Abstract

Increasingly, customers are using Social media and other Internet-based applications such as review sites and discussion boards to voice their opinions and express their sentiments about brands. Such spontaneous and unsolicited customer feedback can provide brand managers with valuable insights about competing brands. There is a general consensus that listening and reacting to the "voice of the customer" is a vital component of brand management. However, the unstructured, qualitative, and textual nature of customer data that is obtained from customer’s poses significant challenges for data scientists and business analysts.

In this paper we propose a methodology that can help brand managers visualize the competitive structure of a market based on an analysis of customer perceptions and sentiments that are obtained from blogs, discussion boards, review sites, and other similar sources. The brand map is designed to graphically represent the association of product features with brands, thus helping brand managers assess a brand's "true" strengths and weaknesses based on the voice of customers. Our multi-stage methodology uses the principles of Topic Modelling and Sentiment Analysis in text mining. The results of text mining is analyzed to represent the differentiating attributes of each brand. We empirically demonstrate the utility of our methodology by using data collected from Edmunds.com – a popular review site for car buyers.

Introduction

Brand management is turning out to be very essential with growing competition, and reviews given by customers are turning out to become first impressions of a brand for several prospects. Keeping the growing importance of brand management and capacity of customer reviews, we chose this topic to elucidate an example on how reviews from various blogs and new sites can be used to know about the brand perception. The dataset compiled for this project serves as a foundation for additional research.

Data Extraction

The Edmunds.com web site allows consumers to post reviews on different car models. Consumers can rate their vehicle’s model, list pros and cons, and write their own review. In this paper we examine consumer reviews and ratings on four popular car brands named as Subaru, Chevrolet, Toyota and Honda. We extracted the data from Edmunds by accessing the Dealers API of Edmunds using the JSON package in Python. Using this API, we were able to fetch reviews and ratings provided by users on various car models based on brand, model name and year. To extract data, we built a tool that automatically iterates through all the model’s available for a car brand in a year. Using this tool, we scrapped 2176 text reviews along with numerical ratings for car models released in 2012 till 2016.
Data Cleaning & preparation
The raw extract data has reviews having stale information, we went through the data and filtered out such reviews. After initial text parsing, we came across terms which doesn’t give any insights, we removed such reviews and we were finally left with a dataset of 2176 review.

Methodology
The process that we followed to come up with the analysis further described is shown in the flow chart below.

Text Parsing & Text Filtering
We used Text Parsing node in SAS® Enterprise Miner to parse and understand the reviews. Through text parsing, we wanted to eliminate terms related to articles and connectors which have high frequency, and also terms that do not actually describe any feature in particular. For example, the stop list can include the names of different car models, so that topic extraction results disregard language use patterns that
revolve around specific car models and focus on high-level concepts. We created a list of such stop words and used it while parsing to exclude them from further analysis. Besides stop words, we used a custom dictionary to detect synonyms and a dictionary to detect multi-word terms. To suit to our analysis goals, we have ignored parts of speech like 'Abbr', 'Aux', ' Conj', 'Det', 'Interj', 'Num', 'Part', 'Pref', 'Prep', 'Pron', 'Prop', entities like 'Address','Currency','Date','Location','Measure','Percent','Phone','Timeperiod' and attributes like 'Punct', 'Mixed', 'Num'. The configurations of text parsing node are shown below.
The terms table output from Text Parsing node shows the frequency of occurrence of terms. The output of text parsing node above shows the term by frequency output, it is evident from the output that highly frequent terms again turned out to be articles and other words which doesn’t explain any special feature. Again terms with low frequency also don’t give much information as there are not many occurrences. So we used a Text Filter node with term weight as ‘Inverse Document Frequency’ instead of frequency to assign weights to the terms, and set minimum number of documents to 4, so that terms that appear in less than four documents in the collection are filtered out. The default setting in the text filter node filters out all the words with low weight. We have also enabled spell check property to accommodate any typo errors, we used custom English dictionary to accommodate spell checks. The text filter node generates term by document matrix, viewing results in interactive filter mode will enable us to explore further using concept links.

**Text Topic**

The Text Topic node enables you to create topics of interest from a list of terms. Text Topic node extracts topics from text documents. A document can have multiple topics. Parameters of the text topic node can be manually set based on the size of the data and the frequency of unique terms. For this project the numbers of terms in a topic are limited to 15. Below term topic matrix shows us the distribution of various terms together as topics.

<table>
<thead>
<tr>
<th>Topic</th>
<th>No. of Terms</th>
<th>No. of Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comfort</td>
<td>37</td>
<td>756</td>
</tr>
<tr>
<td>Interior</td>
<td>23</td>
<td>685</td>
</tr>
<tr>
<td>Performance</td>
<td>55</td>
<td>1138</td>
</tr>
<tr>
<td>Safety</td>
<td>29</td>
<td>483</td>
</tr>
<tr>
<td>Technology</td>
<td>24</td>
<td>554</td>
</tr>
</tbody>
</table>

**Fig6. Custom Topics**
Custom Topic Analysis

We further analyzed each of these topics by brand to find the proportion of customers talking about these attributes by brand. The bar chart below shows the percentage of customers talking about each of the topics by brand.

![Bar chart showing the percentage of documents within each brand for comfort, interior, performance, safety, and technology. Each brand is represented with different colors: Chevrolet (blue), Honda (orange), Subaru (gray), and Toyota (yellow).]
Brand wise regression analysis

We further analyzed to see the sentiment of customers towards the brands in each of these areas. Consumers have provided ratings on a scale of 1-5 on categories such as Performance, Comfort, Technology, Interior, and Safety. Basically, numerical ratings provided by consumers are having maximum 25 percent missing values. By understanding the data, we have considered median value within each brand to impute missing values for a brand.

We have categorized all the 'reviews which contains any of the terms mentioned under each of those topics and with the corresponding topic rating less than three as negative and all the reviews that contain any of the terms related to the custom topics discussed above and the respective topic rating greater than three as positive. We performed a regression analysis to confirm the same.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Binary Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Negative</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Positive</td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Fig9. Brand Sentiments
Fig 10. Regression results for Chevrolet brand

The top two important predictor variables based on LogWorth value for this brand are Performance Rating and Comfort Rating. About 77% of variance is explained by this model.

Fig 11. Regression results for Honda brand

The top two important predictor variables based on LogWorth value for this brand are Comfort Rating and Technology Rating. About 65% of variance is explained by this model.

Fig 12. Regression results for Toyota brand
The top two important predictor variables based on LogWorth value for this brand are Performance Rating and Technology Rating. About 72% of variance is explained by this model.

![Parameter Estimates](image)

**Fig13. Regression results for Subaru brand**

The top two important predictor variables based on LogWorth value for this brand are Performance Rating and Interior Rating. About 76% of variance is explained by this model.

**Conclusion**

We would like to conclude by summarizing the results obtained above and insights derived from the same. A methodology to facilitate brand management using textual data on brand sentiments is developed. This exercise reveals the strengths and weaknesses of brands and provides valuable diagnostic information. Subaru has the fewest reviews. Highest dissonance levels in terms of interior, performance and technology categories. Extreme positive consensus about safety. Niche marketing is recommended. Chevrolet has the highest reviews, moderate dissonance on all factors. Mass marketing is seen.

**Limitations and Future Research**

In this paper we only focused on four car brands and we identified brand sentiments using customer ratings on various aspects. In our future work we are planning to do sentiment mining on the text reviews to compare and contrast the results obtained from numeric ratings. We also would like to continue our research to see if brands can be positioned based on brand sentiment.

**References**


CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the authors at:

Praveen Kumar Kotekal
Email: praveen.kotekal@okstate.edu
MS in Business Analytics Program
Oklahoma State University, Stillwater, OK

Dr. Amit Ghosh, Cleveland State University
Email: A.GHOSH@csuohio.edu
Amit K. Ghosh is Chair & Professor of Marketing Department.

Dr. Goutam Chakraborty, Oklahoma State University, Stillwater OK
Email: goutam.chakraborty@okstate.edu
Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU business analytics certificate at Oklahoma State University.

SAS® and all other SAS® Institute Inc. product or service names are registered trademarks or trademarks of SAS® Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.