# How Text Mining Approach can be Used to Understand Automobile Purchase

Shaoqiong Zhao[1]

[1] School of Business, Accounting, and Economics, Carroll University, United States

Correspondence: Shaoqiong Zhao, School of Business, Accounting, and Economics, Carroll University 100 N. East Ave, Waukesha, WI 53186, United States.

Received: February 3, 2018 Accepted: March 26, 2018 Online Published: March 27, 2018

doi:10.5430/jbar.v7n1p43 URL: https://doi.org/10.5430/jbar.v7n1p43

**Abstract**

Word of mouth has long been recognized to be an influential variable in marketing. With the growth of Internet applications, traditional word of mouth has evolved into the online form where individuals spread their perceptions via the written word. With the rapid growth of comments by consumers over the Internet, in-depth purchasing related information is available to markers. In this paper we try to extract the essence of consumers' attitudes from the online reviews posted on kbb.com through text-mining approach, which is the most popular and highly visited website in automobile industry. Thus we can identify the key features that are related to the prediction of positive/negative overall attitudes of online users. Then with the diagnostic of these identified key features through Gini indexing, they can be used to help to marketers in designing their keyword choices for more effective application of search engine marketing strategies for positive associated key features while identification of the negative associated key words will lead to discovery of problematic areas.

**Keywords:** online review, text-mining, key features, prediction, diagnostic, application

## 1. Introduction

It is widely known that user recommendations play a role in affecting potential purchasers of products and services. (Senecal S, Nantel J., 2004) found that individuals who consulted online product recommendations selected recommended products twice as often as individuals who did not consult recommendations. This is one reason why user reviews are prominent in numerous web sites (e.g., at Amazon.com). With the rapid growth of comments by consumers over the Internet, in-depth purchasing related information is available to markers. The wide availability of lengthy and numerous text-based online reviews provides a treasure trove of information that can potentially reveal a much wider set of variables that determine whether a recommendation is made or not. With the large amount of information, data-mining methodologies are needed to uncover the hidden information in order to discover the patterns in customer behavior (Sebastiani F., 2002) (Lee SJ, Siau K., 2001) (Hoontrakul P, Sahadev S., 2008).

However, as there has been ample research on text mining, they aim at improving the methodology it self a lot like classification performance of text classifier, feature selection methods, etc. (Pang B, Lee L, Vaithyanathan S., 2002) (Bast E, Kuzey C, Delen D., 2005). In the literature of marketing, (Ghose A, Li B, Ipeirotis P., 2012) applied text mining approach to complement numerical variables to help predict products sales; (Mostafa M., 2013) evaluated consumers' sentiments toward well-know brands using tweets. These researches focus on the prediction power of the words itself and are missing the diagnostic aspect of using reviews, which is of important use to marketers. The identification of these key features can be used to help to marketers in designing their keyword choices for more effective application of search engine marketing strategies for positive associated key features while identification of the negative associated key words will lead to discovery of problematic areas.

In our research we extract the most relevant information from online text reviews and then we conduct pre-processing and indexing in order to get the data formatted for classification analysis. The dependent variable is whether users recommended the product or not while the independent variables in our predictive model were words from online reviews. Then we use Gini-index to help us identify the most relevant features of the products that drive the overall attitude of the consumers.

The rest of the paper is organized as follows: in section 2 we discuss the fundamentals of our text modeling methodology and Gini-index. In section 3 we discuss the details how we implement this for online reviews and

recommendations. In section 4 we analyze the results we get and provide the insights of our discussions. Section 5 provides a summary of our findings and suggests future directions for research.

## 2. Method

In this section we describe the overall approach that we use for analysis of text content. Text classification will use a machine-learning algorithm to classify the word based text documents into one of previous defined categories (Sebastiani F., 2002). In the following sections, we will explain the complete process of how we used text mining.

### *2.1 Preprocessing*

Before a learning method can be applied, several preprocessing steps are required to get the data in ready format for further analysis. First, the raw text is divided into tokens (single word, special symbols, etc.). The second step is case conversion where the words are modified to be all in lower cases. The third step is removal of stop-words. The purpose of stop-words removal is to reduce the size of the classification matrix by reducing the number of irrelevant terms. Lots of very commonly used words like *"the", "I", "to"*, etc., are of little use in classifying documents into predefined categories. Lastly, different variations of a word are converted into a single common form that is termed stem.

### *2.2 Indexing*

The result so far is a term-by-document matrix with each cell representing the raw frequencies of occurrence for each term in each document. The columns of the matrix represent terms (words), and the rows represent documents (reviews for example). (Jones KS., 1972) showed that there is a significant improvement in retrieval performance by using weighted terms vectors. The term weight is generated by multiplying Term Frequency (TF) and the Inverse Document Frequency (IDF) (Jones KS., 1973) (Coussement K, Van D., 2008) (Salton G, Buckley C., 1988).

### *2.3 Classification Technique*

We use the Support Vector Machine (SVM) approach for classification purposes. SVM was invented by Vapnik and Chervonenkis (Vapnik V, & Chervonenkis A., 1964) and has been used a lot in various areas (Pang B, Lee L, Vaithyanathan S., 2002) (Bast E, Kuzey C, Delen D., 2015). SVM are supervised learning models that can classify data into the groups. There are various algorithms used for SVM classification models and the most popular one is the Sequential Minimal Optimization (SMO), which is conceptually simple, easy to implement and fast to compute (Cristianini N, Shawe-Taylor J., 2002).

### *2.4 Evaluation Criteria*

For evaluating the performance of different classification models, we used Area Under the receiving operating Curve (AUC) (Coussement K,Van D., 2008) (Powers DM., 2007).

**AUC** (Metz CE., 1978): In order the get the AUC, we need to first draw the Receiver Operating Characteristics (ROC) curve. ROC curve illustrates the diagnostic ability of a binary classifier as its discrimination threshold varies. ROC considers the sensitivity (true positive rate) and 1-specificity (false negative rate) in a two-dimensional graph (Coussement K,Van D., 2008). The sensitivity is the likelihood of identifying a positive case when presented with one while the specificity is the likelihood of identifying a negative case when presented with one (Gopal K, Sacchettini JC, Ioerger TR., 2007). A ROC curve depicts relative trade-offs between sensitivity and 1- specificity. The area under this ROC curve is calculated to compare the performance of a binary classifier (Hanley JA, McNeil BJ., 1982). When classifying randomly, the ROC curve is a line joining points (0, 0) and (1, 1) with the area under the curve equals 0.5. In general, any classification performance should be better than a randomly made classification.

### *2.5 Gini Index*

Gini index was proposed and sudied by Aggarwal & Chen (Aggarwal CC, Chen C, Han JW., 2010). It aims to decide which feature variables are decision variables for a decision support appication. In the training data the key decision variables are identified and trained to predict the decisions classes. Training dataset $D_{trian}$ contains n reviews and each review q belong to a predefiend class with lables s which is drawn from the set {1…k}. Overall we have a dxn feature-review matrix with each feature is denoted i with i range from 1 to d and each review is denoted by q with q range from 1 to n. In our case since the lables will be a binary situation of reconmend or not. Now the Gini index is calculated to define the level of class dscrimination among the data points of each feature as follows:

$$G(i) = \sum_{s=1} f(i,q,s)^2$$

Then we can use gini index to help us find the key features that are important to the decisions. With a bigger Gini index, it indicates a higher discriminating abillity of that word.

## 3. Results

In our paper, we tried to first to predict whether consumers will recommend or not of a car as a binary classification problem using only text data (reviews posted for the car) from consumers on kbb.com. The good prediction performance indicates that the written reviews are a good flection of consumers' thoughts toward the cars. Then we tried to extract top key features with recommendation or not from the accurately predicted models for managerial implications using Gini index.

### *3.1 Data Collection*

In our study, we use data obtained from kbb.com, which is a leading website in automobile industry. We collect the data of three auto models from three carmakers: Ford Focus, Nissan Sentra, and Toyota Avalon. These three cars are chosen because they are very popular models and make in the US market and attracts a good number of reviews. Also they belong to the same strategic group with the same target market, which makes sense to compare and contrast them.

### *3.2 Empirical Analysis*

The first part of the analysis is the prediction of the positive/negative polarity of the reviews as indicated as recommend or not. We use overall ratings to classify the reviews into recommend (9 and 10 out of 10) or not-recommend (below 9) as our dependent variable and form this binary text classification problem. We follow the steps described in section 2. A common 10-folds cross validation for classification testing and prediction is used to avoid over-fitting and also train-test purpose. The classification model confirms our expectation that there is information content in the reviews that can help predict consumers' overall attitude toward the cars: whether the consumers recommend it or not.

The second part of the analysis is a key focus of our analysis: diagnostics of the key features. We used Gini index (discriminant) and the frequency (important) together to identify the top features associated with positive/negative attitudes.

#### 3.2.1 Classification Performance

In this section, we report the performance results of online reviews' classification models.

Table 1. Prediction results of Bally, TI and Venetian (Accuracy)

| Automobile | Ford Focus | Nissan Sentra | Toyota Avalon |
|---|---|---|---|
| ROC | .751 | .704 | .695 |

As shown in Table 1, the predictive performances of the three cars using ROC are reported. All the ROCs are better than random classification performance 0.5 indicating a good prediction.

Based on accurate prediction, we can do further analysis of key words and the applications in the following sections. Prediction itself also has managerial applications. Especially on the aggregate level, the over all tends of recommending or not can indicate the changing directions of the performance of the company so company can make some arrangements accordingly.

#### 3.3.2 Diagnostics

The second part of the result focuses on the diagnostic use of the text mining methodology and Gini index analysis.

For each feature, we calculated the Gini index of that feature and select only the ones with a Gini value higher then 0.75 (Metz CE., 1978) and also frequency is higher than the average frequency of the words appearence.

Table 2. Gini index selection

| Ford Focus | Nissan Sentra | Toyota Avalon |
|---|---|---|
| voice (p) | efficiency (P) | roomy (p) |
| junk (n) | room (p) | matainance (p) |
| navigation (p) | family (p) | acceleration (p) |
| fix (n) | touch (n) | fix (n) |
| command (p) | gear (n) | power (p) |
| stereo (p) | camera (p) | size (p) |
| sport (p) | leg (p) | sport (p) |
| package (p) | horsepower (p) | dealership (p) |
| mountain (p) | engine (n) | fuel (p) |
| suspension (p) | button (p) | look (p) |
| leather (p) | freeway (n) | hybrid (p) |
| recall (n) | city (n) | camera (p) |
| repair (n) | music (n) | cruise (p) |
| warranty (n) | performance (n) | engine (p) |
| service (n) | Style (p) | technology (p) |
| friend (p) | dealer (n) | brake (p) |
| dealership (n) | brand (p) | cabin (p) |
| performance(p) | fuel (p) | family (p) |
| family (p) | mile (p) | leg (p) |
| trade (n) | power (n) | style (p) |
| color (p) | highway (p) | price (p) |
| convenience (p) | base (p) | luxury (p) |
| clutch (n) | design (p) | space (p) |
| shudder (n) | wife (p) | mile (p) |
| maintenance (p) | | radar (p) |
| noise(n) | | vibration (n) |
| economy (p) | | comfort (p) |
| hatchback (p) | | option (p) |
| Option (p) | | dash (p) |
| camera (p) | | seat (p) |
| upgrade (p) | | sound (p) |
| size (p) | | trunk (p) |
| titanium (p) | | wife (p) |
| computer (p) | | color (p) |
| city (p) | | console (n) |
| design (n) | | panel (n) |
| toyota (n) | | transmission (p) |
| nissan (n) | | |
| price (p) | | |
| highway (p) | | |
| brand (n) | | |

**p: positive**

**n: negative**

As shown in table 2, we can identify the features strongly associated with recommending to others. These can really help us identify what aspects the car is doing well and valued by the consumers so they can be used as ad-words for online advertising. We can also identify the features strongly associated with not recommending, which can lead to the problematic aspects identification.

For Ford Focus, overall brand image is negative with problems is repair/fix/warranty/service/dealership and low trade-in value. Consumers do like some of the car features like navigation systems (command, voice, etc.) and the low price (economy). For Nissan Sentra, engine/power seems to be a downside while other accessory features (design, style, camera, etc.) are valued as positive and the fuel efficiency. For Toyota Avalon, overall very positive attitudes of the consumers toward various aspects of the car except a few problems of vibration and the console part.

## 4. Discussion

In the previous section, we empirically show that the text mining models can classify the users' recommendations for the cars very well. What the consumers put on websites can truly represent their real thoughts about their experience with the cars. Additionally we identified those top features, which are really important from the viewpoint of providing diagnostic information to the companies. These features positive or negative can both be of use to managers from different angles. Especially for the negative ones, they lead the direction of problematic areas. For Ford they should really work on to improve the reliability of the car as well as the perception of the consumers towards the car. For Nissan Sentra, the biggest problem sits on the engine, which can become a very critical issue. For Toyota Avalon, the design of the console should be looked and this should be an easier area to improve.

Now we would like to compare the three cars to see the difference across the three cars.

Table 3. Comparisons of 3 automobiles (Gini index based)

| Ford Focus | Nissan Sentra | Toyota Avalon |
|---|---|---|
| **Performance** (performance) | **Performance** (horsepower, engine (n), performance (n), power (n)) | **Performance** (power, engine, vibration (n), transmission) |
| **Quality** (voice, navigation, command, stereo, package, suspension, leather, service (n), color, convenience, clutch (n), shudder (n), noise (n), option, camera, upgrade, size, titanium, computer, design (n)) | **Quality** (room, touch (n), gear (n), camera, leg, button, music (n), base, design) | **Quality** (roomy, acceleration, size, look, camera, cruise, technology, brake, cabin, leg, space, radar, option, dash, seat, sound, trunk, color, console, panel) |
| **Reliability** (fix (n), recall (n), repair (n), warranty (n), dealership (n), Maintenance,) | **Reliability** (dealer) | **Reliability** (Maintenance, fix (n), dealership) |
| **Styling** (sport, mountain, friend, family, hatchback, city, highway) | **Styling** (family, freeway (n), city (n), style, highway, wife) | **Styling** (sport, family, style, wife,) |
| **Value** (junk (n), trade (n), value, economy, toyota (n), nissan (n), price, brand (n)) | **Value** (efficiency, brand, fuel, mile,) | **Value** (fuel, hybrid, price, luxury, mile) |
| **Comfort** () | **Comfort** () | **Comfort** (comfort) |

**n: negative**

**otherwise positive**

From table 3, we can see, these three hotels have lots of important features in common fall into the following aspects including performance, quality, reliability, styling, value and comfort. However among the common categories, there are same as well as unique features to each car. For the same features listed, then ranking is not the same either really indicates the attitudes differences across the three cars. Lots of product features are indicated under quality with unique features to each car. For Ford Focus, there are several negative features including service, clutch, shudder, noise and design, which should draw the manager's attention. For Nissan Sentra, it is the tough, gear and music that are problematic. While for Toyota Avalon, all features are listed as positive. Engine and powers seems to be a problem for Nissan Sentra while vibration is problematic for Toyota Avalon. Nissan Sentra has the best reliability comparing to the other two both has negative features related to reliability. Dealership is also a downside for Ford.

All three cars are considered as for the family while Nissan and Toyota particularly indicated the wife preference. From consumers' review Nissan and Toyota are seen as major competitors to Ford while Ford's brand image lean more towards to the negative comparing with the other two brands. The two Japanese brands are seen as efficient brands are consistent with the long history. Ford is catching up on this but the trade value of the car seems still low. Toyota is the only one identified as comfortable by consumers out the three brands studied.

The comparisons of the three major economy cars in the automobile industry indicated the current brand positioning of the three brands. Ford is still considered less reliable comparing to the two Japanese cars with Toyota seen as the leader. But we do see a small change in the minds of the consumers of improved performance and maintenance of Ford. Ford has spent resource in improving this and it worked. So they should continue on doing so. The internal features are positively valued by consumers and can be used as key words of advertising. Nissan has a big problem is engine/performance. This is critical for automobiles. Nissan should really pay attention to this area and started thinking on how to solve this issue before it is too late. Toyota is still recognized as the leader in this market and rated well on most aspects. One particular standing-out feature different from the other two is comfort. As now it is becoming harder and harder to differentiate and break the clutter, this feature might shine as the unique value proposition.

## 5. Conclusion

Online reviews of products and services are present all over the Internet. Potential consumers value these greatly. Marketers can also get valuable information from reading these reviews. These reviews predominantly contain text-based information. In our present research we utilize text-mining methodology to show that consumers' attitudes can be accurately predicted by text mining.

In addition to making predictions of recommendations, marketers would benefit tremendously if they can identify key words from many thousands of reviews; we suggest a framework by which companies can get this important *diagnostic* information. This framework consists of reliance on the importance of words based on frequency of occurrence and a new way to look at how certain words have greater power to discriminate/distinguish between existence and non-existence of recommendations (Gini index). Words identified by this diagnostic approach will be of use to advertising managers when they plan on designing messages appropriate for search engine advertising as in Google Adwords; a single ad here can use only a small number of words, and the choice of the keywords could become crucial from the viewpoint of revenue generation.

While the contribution of this research is clear, there is still limitation of it. In this research we performed diagnostic analysis using online reviews and suggested managerial applications of it. It would be more beneficial if we can actually test the using of the key words identified from the research can improve the advertising effect. Also we compared the three car brands. It will be more beneficial we can use the identified features to construct the positioning map of the three brands.

The potential future directions for this research stream are numerous. The overall methodology designed in this paper is a foundation that can be applied to a variety of marketing situations. In today's digital era consumers freely express their opinions about products and services on many websites. This provides numerous information sources that can help academicians and practitioners in analyzing consumer attitudes. We can extend this methodology to study a tremendous variety of research questions that would benefit from the analysis of text content posted by web users all over the Internet. Advertisers and marketers would be among the prime beneficiaries once they can glean the appropriate information from text-based reviews.

## References

Aggarwal CC, Chen C, Han JW. (2010). The inverse classification problem. *Journal of Computer Science and Technology, 2010 May, 25*(3), 458-468. https://doi.org/10.1007/s11390-010-9337-x

Bast E, Kuzey C, Delen D. (2005). Analyzing initial public offerings' short-term performance using decision trees and SVMs. *Decision Support Systems, 2015 May; 73*, 15-27. https://doi.org/10.1016/j.dss.2015.02.011

Coussement K, Van D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management, 2008 Apr, 45*(3), 164-174. https://doi.org/10.1016/j.im.2008.01.005

Coussement K,Van D. (2008). Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems, 2008 Mar, 44*(4), 870-882. http://dx.doi.org/10.1016/j.im.2007.10.010

Cristianini N, Shawe-Taylor J. (2002). *An Introduction to Support Vector Machines*. Cambridge: Cambridge University Press.

Ghose A, Li B, Ipeirotis P. (2012). Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science, 2012 Apr; 31*(3), 493-520. https://doi.org/10.1287/mksc.1110.0700

Gopal K, Sacchettini JC, Ioerger TR. (2007). Database approaches and data representation in structural bioinformatics. Proceedings of the 7th IEEE International Conference on IEEE; 2007 Oct 14-17; Bioinformatics and Bioengineering, Boston, MA, USA. 2007. 425p. https://doi.org/10.1109/BIBE.2007.4375597

Hanley JA, McNeil BJ. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology. 1982 Apr; 143*(1), 29-36. https://doi.org/10.1148/radiology.143.1.7063747

Hoontrakul P, Sahadev S. (2008). Application of data mining techniques in the on-line travel industry: A case study from Thailand. *Marketing Intelligence & Planning, 26*(1), 60-76. https://doi.org/10.1108/02634500810847156

Jones KS. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation, 28*(1), 11-21. https://doi.org/10.1108/eb026526

Jones KS. (1973). Index term weighting. *Information Storage and Retrieval, 1973 Nov, 9*(11), 619-633. https://doi.org/10.1016/0020-0271(73)90043-0

Lee SJ, Siau K. (2001). A review of data mining techniques. *Industrial Management & Data Systems, 101*(1), 41-46. https://doi.org/10.1108/02635570110365989

Metz CE. (1978). Basic principles of ROC analysis. *Seminars in nuclear medicine, 8*(4), 283-298. https://doi.org/10.1016/S0001-2998(78)80014-2

Mostafa M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 2013 Aug, 40*(10), 4241-4251. https://doi.org/10.1016/j.eswa.2013.01.019

Pang B, Lee L, Vaithyanathan S. (2002). Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing; 2002 Jul 79-86; Association for Computational Linguistics, Stroudsburg, PA, USA. (vol. 10). https://doi.org/10.3115/1118693.1118704

Powers DM. Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63.

Salton G, Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

Sebastiani F. (2002). Machine learning in automated text categorization. *Journal of the ACM Computing Surveys, 2002 Mar, 34*(1), 1-47. https://doi.org/10.1145/505282.505283

Senecal S, Nantel J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing, 80*(2), 159–169. https://doi.org/10.1016/j.jretai.2004.04.001

Vapnik V, & Chervonenkis A. (1964). A note on a class of perceptrons. *Automation and Remote Control, 1925*, 103-109.