Investigating Consumer Information Search Behavior and Consumer Emotions to Improve Sales Forecasting

**Full Paper**

Christoph Skodda  
Goethe University  
Frankfurt/Main, Germany  
christoph.skodda@web.de

Janek Benthaus  
Goethe University  
Frankfurt/Main, Germany  
jbenthaus@wiwi.uni-frankfurt.de

**Abstract**

Sales forecasting is an essential topic for organizations and accurate forecasts can help to establish competitive advantage. Organizations can integrate a variety of information sources into their forecasting process. In this regard, the information search behavior of potential consumers as well as their emotions about a brand are supposed to improve forecasts. In this study, we combine these two data sources to forecast sales figures of two large automobile manufacturers. We analyze data from Google's search engine as well as microblogging data from Twitter to perform in-sample and out-of-sample analyses. The results show that the conjunction of these data sources allow superior sales forecasting in contrast to an individual consideration of information search behavior and consumer emotions. Hence, data about consumer information search behavior should be enriched with the emotional state of consumers to support organizational decision making.

**Keywords**

Consumer Information Search Behavior, Sales Forecasting, Sentiment Analysis, Decision Support.

**Introduction**

Today's organizational environment is increasingly characterized by complexity and a need to gather and process relevant information to support decision-making (Schenkel and Teigland 2008). Under these conditions, highly dynamic environments require companies to constantly adapt their decisions to new events (e.g., introducing new products) to achieve competitive advantage (Overby et al. 2006). In order to be able to make appropriate decisions in quickly changing environments, processing information for decision support close to real-time is important (Chaudhuri et al. 2011). For this purpose information can be derived from a company's environment with the help of "their employees, formal and informal ties with outsiders such as customers" (Anand et al. 2002). A valuable information source for companies is analyzing the consumer behavior about ongoing purchase decisions. In the process of purchasing a product, consumers typically gather information to increase their knowledge about the product to make a profound purchase decision (Mullins and Walker 2013). Online search engines are intensively used as starting point of a search and consumers reveal information about their search behavior (Otterbacher 2008). Consequently, consumer information search is a constantly explored research area in consumer and marketing science (Schmidt and Spreng 1996). Research on consumer search behavior has shown that it can provide deeper insights into the process of making purchase decisions and provide a basis to forecast behavior of consumers (Blackwell et al. 2005; Morrison 1979). Nevertheless, the progress of research is limited and only a few studies (e.g., Seebach et al. 2011; Shmueli and Koppius 2011) have dealt with using online information about consumer information search behavior for forecasting purposes.

The use of Internet platforms and services (i.e., social media or online shopping) lead to an extensive creation of publicly available information (Chaudhuri et al. 2011; Kim and Ratchford 2012; Watson and Wixom 2007). Companies are confronted with a steadily increasing amount of available information
regarding their products and services (e.g., shared opinions, consumer reviews, blog posts) that can be analyzed for decision support. The expressed opinions of the public affects the purchase decisions of individuals (Feldman 2013; Sobkowicz et al. 2012). Fleishman-Hillard and Harris Interactive (2012) found that the Internet is today the most source for product search today and 69 percent of the Internet users give feedback on brands, products, and services on websites. Hence, consumers use the Internet to receive suitable information to improve decision-making (Liu 2010; Liu et al. 2010; Pang and Lee 2008). The identification of emotions of an author about a specific topic can be done using a sentiment analysis. Previous research on opinion analysis investigated in which way and sentiment algorithms were able to detect sentiment in blogs, consumer reviews, etc. (Bollen et al. 2011a; Bravo-Marquez et al. 2013; Stieglitz and Dang-Xuan 2013; Thelwall et al. 2012).

It thus appears that it is becoming increasingly important for companies to get access to data regarding consumer information search behavior prior to a purchase decision and the overall emotions regarding the company or specific products and services. Combining information about consumer search behavior and the consumer sentiments regarding products reveal new opportunities for forecasting sales. Therefore, this paper aims to demonstrate a conjunction of online search data from Google and microblogging data from Twitter that give indications of consumer information search behavior and emotions about a brand. Empirically, we will investigate how such data can be used to forecast sales in the automobile industry. Hence, in this paper, we will investigate the following research question: How information about consumer search behavior and consumer sentiments can be used to forecast sales? Our study contributes to the research on sales forecasting by presenting the connection of data about consumer information search behavior and consumer emotions.

The remainder of this paper is organized as follows. Chapter 2 provides the theoretical foundation. Afterwards, an empirical analysis will be conducted to investigate the relationships of the amount of consumer information search and expressed sentiments with sales figures. Finally, a conclusion based on theoretical and practical contributions as well as limitations and implications for future research will be presented.

**Theoretical Background**

The consumers need for additional information about products or services making a purchase decision (Grant et al. 2007) is described in the process of consumer decision-making depicted in Figure 1 (Mullins and Walker 2013; Seebach et al. 2011). The process starts with the detection of a consumption need and the consumer focusing on a specific product group, suitable to satisfy this need (Kotler 2001). In the second step, consumers collect information via an information search based on internal information from previous experiences or external information such as the Internet or magazines (Mullins and Walker 2013). After a successful information search, consumers identify a certain set of brands and even more importantly, a set of product attributes which are important for them. The consideration of these different influencing factors is executed with the help of the individual relative importance of each attribute or brand. Based on the prioritization of product attributes and brands consumers are then able to make the purchase decision (Mullins and Walker 2013). The post-purchase evaluation determines the last step of the process and has the aim to identify a certain product satisfaction, which is a comparison of expectations and the performance of the purchased product (Kotler and Keller 2012; Mullins and Walker 2013). In our study, we focus on the second step – information search – where consumers collect information prior to a purchase.

![Figure 1. Consumer Decision-making Process](image)

A purchase decision and the related information search differ in their importance for the consumer based on whether products require high- or low-involvement (Mullins and Walker 2013). If a consumer frequently buys inexpensive or so-called everyday products (e.g., food, beverages, etc.) a limited decision
and information search is necessary prior to the purchase. Such types of products are characterized as low-involvement products and the risk of making a wrong decision is negligible (Assael 1995; Kotler 2001; Mullins and Walker 2013). In contrast to that, the purchase of a high-involvement product which is typically more expensive and infrequently purchased (e.g., an automobile) is characterized by a quite complex purchase decision. These kinds of goods have psychological importance for the consumer because of a direct connection to social requirements that need to be fulfilled and a higher identification with the product or the brand (Martin 1998; Mullins and Walker 2013). Another psychological risk can be ascribed to the increase of product choices which might lead to information overload and a more difficult decision-making process (Mullins and Walker 2013; The Economist 2010). Besides that, financial risks are highly relevant because high-involvement products are in general costly and lead to a loss of purchasing power (Lamb et al. 2011). Thus, a wrong purchase decision and corresponding feelings of guilt could lead to anxiety (Assael 1995).

For this reasons, consumers typically feel uncertain about making a high-involvement purchase decision. The purchase decision has to be made very carefully and under high effort (e.g., time, money) (Baumgartner 2002). Consequently, consumers need to gather sufficient knowledge and try to reduce the risk of a wrong decision and increase product satisfaction with the help of information search that can be divided into internal and external search (Schmidt and Spreng 1996). Internal search is characterized by information which is stored in memory from previous search processes or personal experiences. In contrast, consumers who perform an external search have to gather information from their environment (Mullins and Walker 2013; Schmidt and Spreng 1996). An external search is necessary if the specific information search process takes place for the first time or the information cannot be recalled from memory (Klein and Ford 2003; Schmidt and Spreng 1996; Seebach et al. 2011). The required effort regarding consumer information search can be described by a determining paradigm formulated by Stigler (1961). According to this, consumers optimize their search strategies until the marginal costs of purchasing the product are equal to the marginal benefits arising from acquiring the product. Klein and Ford (2003) stated that the invested search time is a critical factor for determining the cost of a search. However, with the increasing use of the Internet as a search medium, the influence of time can be disregarded owing to the fast accessibility of a great amount of information (Klein and Ford 2003).

In addition to the observation of the information search behavior of consumers, it is also important to get a deeper understanding of the public sentiment about a company’s products and services which can influence the purchase decision of individuals. Getting insights from personal convictions or emotional states is an ongoing task of human beings to support the own decision-making processes, or the understanding of valuations or emotions of somebody else (Liu 2010; Liu et al. 2010; Pang and Lee 2008). From a company point of view, product surveys and focus group studies are traditionally used to receive these kinds of insights. Besides the enormous effort in executing structured studies, there are substantial risks involved (e.g., missing trends, wrong classification of consumer groups, response bias) (Gamon et al. 2005). In line with the increased use of the Internet, new sources of information evolve which companies can use to extend the information basis of traditional market research to consider direct and unfiltered feedback from their customers. Hence, companies can enable decision-making by monitoring continuously opinions of the public about a product or brand. During the last decade, especially text-based information (e.g., blog content, product reviews) increased in an extraordinary way. Data-mining engines provide in-depth analyses to support forecast models. Even text analytic engines became a very precious tool to analyze large amounts of text to get valuable insights (e.g., the number of mention of a product in a review, sentiment about a brand) (Chaudhuri et al. 2011). Pang and Lee (2008) states that “sentiment-analysis technologies for extracting opinions from unstructured human-authored documents would be excellent tools for handling many business-intelligence tasks (…)”

Therefore, it seems useful to enrich decision-making processes by considering expressed emotions about a specific brand or product (Gamon et al. 2005). A study by Bae and Lee (2012) revealed that the change in positive and negative emotions in Twitter messages can be a suitable way to predict real-world events. They investigated the change in the positive-negative ratio of Twitter messages about Barack Obama and found out that these results were strongly correlated with respective job approval data (e.g., Gallup Daily). According to Bae and Lee (2012), the expressed sentiment of potential consumers toward brands or organizations could be used for forecasting models. Emotions toward a brand or product or even a general change in the market expectations could be detected and assessed with the help of sentiment analysis. Based on this, a deeper understanding of collective emotional trends is possible and could have predictive
power regarding social or economic factors (Bollen et al. 2011b; Saggion and Funk 2009). In this regard, data-mining engines and text-analytic engines are becoming more and more relevant (Chaudhuri et al. 2011). Text-based information (e.g., blogs, emails, reviews, etc.) as well as insights into consumer search behavior appear as promising and real-time sources (Chaudhuri et al. 2011). Decision making could be improved by including information about consumer behavior. Moreover, online data enables information delivery in a minimum time lag or even real-time. Hence, Seebach et al. (2011) suggested evaluating for combination possibilities of data to improve organizational sensing abilities like sales forecasting.

**Empirical Study: Sales Forecasting using Google- and Twitter-Data**

In our empirical study, we analyze how information about consumer information search behavior and consumer emotions can be used to forecast sales. We focus on the automobile industry which is under permanent worldwide competition. This implies continuous changes in the companies’ environments and thus, adequate responses by the companies are necessary (Seebach et al. 2011). Forecasting of sales could help companies to establish competitive advantage (Dalrymple 1975; Day 1994). The applied research approach, depicted in Figure 2, is based on Fayyad et al.’s (1996) process of knowledge discovery.

![Figure 2. Research Approach of the Empirical Study](image_url)

**Data Collection**

In the first step, we followed the approach of Requena-Silvente and Walker (2007) and used new vehicle registration volumes as substitute for car sales. We choose the German automobile market for our analysis as no other leading automobile market publishes comparable, reliable registration data. The German Federal Motor Transportation Authority (KBA) reports the brand-new vehicle registration volumes on a monthly basis. As this study presents a first analysis of our data, we focused on two of the largest German car manufactures, Mercedes-Benz and Volkswagen. Almost 30 percent of vehicles registered for the first time in Germany are manufactured by these companies. The new vehicle registration numbers for Mercedes-Benz and Volkswagen were collected from June 2012 to July 2013.

**Google Data**

As data source for consumer information search behavior we used data provided by the search engine Google. Search engines are heavily used as a source of information search prior to a purchase and present the regular starting point of a search. Moreover, selecting Google seems to be the most appropriate because Google has a market share of more than 90 percent in Germany (FT 2014). Google provides data about the search volume of a particular search term through its service GoogleTrends (GoogleTrends 2014). This publicly available web service creates detailed information about search queries for specific search strings. Furthermore, filtering predefined categories (e.g., automobile industry, vehicle brands) and different geographical regions is possible as well. Within the chosen time period, the maximum query share is assigned to the value 100 and the remaining queries are in relation to this (Choi and Varian 2012; GoogleTrends 2014). The returned number can be interpreted as the probability of a casual user to search for a certain term within a specific time period and geographical area (Seebach et al. 2011). Accordingly, we used GoogleTrends data to describe the consumers’ effort regarding the information search prior to making a car purchase. The search volume indices for Mercedes-Benz and Volkswagen were collected from June 2012 until July 2013.
Investigating Consumer Behavior to Improve Sales Forecasting

Twitter Data

As a basis for the consumer emotion data, we decided to use messages from Twitter. The popularity of Twitter is marked by the fact that the platform has almost 290 million monthly active users who send 500 million tweets each day (Twitter 2015). Twitter messages have an explanatory power regarding the identification of a market sentiment and enable the prediction of consumer behavior (Asur and Huberman 2010). We used Twitter’s application programming interface to collect tweets about the two car manufactures from June 2012 through July 2013. To investigate the German automobile market, it was necessary to make further language modifications during the data collection process. Therefore, the provided metadata of each tweet is used to select only German messages. After filtering the initial dataset of about 1.3 million tweets to exclude unrelated messages and spam, the final dataset contained about 235,000 tweets for further analysis.

Sentiment Analysis of Twitter Messages

To extract emotions from the Twitter messages, we applied an supervised sentiment analysis using the tool SentiStrength (Thelwall et al. 2012). The word list of SentiStrength already includes values of polarity for every word on a scale form “+5” (positive) to “-5” (negative). However, Twitter messages about a certain automobile manufacturer contain context specific vocabulary. Therefore, three researchers coded 5,000 random tweets to adjust polarities of the included word list to our research context. Thereby, we followed the process suggested by Morris (1994) for the manual coding which is presented in Figure 3. Afterwards, the sentiment analysis was conducted using the adjusted word list. The SentiStrength algorithm assigns each message both polarity values (positive/negative) which were separately summed up on a daily basis for further analysis. In total, more than 77,000 tweets contained emotional content whereas the positively rated amount of messages outweighed the negative ones. Tweets with equal positive and negative emotions (balanced) represented only a small share.

Data Transformation and Preliminary Analysis

Google and KBA data has different reporting granularity (monthly vs. daily) that makes it necessary to adjust the datasets to get the same periodicity (Seebach et al. 2011). Applying cubic spline interpolation (Wahba 1990), daily values were computed and added up to a monthly search amount. Both data types show strong seasonality. To avoid false results, deseasonalized values were computed (Chatfield 2001).

First, for every month $X_t$, the difference regarding the same month of the preceding year $X_{t-12}$ was calculated. The results are the deseasonalized values $\Delta_{12}X_t = (X_t - X_{t-12})$. The last step of the data
transformation process implies the improvement of comparability. In order to achieve this, year-on-year growth rates for the Google search volume indices and for the new vehicle registrations were computed by dividing $\Delta X_t$ by $X_{t-12}$. In the remainder of this paper the year-on-year growth rates of the Google search volume indices will be labeled as SVI and the growth rates of the new vehicle registration volumes as NVR.

\[
year_{on}\_year_t = \frac{X_t - X_{t-12}}{X_{t-12}}
\]

Next, the results of the sentiment analysis were also adjusted to meet the same periodicity as the KBA data. Besides the daily sum of positive and negative sentiments, the amount of positive and negative messages and the overall amount of all tweets were calculated by the sentiment analysis. Unfortunately, there were some data gaps because of server interruptions. Therefore, linear interpolation was used to estimate the missing values (Wahba 1990). Seasonality adjustments were not necessary because neither the amount of messages nor the distribution of negative and positive sentiment underlies a seasonal fluctuation. Within our dataset, more than 65 percent of all messages were not identified as positive or negative nor were they balanced. However, we assumed that even this share of data had a significant explanatory power and should be part of the empirical analysis. Thus, to compare sentiment values with the new vehicle registration volume, a ratio of the sum of positive sentiment divided by the amount of all posted messages within one day was computed. These daily values were summed up and a monthly average was created. Therefore, an overall change in emotion is measureable by considering the positive sentiment regarding all written messages divided by all other sentiment classifications (negative, neutral, and non-subjective). In the following, we will refer to this as positive sentiment measure (PSM).

<table>
<thead>
<tr>
<th>Car Manufacturer</th>
<th>Time Lag (in Month)</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercedes-Benz</td>
<td>6</td>
<td>0.66*</td>
</tr>
<tr>
<td>Volkswagen</td>
<td>3</td>
<td>0.5*</td>
</tr>
<tr>
<td>Notes: * p&lt;0.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Time Lag and Correlation Results SVI and NVR

Reducing this risk of purchasing a high-involvement product leads to a long-term information search which should take place noticeably before the purchase. The next step involves an investigation of the best combination of search terms and categories. Both data are time series and therefore a cross-correlation is adequate (Gruhl et al. 2005; Seebach et al. 2011) to detect the optimal time lag between search query and car registration. Furthermore, correlations between the SVI and NVR data were computed with a lag of one to six months. The best results, in case of both car producers, were found by using the brand names “Mercedes + Benz” and “VW + Volkswagen” within the Google category “vehicle shopping”. As seen in Table 1, a time lag of three to six months induced correlation coefficients between 0.50 and 0.66. All values are significant at the 0.05 p-level. Moreover, these results are in line with the findings of Seebach et al. (2011) and demonstrate a strong positive relation between the spent effort of consumers in information search prior to an automobile purchase.

<table>
<thead>
<tr>
<th>Car Manufacturer</th>
<th>Time Lag (in Month)</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercedes-Benz</td>
<td>0</td>
<td>0.48*</td>
</tr>
<tr>
<td>Volkswagen</td>
<td>1</td>
<td>0.55*</td>
</tr>
<tr>
<td>Notes: * p&lt;0.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Time Lag and Correlation Results PSM and NVR

Previous studies investigated the time difference between a change in sentiment and a certain response in sales or other events (Bae and Lee 2012; Gruhl et al. 2005). Very often the time lags are measured in terms of a couple of days. Therefore, it could be expected that the cross-correlation between the variable PSM and NVR is very small and does not exceed a lag of one month. However, the creation of a measurement with a smaller unit was not possible because the KBA data is reported monthly. As
expected, the calculated time lags were very small (see Table 2). In case of Mercedes, even a lag of zero could be identified with a corresponding correlation of 0.48. The PSM of Volkswagen led the NVR by one month, showing a correlation of 0.55. Therefore, there was a significant positive correlation between the sentiments of two car manufacturer and the corresponding new vehicle registration volume.

**Empirical Analysis**

Based on the preliminary analyses, we calculated three regression models: Model 1 (GoogleTrends), Model 2 (Twitter) and Model 3 (GoogleTrends + Twitter). All lagged variables for the SVI and for the PSM were included for an in-sample and out-of-sample analysis. For Model 1, a multiple regression model will be executed to get a suitable set of variables (Chatfield 2001; Crane and Crotty 1967; Shumway and Stoffer 2011). Therefore, for every lag from one to six months (i= 1,..6), a single new variable was created by shifting the SVI according to the lags and was used as independent variable to investigate the relationship with NVR as dependent variable.

Regression Model 1: \(\text{NVR}_{\text{mercedes/vw}}(t) = \alpha + \beta \text{SVI}_{\text{mercedes/vw}}(t) + \epsilon\)

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Mercedes-Benz</th>
<th>Volkswagen</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVI(_{t+1})</td>
<td>-</td>
<td>1.143**</td>
</tr>
<tr>
<td>SVI(_{t+2})</td>
<td>-</td>
<td>-0.957*</td>
</tr>
<tr>
<td>SVI(_{t+3})</td>
<td>-</td>
<td>1.466**</td>
</tr>
<tr>
<td>SVI(_{t+4})</td>
<td>-</td>
<td>-0.647*</td>
</tr>
<tr>
<td>SVI(_{t+5})</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SVI(_{t+6})</td>
<td>0.859*</td>
<td>-</td>
</tr>
</tbody>
</table>

\(R^2\) 0.441 0.823

*Notes: * p<0.05; ** p<0.01

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Mercedes-Benz</th>
<th>Volkswagen</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSM(_{t})</td>
<td>0.836*</td>
<td>-</td>
</tr>
<tr>
<td>PSM(_{t+1})</td>
<td>-</td>
<td>0.373*</td>
</tr>
</tbody>
</table>

\(R^2\) 0.232 0.307

*Notes: * p<0.05

Model 2 should reveal insights into the relation between sentiment and new registration volume. In this case, the independent variable for a one-month lag (PSM\(_{t-1}\)) was formed by shifting the dataset about one time period. The variable for the zero lag (PSM\(_{t}\)) did not need any adjustment. As reference value the NVR was used again as a dependent variable. Regarding the variable NVR, the transformation process led to a total loss of 18 observations. One observation was lost due to the generation of the new sentiment variables (PSM\(_{t}\), PSM\(_{t-1}\)). Thus, the final observation period for the regression analysis comprised 13 data points for each manufacturer.

Regression Model 2: \(\text{NVR}_{\text{mercedes/vw}}(t) = \alpha + \beta \text{PSM}_{\text{mercedes/vw}}(t) + \epsilon\)

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Mercedes-Benz</th>
<th>Volkswagen</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSM(_{t})</td>
<td>0.836*</td>
<td>-</td>
</tr>
<tr>
<td>PSM(_{t+1})</td>
<td>-</td>
<td>0.373*</td>
</tr>
</tbody>
</table>

\(R^2\) 0.232 0.307

*Notes: * p<0.05
Finally, Model 3 integrates both variables SVI and PSM to analyze the conjunction of consumer information search behavior and consumer sentiment.

**Regression Model 3:** $\text{NVRmercedes/vw}_t = \alpha + \beta_1 \text{SVImercedes/vw}_t - 1 + \beta_2 \text{PSMmercedes/vw}_t - 1 + \varepsilon$

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Mercedes-Benz</th>
<th>Volkswagen</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVI$_t$</td>
<td>-</td>
<td>1.074**</td>
</tr>
<tr>
<td>SVI$_{t-2}$</td>
<td>-</td>
<td>-0.966*</td>
</tr>
<tr>
<td>SVI$_{t-3}$</td>
<td>-</td>
<td>1.374**</td>
</tr>
<tr>
<td>SVI$_{t-4}$</td>
<td>-</td>
<td>-0.698*</td>
</tr>
<tr>
<td>SVI$_{t-5}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SVI$_{t-6}$</td>
<td>1.009**</td>
<td>-</td>
</tr>
<tr>
<td>PSM$_t$</td>
<td>0.661*</td>
<td>-</td>
</tr>
<tr>
<td>PSM$_{t-1}$</td>
<td>-</td>
<td>0.158*</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.564</td>
<td>0.855</td>
</tr>
</tbody>
</table>

*Notes:* *p<0.05; **p<0.01

**Table 5. Regression Results Model 3**

The results of the regression analyses were used to produce in-sample and out-of-sample predictions for a time horizon of one month. To gain deeper insights into the predictive accuracy of the developed regression model and to be able to compare the models, different error measures have to be determined. In line with the study of Seebach et al. (2011) and Hyndman et al. (2008), the mean absolute error (MAE) for all calculated predictions was used. Hence, actual sales for time $t$ ($A_t$) will be subtracted by the forecasted sales at the same time ($F_t$) and divided by the total amount of cases:

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |A_t - F_t|$$

The MAE will be applied to evaluate the regression. Moreover, we will compare the different regression models with each other in terms of prediction accuracy. Thus, the scale-dependent and easy-to-comprehend MAE seems to be the most suitable (Hyndman et al. 2008). Moreover, we were looking for a measurement more sensitive to outliers and decided on the root mean square error (RMSE). In that case, the sample standard deviation of the difference between the actual sales for time $t$ ($A_t$) and the forecasted sales for time $t$ ($F_t$) is used as error term:

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (A_t - F_t)^2}$$

The prediction of sales forecast should be a trigger for an improved organizational decision-making process. Following Armstrong and Collopy (1992), both chosen measures are useful in that case because they describe the dimension of the error and not just percentages as with relative measures.
Investigating Consumer Behavior to Improve Sales Forecasting

As presented in Table 6, the regression model for Google (Model 1) and Twitter data (Model 2) provided strong explanatory power. The conjunction of both data sources even lead to the highest $R^2$ values of 0.56 for Mercedes-Benz and 0.86 for Volkswagen within all models. The same took place for the error measures MAE and RMSE, which are on a very low level and reached their minimum in the third model. To enhance the statement about predictive ability, each regression model was transferred to enable out-of-sample predictions. The depicted values in Table 7 revealed that the prediction errors even of an out-of-sample analysis were within every regression model on a low level, thereby signaling a strong predictive power. Again, the minimum was reached by the third model where Google and Twitter data were combined.

**Table 6. Results of in-sample analyses**

<table>
<thead>
<tr>
<th>Producer</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercedes</td>
<td>0.441</td>
<td>0.232</td>
<td>0.564</td>
</tr>
<tr>
<td>VW</td>
<td>0.823</td>
<td>0.307</td>
<td>0.855</td>
</tr>
</tbody>
</table>

Notes: Bold values signal the lowest errors

<table>
<thead>
<tr>
<th>Producer</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercedes</td>
<td>0.048</td>
<td>0.068</td>
<td>0.041</td>
</tr>
<tr>
<td>VW</td>
<td>0.035</td>
<td>0.078</td>
<td>0.030</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Producer</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>0.057</td>
<td>0.083</td>
<td>0.050</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.044</td>
<td>0.095</td>
<td>0.040</td>
</tr>
</tbody>
</table>

Notes: Bold values signal the lowest errors

**Table 7. Results of out-of-sample analyses**

**Discussion of Empirical Results**

In the case of both tested car manufacturers, strong results of an in-sample and out-of-sample analysis were reached. Moreover, it was illustrated that the calculated sales forecasts showed low prediction errors and improves the results of a study by Seebach et al. (2011). In summary, it could be demonstrated that data about the online search behavior of potential consumers enables prediction of car sales. In the next step, the ability to forecast sales figures using information about consumer emotions was tested. The emotional state toward products could have predictive power as assumed by Saggion and Funk (2009) and Bollen et al. (2011b). In order to identify emotional trends, the sentiment algorithm SentiStrength was successfully adjusted to the automobile context. The second regression model provided significant results for both producers (Mercedes-Benz, Volkswagen). However, the values for the coefficient of determination were smaller compared to the first model. The in-sample and out-of-sample prediction errors suggested sufficient predictive accuracy regarding the forecast about car sales. Finally, the conjunction of search engine and microblogging data was investigated in model 3. In case of both car manufacturers all parameters were significant and the explanatory power, measured by $R^2$, reached a sufficient level. The same applies to the prediction error measures in case of an in-sample and out-of-sample analysis. It was demonstrated that the Google regression model could be improved by adding Twitter data as a second explanatory variable. Therefore, the combination of Google and Twitter data to forecast car sales was outperforming the individual consideration of the data sources.
Conclusion

The purpose of this paper was to investigate the sales forecasting ability of data about consumer information search behavior and consumer sentiment. In this regard, our study has two major theoretical contributions. First, we contribute to the literature by demonstrating the connection possibilities of consumer information search and consumer emotions. However, the latter demands more and more real-time and unstructured data to enable and improve management decision-making. This need is intensified by the increasing use of the Internet as a search and content sharing medium. It has been pointed out that data about consumer information search and insights into collective emotional trends to a brand are a valuable source to better recognize potential consumer behavior. The second theoretical contribution is based on the empirical analyses. It was demonstrated how Google search volume data and results of a sentiment analysis of Twitter messages are used to predict car sales. Within this process, a sentiment measure which represents the change of positive sentiment regarding all posted messages was developed. Contrary to other studies (Bae and Lee 2012; Lai 2010; Stieglitz and Dang-Xuan 2013), every written tweet, independent of the assigned sentiment, is included in the measure. Therefore, the explanatory power of negative, neutral, and non-subjective messages is not neglected and part of the sentiment calculation.

Furthermore, our study has implications for practice. The empirical analysis investigates different factors that describe the potential consumer behavior directly (e.g., Google search queries) or how it might be influenced indirectly (e.g., sentiment of Twitter messages). Therefore, organizations should use publicly available online data to improve their existing prediction models. Furthermore, companies should take into account the following opportunities. At first, the greater dependence on the Internet led to an increase in online search prior to a purchase (Klein and Ford 2003). Moreover, the executed measurement of the amount of specific search queries, the change over time, and capabilities of content creation should be used as well. Owing to this, companies should adapt, their marketing activities to provide suitable offline and especially online information to directly influence the search results of a consumer. Secondly, user-generated content serves as an information source during the search process and also reveals the general emotional state about a product or brand. Companies should use these functionalities, by sharing information to receive the best possible overview of the published enterprise-related content to influence decision-making of individuals.

This paper is exposed to some limitations which provide also opportunities for future research. First, the required transformation procedures (e.g., interpolation) of Google and Twitter data to estimate, for example missing values, will add bias to the empirical results. Second, there exist a lot of other sources of information in addition to the ones examined. Therefore, the investigation of online information search by gathering data of the primarily used search engine Google does not represent all search efforts of a potential consumer. The same applies for the received sentiment results of Twitter messages. Emotional expressions regarding a brand or product can also be extracted from consumer reviews (e.g., Amazon) or discussion forums. A further bias regarding the data selection was caused by the application of the KBA data as proxy for car sales. There is an unknown time gap between the actual purchase and the registration of the automobile. Third, we developed a new measure to express the change of positive sentiment regarding all written messages was developed. Further analyses are necessary to confirm the validity of this measure.

References


Investigating Consumer Behavior to Improve Sales Forecasting


