerce Research and Applications* (12:4), pp. 224–35.
- Preis, T., Moat, H. S., and Stanley, H. E. 2013. "Quantifying Trading Behavior in Financial Markets Using Google Trends," *Scientific Reports*, 3.
- Rui, H., Liu, T., and Whinston, A. 2012. "Whose and What Chatter Matters? The Impact of Tweets on Movie Sales," working paper.
- Schlosser A.E., 2005. "Posting versus Lurking: Communicating in a Multiple Audience Context," *Journal of Consumer Research* 32, pp. 260-265.
- Schulze, C., Schöler L., and Skiera, B. 2014. "Not All Fun and Games: Viral Marketing for Utilitarian Products," *Journal of Marketing* (78:1), pp. 1-19.
- Schweidel, D. A., and Moe, W. W. 2014. "Listening In on Social Media: A Joint Model of Sentiment and Venue Format Choice," *Journal of Marketing Research* (51:4), pp. 387-402.
- 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.
- Sundaram, D.S., Mitra, K., and Webster, C. 1998. "Word-Of-Mouth Communications: a Motivational Analysis," *Advances in Consumer Research* (25), pp. 527-531.
- Sridhar, S., and Srinivasan, R. 2013. "Social Influence Effects in Online Product Ratings," *Journal of Marketing* (76:5), pp. 70-88.

- Urban, G. L., Hauser, J. R., and Roberts, J. H. 1990. "Prelaunch Forecasting of New Automobiles," *Management Science* (36:4), pp. 401-21.
- Urban, G. L., Hulland, J. S., and Weinberg, B. D. 1993. "Premarket Forecasting for New Consumer Durable Goods: Modeling Categorization, Elimination, and Consideration Phenomena," *Journal of Marketing* (57:2), pp. 47-63.
- Viswanathan, S., Gosain, S., Kuruzovich, J., and Agarwal, R. 2007. "Online Infomediaries and Price Discrimination: Evidence From the Auto-Retailing Sector," *Journal of Marketing* (71:3), pp.89-107.
- Vosen, S., and Schmidt, T. 2011. "Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends," *Journal of Forecasting* (30:6), pp. 565-578.
- Wang, F.-K., Chang, K.-K., and Tzeng, C.-W. 2011. "Using Adaptive Network-Based Fuzzy Inference System to Forecast Automobile Sales," *Expert Systems with Applications* (38:8), pp. 10587-10593.
- 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*.
- Zaichkowsky, J. L. 1985. "Measuring the Involvement Construct," *Journal of Consumer Research* (12:3), pp. 341-352.
- 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.