MINE YOUR OWN BUSINESS: MARKET STRUCTURE SURVEILLANCE THROUGH TEXT MINING

Ronen Feldman*
Hebrew University

Jacob Goldenberg
Hebrew University

Oded Netzer
Columbia University

January 2010

* Ronen Feldman is Associate Professor of Information Systems, Jerusalem School of Business Administration, Hebrew University (Ronen.Feldman@huji.ac.il). Jacob Goldenberg is Professor of Marketing, Jerusalem School of Business Administration, Hebrew University (msgolden@huji.ac.il). Oded Netzer is Associate Professor of Marketing, Graduate School of Business, Columbia University (on2110@columbia.edu). The authors thank the Marketing Science Institute and the Wharton Interactive Media initiative for financial support. The authors thank Moshe Fresko, Don Lehmann, Olivier Toubia, and seminar participants at Columbia University for comments on earlier versions of this research.

Mine Your Own Business: Market Structure Surveillance through Text Mining

Ronen Feldman, Jacob Goldenberg, and Oded Netzer

ABSTRACT

Web 2.0 provides gathering places for internet users in blogs, forums, and chat-rooms. These gathering places leave footprints in the form of colossal amounts of data. These data include consumers' thoughts, beliefs, experiences, and even interactions. In this paper, we propose an approach to transform the Web 2.0 to a large, yet readily available, marketing field test. Exploring such online user-generated content offers the firm an opportunity to "listen" to consumers in the market in general and to its own customers in particular. By observing what customers write about the products in the category, the firm can get a better understanding of the market structure, the competitive landscape, and the features discussed about its and the competition's products. The difficulty in obtaining such insights is that consumers' postings are often not easy to syndicate. A decoding mechanism is needed in order to transform these raw qualitative data into meaningful knowledge. To address these issues, we developed an advanced text mining approach (called CARE) and combine it with semantic network analysis tools. We demonstrate this approach using two cases—sedan cars and diabetes drugs—generating sensible perceptual maps and meaningful insights, without interviewing a single consumer.

1. INTRODUCTION

Guadagni and Little's (1983) seminal paper has paved the way to a radical change in the marketing science community, in part by exposing the field to the potential in utilizing rich scanner panel data to study consumer and firm behavior. Indeed, following their paper, over two decades of marketing research have led to investigating consumer purchases of coffee, ketchup, tuna, and other grocery products, leading to important methodological and substantive findings. In recent years, marketing scientists have obtained additional new and original datasets, such as click-stream data, contractual-setting customer relationship datasets, and pharmaceutical datasets, leading to new insights and advances. However, academic researchers and marketing practitioners often feel consumer data are still hard to come by, and that the available data sets are often sparse and narrow. In this research, we argue and demonstrate that the marketing field has underutilized direct access to a massive and rich consumer dataset that constitutes a world unto itself: the World Wide Web.

The spread of the Internet has led to a colossal quantity of information posted by consumers on the Web through media such as forums, blogs, and product reviews. This type of consumer-generated content offers the firm an opportunity to "listen in" on consumers in the market in general and to its own customers in particular (Urban and Hauser 2004). By observing what consumers write about products in a category, the firm could, in principle, gain a better understanding of marketing opportunities, the market structure, the competitive landscape, and the features discussed about its own and its competitors' products.

So why haven't we seen an explosion of marketing research tapping into this abundant supply of data? Consumer-generated content on the Web is both a blessing and a curse. The wealth of data presents several difficulties: First, the amount of data provided is overwhelmingly large, making the information difficult to track and quantify. Second, this rich, yet unstructured, set of consumer data is primarily qualitative in nature (much like data that can be elicited from focus groups or depth interviews but on a much larger scale), making it noisy—so much so that it has been nearly impractical to quantify and convert the data into usable information and knowledge. Third, the language consumers use on the Web is often informal and does not adhere to standard grammatical rules, making the task of syndicating such data difficult. In this paper, we propose to use a combination of *text-mining apparatus* and a *network analysis framework* to overcome these difficulties.

Our objective is to open a window into large-scale exploratory consumer data posted on the Web, which will allow firms and researchers to "listen to," collect, and distill consumer voices in the marketplace. We first mine these exploratory data and then convert them into quantifiable information. Due to the complexity involved in consumer forum mining, we developed a text-mining apparatus especially suited for text mining of consumer forums. We combine an automatic conditional random field approach with manually crafted rules, resulting in the proposed CRF Assisted Relation Extraction (CARE) apparatus. The CARE approach provides a methodological contribution over the handful of text-mining applications appearing in the marketing literature, as it goes beyond merely summarizing the quantity and valance of the textual information and toward understanding relationships between products and mapping the discussion itself. We use network analysis techniques to convert the text-mined data into a semantic network, which can in turn inform the firm, or the researcher, about the market structure and some meaningful relationships therein.

In what follows, we describe the current state of research with respect to studying and utilizing online consumer-generated content and the applications of text mining in business and marketing. In section 3, we describe the proposed text-mining methodology. In sections 4 and 5, we demonstrate the use of the text-mining approach in the context of mining a sedan cars forum and pharmaceutical drugs forums. We conclude with a discussion of the potential and the limitations of the current approach and directions for future research.

2. MINING AND UTILIZING CONSUMER-GENERATED CONTENT

2.1 Consumer Forums and Blogs

One can think of consumer-generated content in venues such as forums and blogs as an online channel for word of mouth, which is one of the marketing operationalizations of the somewhat broader concept of social interactions. A wide range of academic papers, industry market research, and anecdotal evidence point to the significant effect of word of mouth on consumer behavior and, in turn, on sales (e.g., Eliashberg et al. 2000, Reichheld and Teal 1996).

Online word of mouth, often known as "internet word of mouth" or "word of mouse," enables consumers to communicate quickly with relative ease. Numerous cyberspaces such as chat rooms, product reviews, blogs, and brand communities invite and encourage consumers to post their ideas, views, and reviews. The level of activity in these channels of communication has grown exponentially in recent years. To gain a concept of the magnitude of the activity in these cyberspaces, in 2008 there were approximately 1.6 million blog postings per day, about double the number in 2007 (Sifry 2008). Eleven percent of U.S. adults participate in message boards or internet forums, and 6 % post product ratings or reviews at least once a month (Rousseau-Anderson 2008).

Firms take differing roles in Internet word-of-mouth activity: companies might participate in viral marketing initiatives as moderators, mediators, or even active participants to trigger the word-of-mouth process and accelerate its distribution (Oberndorf 2000). Active participation of the firm in the creation and management of word of mouth was examined theoretically by Mayzlin (2006), and empirically by Godes and Mayzlin (2009). In this paper, we propose a more passive form of firm participation, and suggest a platform for firms to tap into Internet word of mouth by aggregating the wealth of information posted in online forums and blogs.

Forums and blogs are places where consumers post their opinions, views, and recommendations without knowing whether anyone will be influenced by or even read their posts.

This communication channel shrinks the geographic barriers and time lags often associated with offline word of mouth (Dellarocas 2006, Schindler and Bickart 2005).

In the past few years, academics and practitioners have begun to realize the potential in online consumer forums, blogs, and product reviews. Several studies have investigated the relationship between consumer-generated content and sales. One of the main difficulties in utilizing consumer-generated content for quantitative analysis is that the data are primarily qualitative in nature. Indeed, in their conceptual paper on future directions for social interaction research, Godes et al. (2005) state that one of the difficulties in tapping into user-generated content is the inability to analyze the communication content. One way to overcome this difficulty is to use a controlled lab environment where the topic of the discussion is manipulated (e.g., Verlegh et al. 2004). A more common approach has been to use moments of consumer-generated data, such as magnitude or valence, to represent the discussion. Alternatively, quantitative summaries of content, such as overall product ratings, can be used to represent the content of consumers' opinions. For example, Liu (2006) examined the volume and valence of messages posted on Yahoo Movies Message Board to predict box office sales. Due to limited text-mining capabilities, Liu reported that mechanically analyzing more than 12,000 movie review messages using human reviewers was "an extremely tedious task." Similarly, in studying the effect of volume and dispersion of forum messages on television show ratings, after manually coding (using judges) a sample of their messages for valence and length, Godes and Mayzlin (2004) highlighted the potential of content analysis but concluded that the cost associated with their approach to data collection was prohibitively high and the data reliability was limited. Chevalier and Mayzlin (2006) used the number and average star ratings of reviews and the number of characters mentioned in the review to study the effect of book ratings on sales. Similarly, Dellarocas, Zhang, and Awad (2007) used volume of messages and average user ratings of movies to predict box office sales.

Unlike many product reviews sites, most online consumer forums do not include quantitative summaries of consumers' evaluations, such as star ratings. Furthermore, evidence with respect to overall reliability and predictive validity of online product ratings is mixed (Chen, Wu and Yoon 2004; Godes and Mayzlin 2004). Archak, Ghose, and Ipeirotis (2008) demonstrated the advantage of extracting a more multifaceted view of the content of product reviews (using text mining) to successfully predict product choices. Thus, although the aforementioned studies demonstrate that using summary statistics about online word-of-mouth information can be useful in predicting outcomes such as sales and ratings, they also highlight the need to delve deeper into the content of online discussions.

In this research, we propose going beyond the analysis of consumer star ratings and the effect of moments of consumer-generated content on sales toward uncovering the semantic relationships in the discussion and mapping the consumer-generated content itself. To do so, we utilize recent advances in text-mining techniques.

2.2 Text Mining and Marketing

Text mining (sometimes called knowledge discovery in text) refers to the process of extracting useful, meaningful, and non-trivial information from unstructured text (Dörre, Gerstl, and Seiffert 1999; Feldman et al. 1998; Feldman and Sanger 2006). For example, using what they call "undiscovered public knowledge," Swanson and colleagues found relationships between magnesium and migraine (Swanson 1988) and between biological viruses and weapons (Swanson and Smalheiser 2001) by merely text mining disjoint literatures and uncovering words common to both literature bases.

Text mining has become particularly popular and successful in fields in which meaningful information must be extracted from "mountains" of data in a relatively short period of time. Such fields include security and intelligence organizations looking for signs of irregular activity in the stream of public and less public written media (Fan et al. 2006), and doctors searching for biomedical information in the superabundance of medical information (Rzhetsky et al. 2004). Academics have used text mining in order to automatically meta-analyze the knowledge base on a particular topic (Börner, Chen, and Boyack 2003). With the increasing availability of digitized data sources, the business world has begun to explore the opportunities offered by text-mining tools to collect competitive intelligence, to automatically analyze the infinite stream of financial report data to search for patterns or irregularities (e.g., Feldman et al. 2009), and to syndicate and meta-analyze the wealth of information consumers are posting online (Feldman et al. 2008, Lee and Bradlow 2008). From the practitioner's point of view, several companies, such as Nielsen's BuzzMetrics, have begun offering text-mining services to businesses to help them achieve these goals.

Much of the advent of text mining has been restricted to computer scientists and information system researchers developing advanced text-mining apparatuses. Collaboration between computer scientists and business researchers has often facilitated the dissemination (albeit limited) of these tools to business research (e.g., Das and Chen 2007, Feldman et al. 2009, Lee and Bradlow 2008). These collaborations have in turn led to fruitful research initiatives by opening opportunities to quantitatively explore new sources of business data. In marketing, the first attempts to text mine consumer forums used manual text mining involving humans reading the messages and judging their content (e.g., Godes and Mayzlin 2004, Liu 2006). The authors described this inefficient and inaccurate methodology as a "tedious task" and a "costly and noisy process." Computer scientists such as Dave, Lawrence, and Pennock (2003), Hu and Liu (2004), Liu, Hu, and Cheng (2005), and Feldman et al. (2007) offered a solution to the tedium by building apparatuses that could automatically summarize and quantify consumer reviews. More recently, a handful of studies applied text mining to marketing, among the first were Lee and Bradlow (2008), who used text-mining techniques to automatically extract product attributes and attribute levels for conjoint analysis studies. Archack, Ghose, and Ipeirotis (2008) studied the effect of extracted preference for product attributes on sales of electronics products. Fowdur, Kadiyali, and Narayan (2009) used latent semantic analysis to assess the emotions elicited from movie plots and the relationship of those emotions to sales. Others have mainly used summaries of user-generated content such as quantity and valances to predict product pricing (Shin, Hanssens, and Gajula 2008), movie sales

(Eliashberg, Hui, and Zhang 2007; Pai and Siddarth 2009), and financial performance (Seshaderi and Tellis 2009).

We believe these applications of text-mining techniques to marketing represent just the tip of the iceberg, and our research adds another dimension to these efforts. We focus on utilizing text mining to assess market structure (Rosa, Spanjol, and Porac 2004). Unlike most of the aforementioned research, we focus not on product reviews, but on less structured consumer forums that discuss a product category (e.g., cars or pharmaceutical drugs). Such forums are more qualitative and are less focused than product reviews. Further, consumer forums rarely include quantitative overall assessments, such as product ratings. On the one hand, the unstructured nature of consumer forums makes the extraction of meaningful information from such forums more challenging. On the other hand, forums often provide richer discussions in terms of content and associations. Accordingly, our objective is more open ended than the existing studies, in that we propose a method by which to extract the *semantic network* of the forum discussion in order to understand the discussion itself, as opposed to using summary measures of the discussion as covariates in a predictive model. Finally, most of the earlier studies extracted well-structured information for *single* entities, such as products or product features, one at a time, quantifying their volume and valence. Our approach allows extracting, analyzing, and visualizing information about large numbers of entities and the *relationships* and comparison between them. We describe the proposed text-mining apparatus next.

3. THE TEXT-MINING METHODOLOGY

Our objective is to mine the discussions contained in the user-generated content itself and look for relationships between the semantic components. To do so, we have developed a text-mining apparatus specifically tailored to deal with the difficulties involved in mining consumer forums.

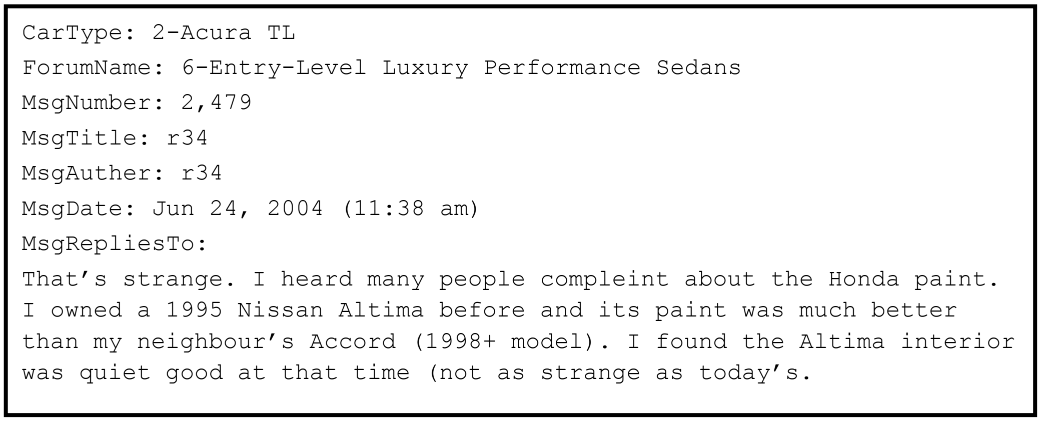
3.1 The Text-Mining Apparatus

Extracting structured products and attributes data, and the relationships among them, from Web forum messages involves the following five main steps:

1. *Downloading*: The Web pages are downloaded from a given forum site in html format.
2. *Cleaning*: html tags and non-textual information such as images and commercials are cleaned from the downloaded files.
3. *Information Extraction*. Products and product attributes are extracted from the messages.
4. *Chunking*: The textual parts are divided into informative units such as threads, messages, and sentences.
5. *Semantic relationships*. Two forms of product comparisons are computed: First, we generate a semantic network of co-occurrences of products in the forum. This analysis can provide an overview of the overall market structure. Second, we extract the snippets wherein products are compared, and examine with what terms products are co-mentioned and the sentiment of the semantic relationship.

Figure 1 depicts a typical message downloaded from a forum used in our first empirical application.

Figure 1 – A Typical Message Downloaded from the Forum Edmunds.com



```
CarType: 2-Acura TL
ForumName: 6-Entry-Level Luxury Performance Sedans
MsgNumber: 2,479
MsgTitle: r34
MsgAuther: r34
MsgDate: Jun 24, 2004 (11:38 am)
MsgRepliesTo:
That’s strange. I heard many people compleint about the Honda paint.
I owned a 1995 Nissan Altima before and its paint was much better
than my neighbour’s Accord (1998+ model). I found the Altima interior
was quiet good at that time (not as strange as today’s.
```

The first non-trivial step in the text mining process is information extraction. Information and semantic relationship extraction consist of eight main steps (see Table 1).

Table 1 - The Information Extraction Process

	Steps in the Information Extraction Process	**Purpose**	**Example** (Using the example in Figure 1)	**Commonly used Techniques***
1	Tokenization	Splitting the text into a series of basic elements, such as words and separators.	“I”, “heard”, “many”, “people”, “compleint”,”about”, “the”, “Honda”, “paint”, “.”	*Lookup tables* for token separators, *Regular Expressions*.
2	Part of speech tagging	Assigning part of speech tags (e.g., noun, verb, adjective) to each token (e.g., word, symbol, punctuation).	“[pronoun]**I**[pronoun] [verb]**heard**[verb] [adjective]**many**[adjective] [noun]**people**[noun] [unknown]**compleint**[unknown] [preposition]**about**[preposition] [def. article]**the**[def. article] [noun]**Honda**[noun] [noun]**paint** [noun] [period].[period]”	Hidden Markov Model (HMM), *Conditional Random Fields* (CRF), Maximum Entropy Models (MEM).
3	Spelling and grammar correction	Correcting spelling and grammar mistakes for the tagged text.	The following errors are corrected in the example in Figure 1: “**Compleint**” -> “**Complain**” “**quiet**” -> “**quite**”	*HMM*, *Edit Systems*, Neural Networks.
4	Chunking of tokens	Combining tagged tokens to verb and noun phrases wherever needed. The noun phrases relate to entities, whereas the verb phrases relate to the relationships between entities.	The nouns Nissan and Altima are combined to the noun phrase: [NP]**Nissan Altima**[NP] Similarly, the following noun phrases are extracted from the example in Figure 1: [NP]**many people**[NP] [NP]**Honda paint**[NP] [NP]**Altima interior**[NP]	Rules based, *CRF*, MEM.
5	Entity extraction	Identifying entity types for the noun phrases; For example, brands, products, and names.	In the sentence “I heard many people compleint about the Honda paint.” [brand] **Honda** [brand] is extracted.	*Rule based*, *CRF*, MEM, HMM.
6	Anaphora resolution	Identifying all indirect references, such as pronouns, and replacing them with the full name of the entity.	In the sentence “I owned a 1995 Nissan Altima before and its paint was much better” Replace “its” with “Nissan Altima”	*Support Vector Machine* (SVM), Knowledge-based approach using heuristics.
7	Relationship extraction	Identifying relationships between entities, including handling negation.	For the relationship “complain”: **[complain]** Actor: **many people** Object: **Honda** Attribute: **paint** **[complain]**	*Rule based*, CRF, HMM, Kernel-based methods, unsupervised machine learning such as Self Supervised Relation Extraction Systems.
8	Sentiment analysis	Detecting the sentiment (positive or negative) between entities.	In the sentence “I found the Altima interior was quiet good at that time” identify: [brand]**Altima**[brand] [attribute] **interior** [attribute] [positive] **quite good** [positive]	*Rule based*, Classification-based methods such as SVM, unsupervised machine learning.

* The technique/s in *italics* are the technique/s used in our text mining apparatus.

The extraction of product names and the terms used to describe products constitutes the process of converting unstructured textual data into a set of countable textual entities (e.g., brands, car models, car characteristics). The computer science literature outlines a plethora of methods for information extraction (see Feldman and Sanger 2006, Pang and Lee 2008 for a review). Most of the reported information extraction techniques rely on unsupervised machine learning methods such as hidden Markov models (Freitag 2000, Ray and Craven 2001), conditional random fields (CRF), and other maximum entropy models (McCallum and Wellner 2005). Our focus is on information extraction methods used to find pairs of entities (e.g., companies, drugs, products, and attributes) mentioned together, sometimes in the context of a phrase such as "better than" that describes the relationship between the entities (Feldman et al. 1998).

Such relationship-based information extraction has been mainly applied to formally written texts such as newswire, articles, scientific abstracts, or partially structured Web pages (e.g., seminar announcements). Entity extraction from consumer forums is much more difficult than that extracted from the corpora used in classic information extraction research (Lee 2007). First, the unit of analysis in forums (e.g., message) is often much shorter than the ones used in newswire analysis, and needs to be interpreted in the context wherein it is discussed. Second, product names can take various forms, and clues, such as capitalization, that exist in formal text are not reliable in consumer forums. Finally, the texts in consumer forums do not necessarily conform to grammar rules and conventions. For our case, brands, or company names, were found to be relatively easy to extract; that is, one can simply string-match names. Models or drug names, however, were much more difficult to extract, as they exhibited significant variation and ambiguity.

To deal with these difficulties, we developed a unique platform called CARE. CARE is a natural language processing engine for extracting complex relations from free text. CARE is a hybrid natural language processing system combining supervised machine learning methods, such as CRFs, and manually created rules to enhance the performance of the machine learning algorithm. The machine learning component is mainly in charge of identifying entities in the text (Steps 2–6 in

Table 1). Entity types may include brand names, product names, company names, people names, locations, products attributes, drugs, symptoms, technologies, and so forth. Automatic methods like CRF provided accuracy levels of above 90% for entities. The key to constructing such entity extraction tools is to note that lists of product brands and models are easy to come by on the Web. The difficulty lies in generalizing the formal product names to the many variants used. Terms referring to products generally consist of a subset of the brand name or model name. Any combination of the three components can be retained or dropped. Various delimiters (e.g., space, hyphen, or nothing) can also be used between them. As an example, here are some of the terms referring to the car model Audi A6 in one of the forums we analyzed: "Audi A6," "A6," "a6," "a 6," "the 6," "the six," "Audi 6," and "the VI". Note the highly erratic capitalization and the elision of various words or spaces. Accordingly, we augment the rule book extracted from manufacturers' websites and the CRF approach with some handcrafted rules to improve the accuracy of identifying product names. For example, the extraction rule "[Audi|audi] [A|a]6" would recognize, among others, "Audi A6," "Audi 6," "audi 6," "6," and "A6" as phrases used to describe the Audi A6. To distinguish between the number 6 to represent the Audi A6 and, say, the Mazda 6, CRFs are utilized to make use of the full information in the sentence. Of course, most of our extraction rules are much more complicated than this one. We found handcrafted extraction rules to be very useful for informal forum messages, wherein building rules by hand is substantially faster than labeling data, writing feature extractors, and training models. The advantage of handcrafted extraction rules is particularly true because we want not only to tag terms as products, but also to resolve which product is being referred to.

Using CRF for relationship extraction (Step 7) in Table 1 would be infeasible due to the size of the required training data (Feldman and Sanger 2006). The CARE platform provides a synergetic balance between an automatic name entity recognition process (based primarily on CRFs) and the manually crafted rules used to define relationships between entities to yield optimal results. CARE rulebooks are comprised of a set of relation-specific linguistic rules written for each domain. For example, the relation "Nissan Altima - paint - better - Honda Accord" in the phrase "Nissan Altima before and its paint was much better than my neighbour's Accord" can be extracted using the following rule:

Product → PRODUCT1 [Token 3] – Attribute →ATTRIBUTE – CompPred → PREDICATE [Token 3] Product → PRODUCT2,

where the follow apply:

- Products are matched using the information extraction procedure described above ("Nissan Altima" and "Honda Accord" in this example),
- Token 3 permits the rule to include up to three predicates (tokens) between the product and the attribute,
- Attribute refers to car attributes extracted using the information extraction procedure described above ("paint" in this example).
- CompPred refers to comparative predicate ("better than" in this example),

For every relationship identified in the text, the sets of pre-defined rules "compete" with each other as the best explanation for the extracted relationship. Each rule receives a weight based on the training text-mined data, which utilizes tagged sentences prepared by human experts, with the highest-weighted rule prevailing.

One of the difficulties with extracting textual information from consumer forums is that grammar, spelling mistakes, and typos are not uncommon. For example, in the sample message in Figure 1, the word "compleint" is misspelled and is grammatically incorrect. We perform the spelling check and corrections after parts of speech tagging and before chunking (Step 3 in Table 1). To identify the correct entity from the misspelled one, we use the following procedure. First, given a misspelled word (e.g., "compleint"), we find the words that are similar enough to have been the originally intended word. The number of possible mistakes that could have lead to the misspelling is extremely large, so we cannot hope to enumerate all possibilities and check whether any of them is a correct word. Instead, we formed a set of hash functions that depend on pieces (substrings) of the word and on the sets of letters, which produces a smaller set of candidates. Second, given two words, we estimate the probability that one word is a mistaken form of another. Here we use an

Edit System that measures the weighted edit distance between two strings (i.e., we count the number of changes needed to move from one word to the other). Finally, because the fixing of many mistakes depends on context, we use the content of the sentence to fix all mistaken words in a most probable fashion. We use a hidden Markov model and the Vitterbi algorithm to optimize the joint probability of possible word mistakes. For instance, in the example in Figure 1, "quiet good" was probably a spelling mistake and should be replaced with "quite good." Following the spelling procedure, part-of-speech tagging (Step 2 in Table 1) is repeated for the misspelled tokens.

Once the entity extraction is completed by the CARE, we conduct an anaphora resolution (Step 6 in Table 1). In an anaphora resolution, pronouns and indirect reference to entities are replaced with the full name of the entities. For example, in the sentence "I owned a 1995 Nissan Altima before and its paint was much better than my neighbour's Accord," the word "its" is replaced with the brand model "Nissan Altima." Similarly, terms such as "the drug" or "the car" are replaced with the product's name.

The accuracy of our information extraction procedure is relatively high. We use human evaluation to test the CARE system. Based on a random sample of 500 messages and manual evaluation of the results in each of the two empirical applications we next describe, we achieved recall (proportion of entities in the original text that are identified and classified correctly) of 88.3% and precision (proportion of entities identified that are classified correctly) of 95.2%, leading to an F1=2×(recall×precision)/(recall+precision) = 91.62%. These figures are in comparison to recall, precision, and F1 measures of 80% to 90% often achieved for products' entity extractions (Ding, Liu, and Zhang 2009). Commercial software applications frequently used in marketing academic papers and practical applications tend to result in lower accuracy.

After extracting the information, the records are divided into chunks at three levels: threads, messages, and sentences. Threads often contain hundreds of messages, whereas messages are short, often with only one or a few sentences or sentence fragments. For the purpose of this study, we use messages as our primary unit of analysis. That is, we look for co-occurrences of pairs or trios of models, brands, and terms in each message.

3.3 Occurrence, Co-occurrence, and Lifts

A basic unit we use for our analysis is the *co-occurrence* of terms. We analyze co-occurrences to look for patterns of discussion in the text-mined data and to form semantic networks and market structure perceptual maps.

Saiz and Simonsohn (2007) provide compelling evidence for the face validity of using the frequency of occurrence of terms on the Web to reflect the “true” likelihood of the corresponding phenomena. For example, Saiz and Simonsohn find high correlation between the frequency of occurrence of terms on the Web related to corruption in various states and the likelihood of corruption in these states.

Going beyond the mere occurrence of terms, we propose assessing the proximity or similarity between terms based on the frequency of their co-occurrence in the text. The notion of using co-occurrence as a proxy for proximity or similarity has roots in the knowledge discovery and co-word analysis literature (He 1999). For example, co-occurrence of words (known as co-word analysis) is frequently used to trace the development of science by tracking the co-occurrence frequency of pairs of words in various research fields (Callon, Law, and Rip 1986).

One premise behind utilizing the co-occurrence of terms to analyze consumer forum discussions is that consumers often compare products (Pang and Lee 2008). Many of the underlying constructs of social sciences, such as preference, utility, and attitude, are hardly absolute scales. Therefore, comparison is a relative measurement that allows consumers to better understand their own thoughts and to explain their evaluations and choices to others (Shafir, Simonson, and Tversky 1993). Indeed, decision experts often advise decision-makers to compare alternatives when making evaluations and decisions (Janis and Mann 1977). Schindler and Bickart (2005) found that direct comparison between brands and products in consumer forums is one of the main information-seeking motives for content generators and readers of these forums.

Comparisons were prevalent and helpful in the automatic analysis of sentences in the forum we mined. For example, given a sentence such as "Toyota is faster than Honda," we can automatically extract the two car manufacturers ("Toyota" and "Honda") and what attribute(s) are being compared, namely, speed. Our measure, which is the basis for much of the analysis we describe in the next section, is therefore a co-occurrence of terms. We start by analyzing the context-free co-occurrence of products in the same message to build a perceived perceptual map of the product as reflected by the forum discussion. We then explore the topics discussed with each of the products.

One limitation of using simple co-occurrence as a measure of similarity between terms is that if one of the terms appears frequently in the forum, its co-occurrence with nearly any term will be higher than that of a term that appears less frequently. For example, in the sedan cars forum described later, the car model "Toyota Camry" appears in our forum with "safety"-related words 379 times, relative to only 18 co-mentions of the car model "Volvo S40" with "safety"-related words. However, consumers mention the Toyota Camry 34,559 times in the forum, relative to only 580 for the Volvo S40. Thus, once we *normalize* for the mere occurrence of each car in the forum, we find that conditioned on one of the cars appearing in the message, the likelihood of "safety" realted words appearing in a sentence that includes Volvo S40 is much higher than such words appearing in a sentence that includes Toyota Camry. Such normalization is called lift. Lift is the ratio of the actual appearance of two terms appearing together to the frequency we would expect to see them together.[1] Thus the lift between terms A and B can be calculated as

$$Lift(A,B) = \frac{P(A,B)}{P(A) \times P(B)} \quad , \tag{1}$$

where P(X) is the probability of occurrence of term X in a given message, and P(X,Y) is the probability that both X and Y appear in a given message.

[1] In the context of brand switching, lifts are sometimes referred to as "flow" (Rao and Sabavala 1981).

A lift ratio lower than 1 suggests the two terms appear together less than one would expect by the mere occurrence of each of the two terms in the forum separately, whereas a lift ratio higher than 1 suggests a higher co-occurrence than one would expect by the separate occurrence of each term. In the example used above, the lift measure between Toyota Camry and "safety"-related words (2.08) is lower than the lift of the Volvo S40 with such words (4.9), suggesting that after normalization, the relationship between Volvo S40 and "safety" is stronger than the relationship between Toyota Camry and "safety."

Next we describe two applications of the proposed approach to a sedan cars forum and diabetes drugs forums.

4. EMPIRICAL APPLICATION 1: SEDAN AUTOMOBILE MARKET

The text-mining process described in Section 3 allows us to take a qualitative and unstructured dataset consisting of millions of (often grammatically incorrect) sentences and convert it into a quantifiable set of terms. We then aggregate the occurrence and co-occurrence of terms in the forum's messages to obtain a measure of relative frequency of discussion for each semantic relationship to create a *semantic network*. In this semantic network, terms appear closer to one another if they are mentioned together more than one would expect from chance (a high lift ratio). An advantage of converting the text-mined co-occurrence data into a semantic network is the resulting ability to present the forum's discussion in a graphical manner. One can then "zoom in" on certain domains or subsets of the network and trace the relationship between terms in more detail. We demonstrate this process using a consumer sedan cars consumer forum.

4.1 Sedan Automobile Data

The first step in text mining involves downloading the data to be mined. The sedan cars forum included 868,174 consumer messages (consisting of nearly 6 million sentences) posted by more than 76,000 unique consumers during the years 2006 and 2007 in the sedan car forum Edmunds.com (http://townHall-talk.edmunds.com/WebX/.ee9e22f/, see Table 2). From this repository, following the procedure described below, we extracted 28 car brands (e.g., "Honda" and "Toyota"), 135 car models[2] (e.g., "Honda Civic" and "Toyota Corolla"), and more than 1,000 common terms (mostly noun phrases and adjectives) used to describe these cars (e.g., "compact," "safe," "hybrid," and "leg room"). We used a web crawler through the manufacturer's websites to create a dictionary for the car models. We focused most of our analysis on the textual unit of a forum message. Within a message, we looked for co-occurrences between two car brands, two car models, or a car brand or model and a term used to describe it (e.g., "Toyota Corolla" and "compact").

Table 2 – Characteristics of the Sedan Car Forum Edmunds.com

No. of messages	868,174
No. of threads	557,193
No. of sentences	5,972,699
No. of unique users	76,587
No. of brands	28
No. of car models	135
No. of terms	1,038

We believe this dataset provides an appropriate platform for our analysis since it involves a rich product category with a large number of products and multiple product dimensions. Furthermore, this category is popular, involving both enthusiasts and lay consumers. These features make the discussion in this category interesting, yet challenging, to map.

4.2 Co-occurrence of Car Models

We begin with an analysis of the co-occurrence and lifts of car models mentioned in the same message. The premise behind the analysis of car models' co-occurrence is that the more frequently consumers mention two car models together in a sentence, the closer those cars are in the consumers' perceptual space. As mentioned previously, this line of reasoning has strong roots in the co-word analysis literature (He 1999). Whether the consumer highlights points of parity (e.g.,

[2] We included in our dataset and corresponding analyses only car models that were mentioned at least 100 times in the forum.

"Toyota Corolla and Honda Civic have similar prices") or points of differentiation (e.g., "Toyota Corolla differs from Honda Civic in terms of mpg"), we postulate that the fact that the consumer consciously compared the two cars highlights a sense of proximity or relationship in her mind (i.e., the dissimilarities between Honda Civic and Lamborghini would be too obvious to write about).

We constructed a 135×135 dyad matrix of lifts between each pair of sedan car models observed in the forum following Equation 1. The relationship between each pair of nodes has a symmetric strength reflected by the magnitude of the lifts (lift(i,j) = lift(j,i)). Examining the characteristics of this network may help us understand the nature of the forum discussion and the centrality of different car models to the discussion. First, note that given the large number of messages (close to 900K), the dyad matrix is relatively dense: 68% of the dyad lifts are positive. That is, on average, each car was mentioned at least once with 68% of the car models. The most compared-against car model in our forum is the Honda Accord, which was compared at least once to 133 (out of 134) car models. The Lexus LS500 is the least compared-against car, with only 13 car models compared thereto. We used several measures of network centrality to measure the centrality of different car models to the discussion. Specifically, we used the following: *degree centrality*—the number of cars models each car is mentioned with at least once; *betweenness centrality*—to what extend each car is on the shortest path between every pair of other car models in the network; and *eigenvector centrality*—to what extent each car is mentioned with cars that are central to the discussion.

Table 3 presents the centrality measures and occurrences of the 30 car models with the highest degree in the network. Cars such as the Honda Accord and the Honda Civic are not only mentioned frequently in the forum—as is evident in the rightmost column in Table 3—but are also central to the discussion and are frequently compared against, as is evidenced by their high centrality measures. On the other hand, several cars that have a lower number of occurrences in the forum (e.g., Chevrolet Corvette and Chrysler Cruiser) have very high centrality measures. Thus, these two relatively unique and eye-catching cars are frequently compared with a large and heterogeneous set of (other) cars.

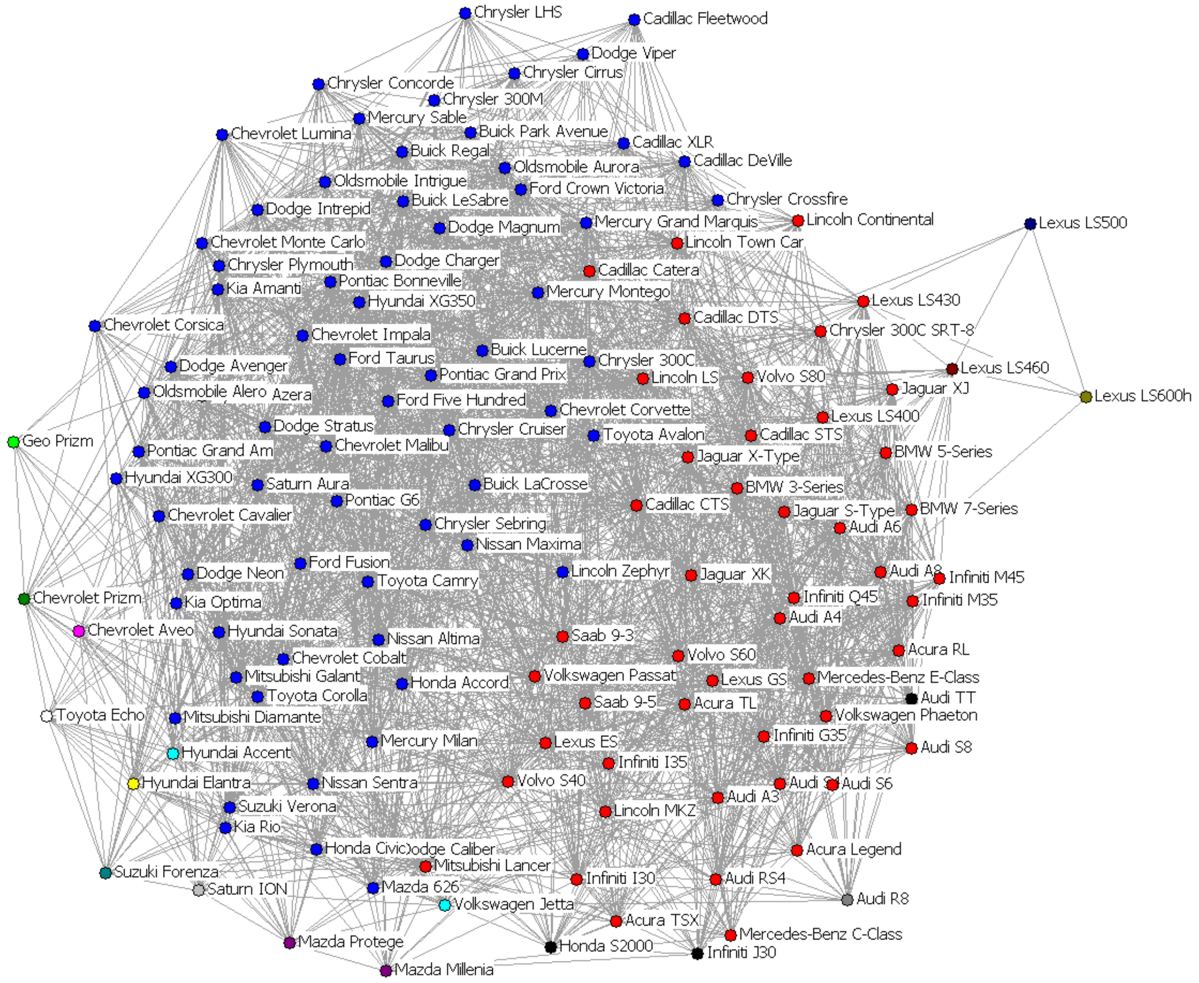
Table 3 – Centrality and Occurrence Measures for the Top 30 Car Models*

	Car model	Degree	Betweenness	Eigenvector	No. of occurrences
1	Honda Accord	99.25	0.954	15.93	58,443
2	Honda Civic	97.76	0.891	15.77	21,467
3	Lexus ES	97.76	0.861	15.78	11,540
4	Toyota Camry	97.76	0.710	15.89	34,559
5	Volkswagen Passat	97.76	0.794	15.86	16,474
6	Infiniti G35	97.02	0.713	15.79	23,247
7	Nissan Altima	96.27	0.605	15.80	13,061
8	Acura TL	95.52	0.545	15.75	29,400
9	Cadillac CTS	94.78	0.597	15.56	8,220
10	Nissan Maxima	94.78	0.533	15.69	10,727
11	Audi A4	94.03	0.547	15.54	13,454
12	Volkswagen Jetta	94.03	0.607	15.48	12,251
13	Toyota Corolla	93.28	0.579	15.41	7,133
14	Chevrolet Impala	92.54	0.457	15.46	11,659
15	Lincoln LS	92.54	0.594	15.36	3,092
16	Chrysler 300C	91.79	0.438	15.35	4,833
17	Toyota Avalon	91.79	0.597	15.25	12,796
18	Pontiac Grand Prix	91.05	0.425	15.24	2,327
19	Chevrolet Malibu	90.30	0.454	15.09	7,235
20	Hyundai Sonata	90.30	0.397	15.22	16,733
21	Audi A6	89.55	0.485	14.96	13,617
22	Ford Taurus	88.81	0.381	15.00	6,907
23	Chevrolet Corvette	88.06	0.364	14.90	2,254
24	Cadillac STS	87.31	0.418	14.67	4,071
25	Acura RL	85.82	0.312	14.65	9,258
26	Chrysler Cruiser	85.82	0.388	14.48	818
27	Nissan Sentra	85.08	0.379	14.34	3,170
28	Chrysler 300M	84.33	0.303	14.43	6,248
29	BMW 3-Series	83.58	0.355	14.16	2,890
30	Ford Fusion	83.58	0.367	14.16	10,227

* The centrality measures are with respect to all 135 car models in the network

Figure 2 presents a visual depiction of the lift matrix for the 135 car models mentioned in the forum, using Kamada and Kawai's (1989) spring-embedded algorithm. This algorithm, similar to multidimensional scaling, minimizes the stress of the spring system connecting the nodes in the network so that car models that are more similar (have higher lift) appear closer to one another in the graph. The edges connecting the nodes (car models) in the graph represent lifts between two car models that are significantly higher than 1 at the 1% level based on the χ^2 test.

Figure 2: Spring-Embedded – Kamada and Kawai – Network Graph of Sedan Car Models



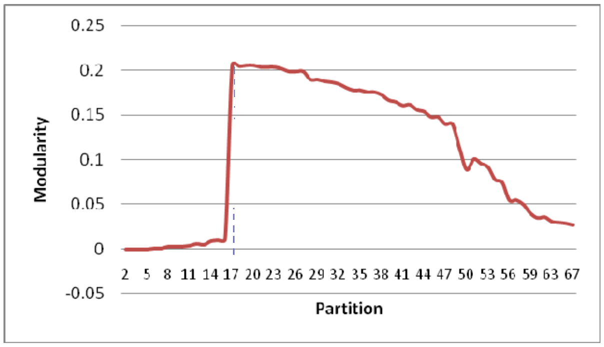
Although Figure 2 is somewhat crowded, it highlights some of the advantages of using the combination of text-mining techniques and network analysis to trace consumer perceptions and discussions. First, using the text-mining apparatus, we were able to measure simultaneous discussions for 135 different car models. Such an endeavor would be prohibitively difficult and costly using traditional marketing research methods. Second, unlike multivariate analysis techniques such as multidimensional scaling, network analysis permits us to study and visualize a large number of entities, thereby providing a comprehensive picture of the forum discussion.

We can gain several interesting insights from Figure 2. First, the car models in the southwest region of the figure are the smallest sedan cars in the market, such as the Toyota Echo and Geo, and the Chevrolet Prizm (note that the Geo and the Chevrolet Prizm refer to the same car under different brand names; not surprisingly, they appear next to each other in the network). As one moves eastward in the figure, the cars increase in size and luxuriousness, all the way to the high-end Lexus models (LS500 and LS600) at the far eastern edge of the network. Furthermore, the compact cars are clustered together in the southwest; the family cars appear in the upper middle; and the luxury cars appear in the west side of the figure. Cars of the same country of origin or the same make often appear close to one another (e.g., Audi A3, S4, and S6). Figure 2 provides a high degree of face validity, depicting the familiar sedan cars market structure, thus indirectly supporting the external validity of consumer forum data and the text-mining methodology used. Recall that the input into this figure is merely the co-occurrence of car models in a message, ignoring the content of the discussion about the two car models. We analyze the content of the discussion later.

To further explore the sedan cars market structure, we looked for clusters of car models in the car models network. That is, we looked for car models that were mentioned together frequently but were mentioned less frequently with other groups of car models. Since we are dealing with segmentation in a network, we adopted the Girvan-Newman community clustering algorithm, commonly used to cluster networks(Girvan and Newman 2002). Unlike typical social networks, in our semantic network, the communities consist of car models (rather than people) that were mentioned together frequently in the discussion. In the Girvan-Newman algorithm, the clusters are defined by a group of nodes that are densely connected within the cluster and less densely connected across clusters. This clustering algorithm relies on the notion of "betweenness," such that edges with high betweenness (edges lying on the shortest path between many nodes) are sequentially removed to create clusters. The Girvan-Newman algorithm uses the notion of modularity to assess the appropriateness of each cluster solution. Modularity is defined by the relationship between the density of edges within clusters and the density expected at random.

The Girvan-Newman algorithm identified 16 clusters of car models; at more than 16 car clusters, the modularity started dropping (see Figure 3). The colors of the nodes (cars) in Figure 2 represent the clusters membership. We identified two large clusters (the blue and red nodes), as well as 14 additional clusters consisting of one or two car models each. The clustering results provided high face validity to our analysis. The cluster of cars in blue includes, for the most part, economy and family cars, whereas the cluster in red includes cars that belong predominantly to the luxury market.[3]

Figure 3 – Modularities for the Girvan-Newman Clustering Algorithm



The breadth of the text-mining data allows us to assess simultaneously 135 car models as demonstrated in Figure 2. The depth of these data allows us to "zoom in" on the discussion of any subset of these car models found to be similar in the text-mining data or of particular interest to the marketer for managerial reasons.

4.3 "Zooming in" on the Discussion

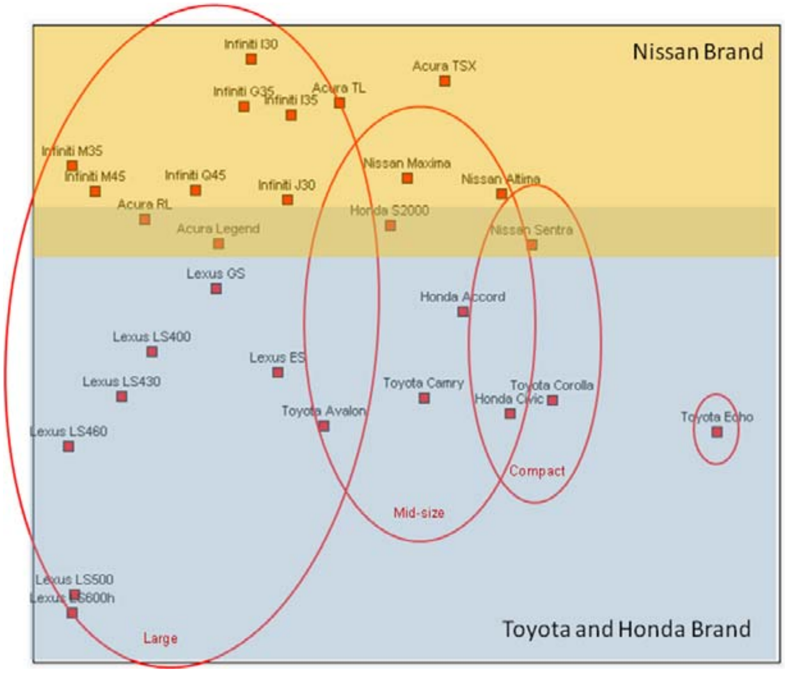
To demonstrate such zooming in, we look at the co-occurrence and lifts between the car models belonging to the competing Japanese brands, Honda, Toyota, and Nissan, and their high-end luxury brand extensions, Acura, Lexus, and Infiniti.

In Figure 4, we use an MDS to analyze the lifts between these car models appearing in the same message. For the most part, the Honda, Nissan, and Toyota brands are separated from their high-end brand extensions (Acura, Infiniti, and Lexus). An exception may be the Toyota Avalon,

[3] We also clustered the car models based on traditional k-means cluster analysis. The result of the k-means 2 clusters analysis were similar to those of the Girvan-Newman algorithm, suggesting the clustering solution found is robust to the clustering mechanism used (equivalence-based or community-based).

which is located closely in the discussion map to the Lexus ES. Indeed, the Avalon was reported in a car review article in the *L.A. Times* as "Avalon gives you Lexus taste on a Toyota budget." The x-axis on this map seems to reflect the size of the car moving from the smallest car in this set (Toyota Echo) to the largest and most luxurious cars (Lexus 460, 500, and 600).

Figure 4 – MDS Discussion Map of the Leading Japanese Car Models

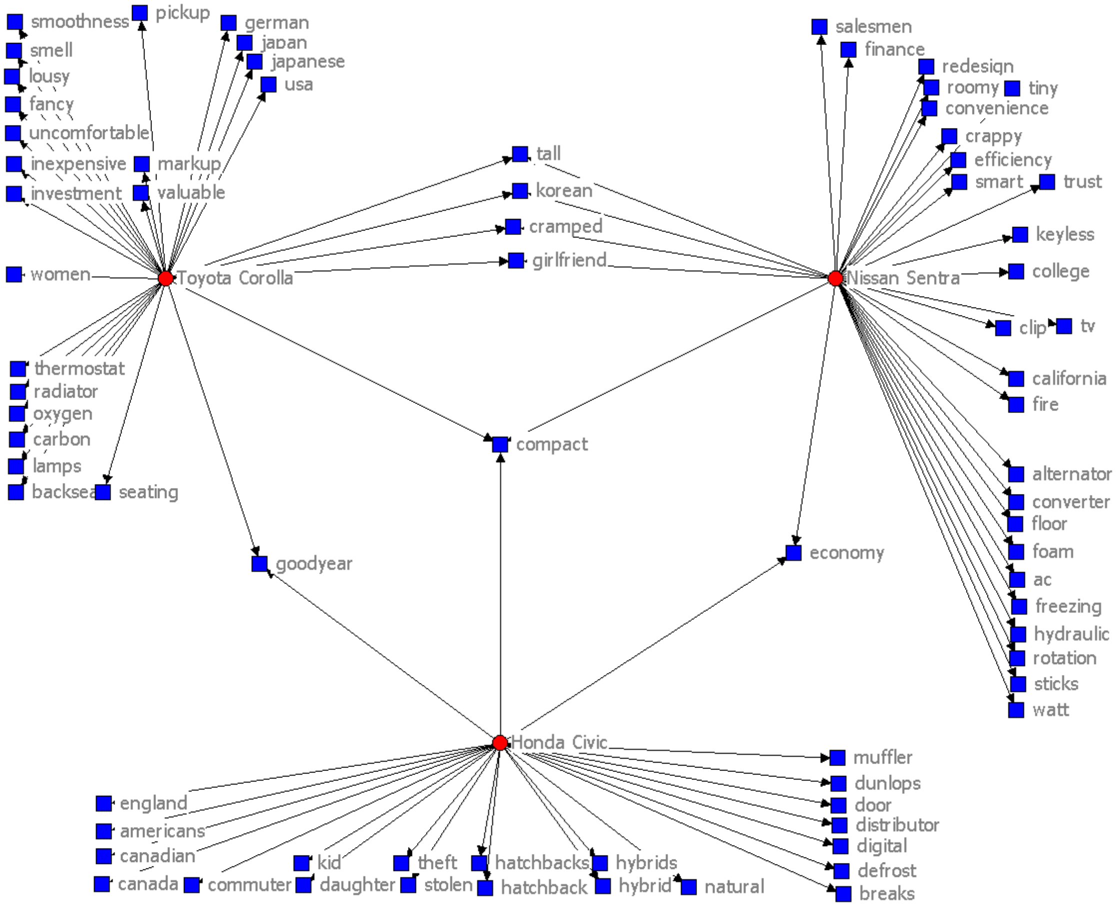


The cars could be grouped based on their size category moving from west to east. Note that among the luxury brands, Infiniti and Acura appear close to one another in the discussion map, whereas Lexus is separated in the southwest section of the map. Similarly, note that Honda Civic and Toyota Corolla are close to one another in the discussion map but are farther away from their Nissan counterpart, Nissan Sentra. Similar to Figure 3, Figure 4 provides a high degree of face validity. Additionally, Figure 4 allows us to assess the competitive perception and top-of-mind association of the car models as reflected by the comparison between them in the forum. Specifically, we identify that for the most part, the high-end competitive brands are perceived closer to one another than they are to the lower-end cars of their own brand, suggesting more top-of-mind competition than cannibalization in these cars' product lines.

Thus far, we have analyzed the co-occurrence of car models with one another in the forum. However, one of the most promising aspects of the text-mining methodology is the opportunity to quantify what consumers wrote about each of the cars and the terms mentioned with each pair of cars. This type of analysis allows us to drill one level deeper into consumers' discussions. Recall that in addition to the 135 car models, we extracted more than 1,000 nouns and adjectives consumers used to describe the cars. Therefore, we can investigate the frequency with which each term co-occurred with each of the car brands or models. As in the previous analyses, we focused on the lift measure to control for relative frequency of appearance of each term and the car brand or model in the forum. Figure 5 depicts the terms that exhibited the highest lift (lift > 5) with three compact Japanese cars: Honda Civic, Nissan Sentra, and Toyota Corolla. All lifts are statistically significant at the 0.01 level.

As Figure 5 shows, the only term appearing with high lift with all three cars is the adjective consumers used to describe these cars' category: "compact." Possibly of greater interest are the terms used to differentiate between the cars. These are terms consumers used frequently with one of the cars but not with the others, suggesting a point of difference in top-of-mind association. Interestingly, the Honda Civic was successful in differentiating itself from the other two cars based on terms like "hatchback" and "hybrid." Indeed, during the period mined, from among the three cars, only the Honda Civic offered hatchback and hybrid models. The Nissan Sentra, on the other hand, was differentiated by its keyless feature. The high lifts for this term with the Nissan Sentra suggest that consumers frequently mention this feature in discussing the car.

Figure 5 – Terms Commonly Appearing with Honda Civic, Nissan Sentra, and Toyota Corolla



Another interesting term forum users mentioned with the Nissan Sentra is "college," possibly suggesting the Nissan Sentra is perceived as a "college car." This market segment may not be obvious to Nissan through a simple demographic analysis of Sentra's buyers because the buyers may be parents of college students. The Toyota Corolla appeared frequently with terms like "valuable," "markup," "inexpensive," and "investment." Indeed, the Toyota Corolla is often presented in the media as a "good bang for your buck" car, constantly reaching Kelly Blue Book's Top 10 Best Resale Value Cars.

The analysis presented in Figure 5 depicts only the lifts between cars and terms. One could zoom in one step further in analyzing the valance and context of the relationship between the car and the terms used to describe it. For example, how often was the mention of Toyota Corolla with the term "smell" positive or negative? This analysis could be very valuable for the firm in detecting mechanical and other problems by "listening in" on consumers in real time. We demonstrate such valence and contextual analysis in the pharmaceutical application below.

In the next section, we use the extracted semantic network to go beyond graphical representation of the relationships between cars or between cars and the terms used to describe them, into statistical analysis of possible drivers of the co-occurrence or lifts between a pair of cars or terms.

4.4 Decomposing the Semantic Network

The network analyses and multivariate methods we have used thus far to analyze the text-mining data help us create perceptual maps that depict cars' similarities or dissimilarities as they emerge from co-occurrence of these cars in the forum discussion. Figure 5 goes a step further in explaining the dimensions underlying the co-occurrence between the three compact Japanese cars. However, one may wish to go beyond visual assessment of the terms used to describe three particular car models and into a more systematic analysis of the determinants of co-occurrence of cars in the discussion. Specifically, we wish to investigate the characteristics of the cars and the terms appearing in the forum that could explain the patterns of co-occurrences observed in our semantic network. To do so, we relate the variation in lifts between the 135 car models to the cars' characteristics (size, brand, manufacturer, country of origin, and price), the independent mentions of the cars in the forum, and their co-occurrence with adjectives and nouns consumers most commonly used to describe them in the forum.

We define y_{ij} as the lift between car models *i* and *j*, following Equation 1. Accordingly, our variable of interest is a vector of (135×134)/2 lifts between pairs of car models ($\mathbf{Y}$). Similarly, each explanatory variable is a vector reflecting the match between each pair of car models with respect to the variable of interest. We define the following explanatory variables as follows:

- **Brand** – $brand_{ij}$ equals 1 if both cars carry the same brand name (e.g., Honda Civic and Honda Accord) and zero otherwise.
- **Manufacturer** – $manuf_{ij}$ equals 1 if both cars are manufactured by the same parent company (e.g., Honda Civic and Acura TL) and zero otherwise.

- **Country of origin** – $country_{ij}$ equals 1 if the country of origin of both cars is the same (e.g., Honda Civic and Toyota Camry) and zero otherwise.
- **Size** – $size_{ij}$ equals 1 if both cars belong to the same size category as defined in the website (size categories used are compact, midsize, and large) and zero otherwise.
- **Price difference** – $price\_difference_{ij} = \left| MSRP_i - MSRP_j \right|$. This number is calculated as the mean absolute deviation between the Manufacturer Suggested Retail Price (MSRP) of the two cars in thousands of dollars. Thus the smaller the difference, the more similar the price of the two cars. The MSRP was elicited from the official MSRP price listed on Edmunds.com, the website that hosts the forum we mined. For cars that did not have a new model in 2008, we replaced the MSRP with Edmunds.com published True Market Value prices for the most recent model of the car.
- **Occurrence** – Lift measures can be sensitive to the base frequency of occurrence of each of the terms. To control for the base occurrence of each car model in the forum, we included the product of the occurrences (in thousands) of the two car models: $occurrence_{ij} = occurrence_i \times occurrence_j$.

Additionally, we wish to relate the co-occurrence between cars to the terms used to describe them in the forum. Because we extracted more than 1,000 adjectives and nouns from the forum, including all terms directly in the model would be impractical. We use Factor Analysis to reduce the dimensionality of the terms' space and identify the underlying topics most relevant to the discussion. Since some of the terms appear relatively infrequently in the forum, we focus on the 100 terms most frequently mentioned in the forum (see Table A1 in Appendix for list of terms). We used Factor Analysis with Varimax rotation to maximize the interpretability of the results. The scree plot (see Figure A1 in the Appendix) suggested there is an "elbow" after three factors. These first three factors explained 46% of the variance in the data. As can be seen in Table A1, the first factor is characterized mainly by upscale terms such as "class," "expensive," "best," "bigger," sport," "smaller," "premium," "performance," and "luxury." Thus we label this factor *upscale*. The second factor loads high on terms that relate to owners' experience with the car such as "problem," "purchased," "owned," "service," "experience," "old," "dealer," and "warranty." Accordingly, we label this factor *experience*. The third factor loads high on terms like "power," "fuel," "gas," "engine," mpg," and "mileage." Thus we label this factor *car efficiency*. Next, we related the lift between each pair of cars to the score of each car on these factors. Specifically, we define the following:

- **Meta-term Factors** – For each factor *k* and each car model *i*, the measure of similarity between each pair of cars (i, j) is calculated by $factor(k)_{ij} = MAD\left|score_{ki} - score_{kj}\right|$, where $score_{ki}$ is the score of car *i* on factor *k*.

From a statistical point of view, we need to address two issues before we can regress the vector of lifts between cars ($\mathbf{Y}$) on the set of explanatory variables. First, the dependent variable (lifts between cars) is not normally distributed. The lifts are censored at zero. Moreover, the lift measure has a mass at zero. Specifically, 32% of car model pairs were never mentioned together, thus resulting in a lift of zero. To address this issue, we use a Tobit censored regression model to relate the network lifts (y_{ij}) to the set of explanatory variables ($\mathbf{x_{ij}}$) such that

$$E[y_{ij} \mid \mathbf{x}_{ij}] = \phi\left(\frac{\mathbf{x}_{ij}\boldsymbol{\beta}}{\sigma}\right)(\mathbf{x}_{ij}\boldsymbol{\beta} + \sigma\lambda_{ij}), \qquad (2)$$

where $\lambda_i = \frac{\phi\left(\frac{\mathbf{x}_{ij}\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{x}_{ij}\boldsymbol{\beta}}{\sigma}\right)}$, and ϕ and Φ are the pdf and cdf of the Normal distribution.

Second, the above Tobit model assumes that observations in the data are independent. That is, the error term σ in Equation 2 is i.i.d across observations. However, in our vector of lift dyads ($\mathbf{Y}$), each car model appears 134 times. Therefore, the error terms in Equation 2 are not independent. Although a simple Tobit regression would produce valid point estimates, the standard errors are likely to be incorrect. To solve this problem, we adopted the Quadratic Assignment Procedure (QAP) method, which involves two steps. In the first step, a standard Tobit is performed to obtain the parameter estimates. In the second step, we permute the rows and columns of the dyad lift matrix and re-estimate the Tobit model. We repeat the permutation 1,000 times to estimate the distribution of the Tobit parameters. This approach has been shown to yield unbiased estimates and standard errors (Krackhardt 1988). We coded the QAP Tobit model in GAUSS. We report the result of the QAP Tobit model in Table 4.

Table 4 – Parameter Estimates from the QAP Tobit Model of Car Models Lifts

	Model 1 - Car Characteristics			**Model 2 - Car Terms (Factors)**			**Model 3 - Car Characteristics + Terms**		
	Coef.	Standardized Coef.	Pseudo P-value	Coef.	Standardized Coefficient	Pseudo P-value	Coef.	Standardized Coefficient	Pseudo P-value
Intercept	0.3543	--	0.000	0.4426	--	0.000	0.4774	--	0.000
Brand	0.5441	0.1017	0.000	--	--	--	0.5328	0.0996	0.000
Manufacturer	0.3447	0.1156	0.000	--	--	--	0.3493	0.1171	0.000
Country	0.0937	0.0427	0.000	--	--	--	0.1039	0.0473	0.000
Size	0.0292	0.0146	0.100	--	--	--	0.0203	0.0101	0.251
Price difference	-0.0176	-0.1852	0.000	--	--	--	-0.0165	-0.1742	0.000
Occurrence	--	--	--	0.0003	0.0176	0.161	0.0000	0.0005	0.998
Factors									
Upscale	--	--	--	-0.0023	-0.0027	0.894	0.0220	0.0258	0.195
Experience	--	--	--	-0.1037	-0.0962	0.000	-0.0790	-0.0733	0.000
Efficiency	--	--	--	-0.0706	-0.0651	0.000	-0.0638	-0.0588	0.000
Sigma	0.2820	--	0.000	0.3587	--	0.000	0.2735	--	0.000
-2log Likelihood	14,213			15,976			13,943		

* The two-tailed Pseudo P-value is calculated based on the proportion of times the absolute value of the estimated coefficient was larger than the absolute value of the QAP permuted coefficient estimate across the 1,000 iterations.

Model 1 regresses the lifts between car models on the similarity in car characteristics only; these are car characteristics that are exogenous to the forum discussion. Not surprisingly, cars sharing similar characteristics tend to be compared with one another in the forum more frequently. Looking at the standardized coefficients, we see that price similarity has the strongest relationship to co-mention of the cars. The effect of same-manufacturer similarity on the lift measure is almost three times that of country of origin. Thus, discussion of any pair of cars is mainly driven by their price tier, followed by their manufacturer and brand name. Size of the car had only a marginal relationship to cars' co-mention. This result is surprising because Edmunds.com organizes the cars in its website based on size. Note, however, that the price tier partially captures the effect of car size. The website management may be able to use the results of this regression to better organize the website by groups that are meaningful to users.

Next, we looked at how information that is endogenous to the forum can help explain the lifts between cars. Specifically, we included in the QAP Tobit model the independent occurrence of each car in the forum, and the similarity between the scores of each pair of cars on the three factors, generated from the Factor Analysis. We predict that cars mentioned with similar terms (low MAD between their factor scores) are more likely to be mentioned together in the forum (high lifts). Thus we expect a negative relationship (coefficient) between factors MAD and lifts. As Model 2 in Table 4 shows, both the "experience" and "efficiency" factors MAD had a significant negative effect on the lifts. That is, cars that share a similar pattern of mention with terms related to the experience with the car and car efficiency were also more likely to be mentioned together in a sentence, providing insight into the topic used in the forum to relate the two cars. In Model 3, we included both the car characteristics, which are exogenous to the forum, and the occurrence and factor scores in a single model. The nature of the results did not change from the separate nested models. The likelihood ratio test for nested models suggests the full model (Model 3) explained the lift measures significantly better than each of the nested models ($\chi^2_{d.f.=4} = 270$, $p < 0.001$ relative to Model 1 and $\chi^2_{d.f.=5} = 2033$, $p < 0.001$ relative to Model 2). Thus the similarity in the cars' characteristics and price as well as the discussion about the cars help explain the degree of comparison between cars in the discussion.

5. EMPIRICAL APPLICATION 2: DIABETES DRUGS FORUMS

In the previous sedan cars study, our text-mining analysis relied mainly on entity extraction and the notion of co-occurrence. The purpose of this study is to further examine this approach as well as go beyond mere co-occurrence and into deeper textual relationship and sentiment analysis, such as a mention of a drug in relation to an adverse drug reaction, or patients mentioning co-taking of two drugs. One of the premises of text-mining is to provide a quantifiable large-scale exploratory research. Thus, in addition to the market structure analyses the previous study presented, we wish to examine also whether text mining has the capacity to add insights that may be difficult or expensive to obtain using traditional methods. Additionally, one of the advantages of forums data and the text-mining apparatus is that data stream and can be analyzed in real time. In the previous study, we treated the data as static and did not investigate how the discussion changed over time. In the current study, we wish to demonstrate such dynamic analysis. We choose to focus in the second application on pharmaceutical drugs. Contrary to car usage, drug consumption involves physiological as well as psychological reactions, often leading to high involvement and to necessity of consumption, making the forums very active and involved. Specifically, we study forums discussing diabetes drugs. We selected the diabetics field since diabetics is a worldwide disease with multiple pharmaceutical treatments and an active and involved group of patients sharing their experiences over multiple forums.

5.1 Diabetes Drugs Data

We downloaded the entire forum discussion from five of the largest diabetes drugs forums. Table 5 describes the number of threads, messages, and sentences in each forum as well as the number of unique contributors to each forum. Overall, we mined a total of 670,000 messages (over 5 million sentences). We used a web crawler through medical websites to create a dictionary for the drugs' vocabulary.

Table 5 – The Diabetes Drugs Forums Data

Forum	Threads	Messages	Sentences	Users	Dates
DiabetesForums.com	17,229	228,690	1,449,757	4,881	02/2002-05/2008
HealthBoards.com	4,418	24,934	216,220	3,723	11/2000-05/2008
Forum.lowcarber.org	22,092	325,592	3,106,362	7,172	10/2002-05/2008
Diabetes.Blog.com	61	29,359	227,878	3,922	07/2005-05/2008
DiabetesDaily.com	5,884	62,527	380,158	2,169	05/2006-06/2008
Total	**49,684**	**671,102**	**5,380,375**	**21,867**	

5.2 Analyzing Adverse Drug Reactions

First, we would like to use consumer forums to asses consumer perceptions of a phenomenon termed “adverse drug reaction” (ADR), which is medical damage caused by taking a given medication at a normal dose. ADR is more commonly referred to as "side effect;" however, side effects can be both negative and positive (see the famous case of Viagra), whereas ADRs refer to only negative effects. Every drug has the potential for ADR (Roden 2008). Recent estimations are that ADR is a cause of 3% to 5% of all hospitalizations (around 300,000 hospitalizations annually in the United States). Prior to approval and market introduction, ADRs are examined in clinical trials on a sample of patients. Because of the relatively small number of patients studied in clinical trials, their short duration, and idiosyncratic conditions, clinical trials often miss ADRs. Accordingly, there are several mechanisms for surveillance of ADRs, such as cohort and case studies, population statistics, and anecdotal reporting from journals and doctors (see Table 2 in Edwards and Aronson [2000] for a comprehensive list). Most of the post-marketing ADR reporting is put in the hands of physicians to formally report to organizations such as the World Health Organization or the Food and Drug Administration, or informally disseminate the information among their peers. To the best of our knowledge, no automatic mechanism is in place for patients to directly express their ADR concerns. Furthermore, patients are constantly searching for ADR information. The package inserts are not always updated and often consist of long checklists, primarily motivated by legal concerns, making it difficult for patients to see the forest for the trees. The difficulty find the prevalence of ADRs may be one reason for the popularity of pharmaceutical or disease-focused forums, where patients can share their common experiences with drugs. The forums provide patients' firsthand experiences with the drugs and act as a living environment that keeps updating itself overtime. Similar to car forums, we have found that by tapping into these forums, drug companies can gain a real-time window into emerging consumers' views. The costs of such an approach using text mining is much lower and less sensitive to sample-size concerns relative to traditional medical post-marketing research methods.

To explore the ADRs mentioned with each diabetes drug, we have to go beyond the co-occurrence between drug i and drug j or drug i and ADR k and explore the full relationship between the drug and the ADR that co-occurred in the same message. For example, in the following three sentences, the drug "Actos" co-occurred with the ADR "nausea;" however, only the first sentence refers to "nausea" being an ADR for "Actos": (1) "I had a terrible nausea after taking Actos;" (2)

"Unlike other drugs Actos does not cause nausea;" (3) "I switched from Actos to Lantus and had a terrible nausea." Thus the need to identify contextual relationship between drugs and ADRs goes beyond the co-occurrence text mining used in the cars' study. The CARE apparatus described in Section 3, which we developed and trained specifically for consumer forums, combines both machine learning and handcrafted rules to achieve this goal. The use of the CARE apparatus is a methodological contribution over the text-mining applications in the existing marketing literature, which typically focuses on occurrence and valance summaries. The CARE apparatus allowed us to identify (1) drug-drug co-occurrence, (2) drug-ADR sentiments, (3) drug-usage sentiments, (4) drug-doctor sentiments, and (5) time trends.

We used the CARE apparatus to create a list of all ADRs that were mentioned in negative relationship with each of the diabetes drugs. We then converted the co-occurrence data to lifts, normalizing for the independent occurrence of each drug and each ADR in the forum. The first three columns in Table 6 list all the drug-ADR relationships that had a lift significantly larger than 1 at the 95% level (recall that lift=1 means the co-occurrence is equal to the co-occurrence expected by mere chance). In order to evaluate the validity of the extracted drug-ADR relationships, we also collected ADR information for each diabetes drug from WebMD, the leading health portal in the United States (comScore 2007). On WebMD, the ADRs are rated by their frequency of occurrence and severity. The last two columns in Table 6 report for each drug-ADR relationship found in the forums data, whether the ADR was reported for the particular drug in WebMD, and if so the WebMD ratings for the frequency (common, infrequent, or rare), and severity (severe, less severe) of the ADR. Most (70%) ADRs identified as appearing frequently with each drug in the forum were associated with frequent and/or severe known ADRs. The nine ADRs in this list that were both common and severe had an average higher lift (4.2), relative to the average lift of the remaining ADRs (2.4). Thus, this analysis provides an external validity for the user-generated content and the text-mining apparatus.

Table 6 – Drug-ADR Relationships Extracted from the Forums

Forum Mentions			WebMD	
Drug	**Effect**	**Lift**	**Frequency**	**Severity**
Actos	Fluid retention	12.60	Infrequent	Severe
Actos	Weight gain	5.04	Rare	Severe
Actos	Swelling	5.03	Infrequent	Severe
Actos	Heart disorders	2.87	Rare	Severe
Amaryl	Bad sugar	2.02	Common	Severe
Apidra	Bad sugar	21.68	Common	Severe
Avendia	Weight gain	3.74	Rare	Severe
Avendia	Pain	1.68	Common	Less severe
Byetta	Nausea	2.00	Common	Less severe
Byetta	Hair loss	1.95	Doesn't exist	
Byetta	No appetite	1.88	Infrequent	Less severe
Byetta	Gastro problems	1.49	Common	Less severe
Byetta	Drowsiness	1.44	Rare	Less severe
Byetta	Dermatitis	1.43	Rare	Severe
Byetta	Dizziness	1.35	Infrequent	Less severe
Byetta	Skin problem	1.29	Rare	Severe
Byetta	Cold symptoms	1.25	Doesn't exist	
Byetta	Headache	1.25	Infrequent	Less severe
Glipizide	Weight gain	2.31	Common	Severe
Glipizide	Bad sugar	1.82	Common	Severe
Glucophage	Digestive disorders	3.19	Common	Less severe
Gylburide	Weight gain	2.51	Common	Severe
Gylburide	Bad sugar	1.89	Common	Severe
Humalog	Bad sugar	1.96	Common	Severe
Januvia	Cold symptoms	2.12	Infrequent	Severe
Januvia	Paresthesias	2.37	Doesn't exist	
Lantus	Mood problem	2.89	Doesn't exist	
Lantus	Paresthesias	1.77	Rare	Severe
Lantus	Weight gain	1.46	Doesn't exist	
Lantus	Bad sugar	1.41	Common	Severe
Levemir	Pain	2.00	Doesn't exist	
Levemir	Weight gain	1.66	Infrequent	Less severe
metformin	Digestive disorders	2.04	Common	Less severe
metformin	Cramps	1.88	Common	Less severe
metformin	Kidney problems	1.88	Doesn't exist	
metformin	Mood problems	1.37	Doesn't exist	
metformin	Bad sugar	1.28	Rare	Severe
Novolog	Cold symptoms	2.36	Doesn't exist	
Novolog	Bad sugar	1.87	Common	Severe
Symlin	Nausea	1.72	Common	Less severe

Possibly more interesting are the six ADRs that were frequently mentioned in the forums but are not reported in WebMD (Byetta - hair loss and cold symptoms; Januvia – paresthesias; Lantus - mood problems and weight gain; Levemir – pain; Metformin – kidney problems and mood problems, and Novolog – cold symptoms). Patients' mentions of these ADRs in the forums should, at the very least, raise a flag for health officials to track these possible drug reactions. Note that several of the ADRs not reported are "softer" ADRs, such as hair loss and mood problems, which may not be considered medically serious but are ones patients are sensitive to. The relationship between diabetes and mood problems such as depression is well documented (Anderson et al. 2001). Thus, although the patient's association of her psychological condition, which is related to her physiological condition, with a particular drug may be a misattribution, the pharmaceutical firm should be aware of such common misattributions in marketing their drugs.

One can also create a "snake plot," commonly used to conduct gap analysis and brand perceptions, to compare the different drugs with respect to their likelihood of being mentioned with various ADRs in the forum. In Figure 6, we demonstrate such ADR-based drug positioning analysis for three of the leading diabetes drugs: Byetta, Lantus, and Metformin. Lantus seems to be causing fewer ADRs than Byetta and Metformin (for most ADRs, it has a lift lower than 1). However, two of the ADRs that have high lift with Lantus are directly related to diabetes problems: weight gain and bad sugar. Furthermore, mood problems seem to be an issue for Lantus users (which may be a result of the fact that this drug is an invasive injection treatment). Byetta users commonly discuss hair loss and light ingestion problems, such as nausea and no appetite. Metformin users, on the other hand, report more severe digestive disorders, such as kidney problems and cramps.

Similarly, we can depict the association between ADRs and drugs by a two-mode semantic network. Figure 7 presents such a map for Actos, Byetta, Lantus, and Metformin. The edges on the network depict lifts greater than 1. Such a map can illuminate points of differentiation and points of parity between the drugs. For example, although Actos shares many of the ADRs with Metformin, heart and respiratory disorders are common for Actos but not for Metformin. Such points of differentiation can serve marketers in positioning the drugs vis-à-vis each other.

Figure 6 – Lifts between ADRs and Byetta, Lantus, and Metformin

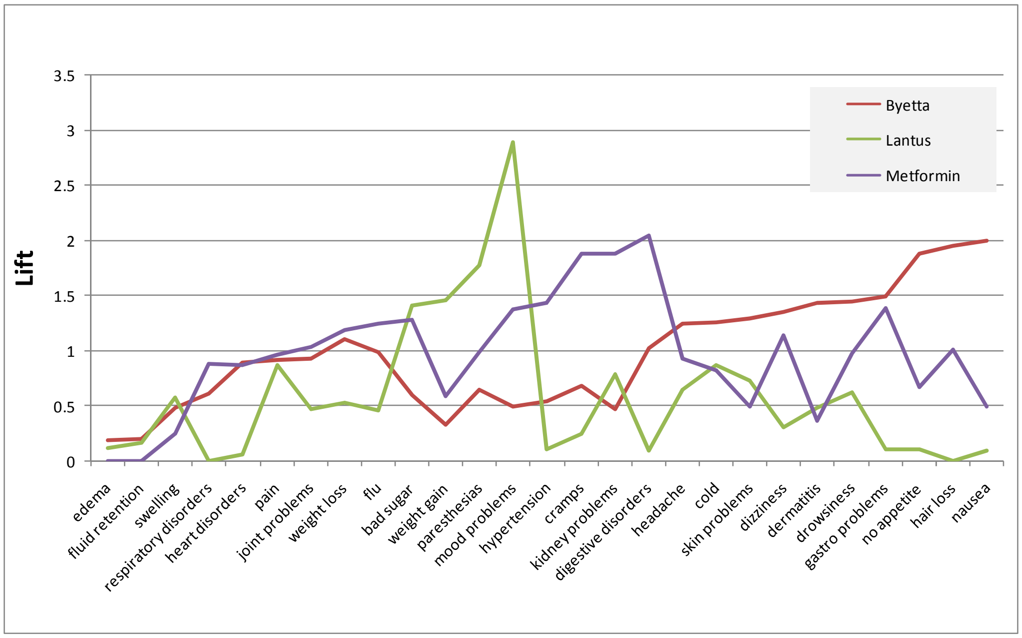
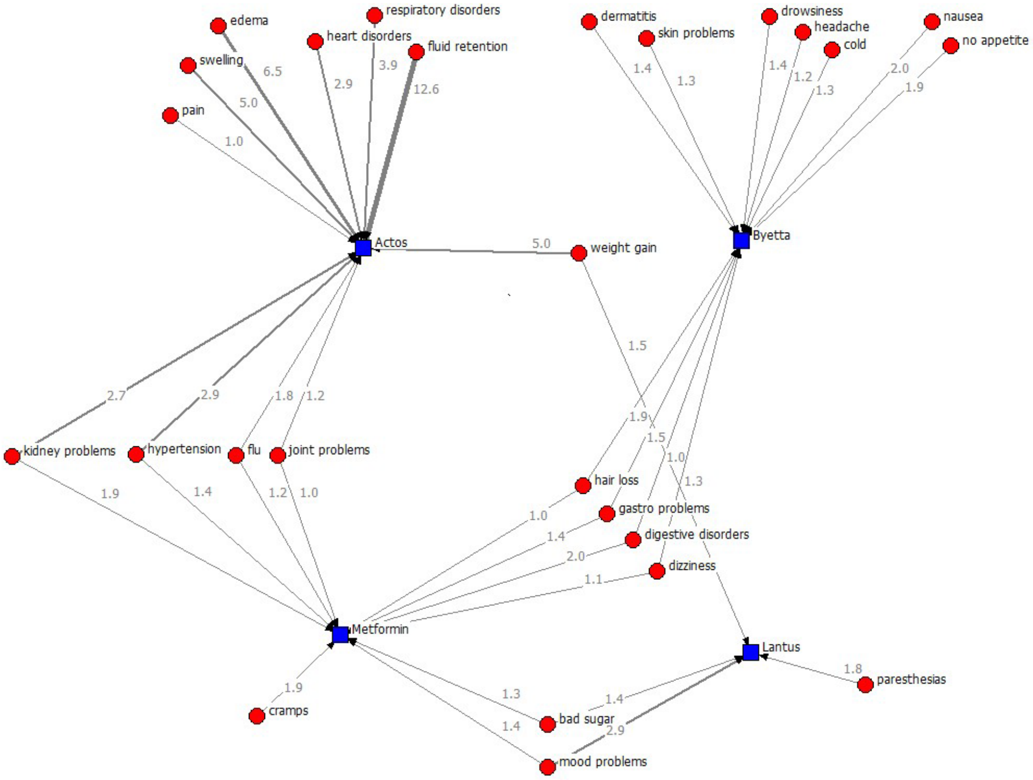


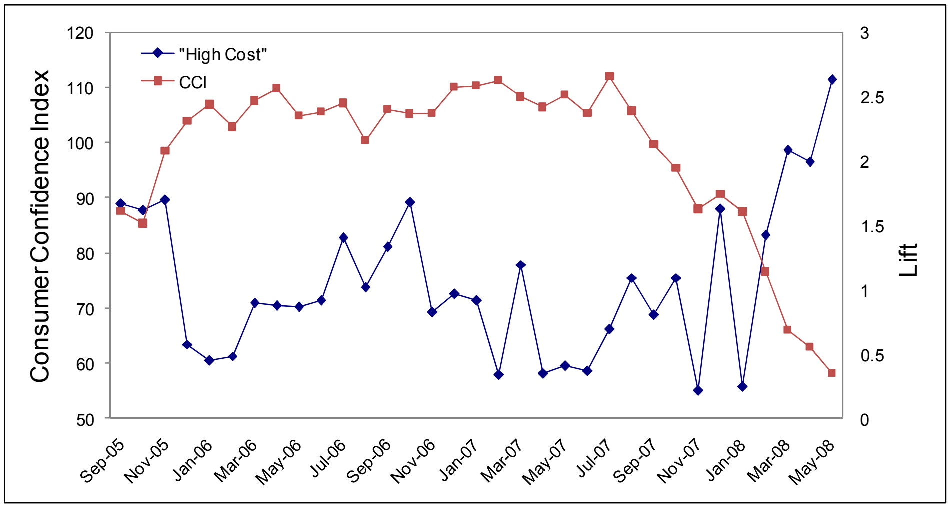
Figure 7 – A Semantic Network of Drugs-ADRs relationships



5.3 The Dynamics of the Discussion

One of the advantages of "listening in" on the content consumers are generating on forums is that this source of data keeps streaming in real time. Moreover, the time stamp of each forum message is readily available and can be easily textmined, allowing the researcher to assess the dynamics in the discussion. To demonstrate such dynamics, we looked at one of the more expensive drugs in the category (Byetta) and tracked how frequently forum users mentioned the drug in reference to high cost or being expensive over time. Figure 8 presents the lifts between Byetta and high-cost-related sentiment over a period of almost three years between September 2005 and May 2008.

Figure 8 – Mentions of Byetta with "High-Cost" Sentiment vs. the Consumer Confidence Index



Byetta mentions with high cost were moderately high during September to November of 2005, then until December of 2007, the mentions with high costs fluctuated at an average lower level and then dramatically increased in 2008. One possible reason for these observed dynamics is price variations over time, though we could not find any evidence for a significant increase/decrease in Byetta's price during that period. To further investigate the source of dynamics, we plot in Figure 8 the Consumer Confidence Index (CCI) during the aforementioned period. The CCI is based on a longitudinal monthly survey that measures consumers' assessment (confidence) of the state of the U.S. economy and consumers' saving and spending behavior (see also Fowdur, Kadiyali, and

Narayan 2009, for use of the CCI). The correlation between Byetta mentions with high cost and the CCI is strong and negative (average correlation of -0.7, $p<0.05$). When consumers' confidence in the U.S. economy was low (late 2005 and the first half of 2008), the high price of Byetta relative to its competitors played a more important role in patients' discussions of the drug. This finding may suggest that in marketing the drug, the pharmaceutical firm needs to pay close attention to the state of the economy and to the effect it may have on consumers' perception of the drug. More generally, this analysis provides a first step in exploring the potential in analyzing the dynamics in consumers' discussion, using the proposed text-mining apparatus.

6. DISCUSSION

In this paper, we propose a "sonar" for marketing researchers to listen to consumers' ongoing discussions over the World Wide Web. The task of mining and quantifying the wealth of online data consumers generate on the Web is difficult for several reasons. First, the discussion is spread across the universe of Web. Second, the quantity of data posted is large, unstructured, and often masked by less relevant graphical and textual data. Finally, the data consumers post are primarily qualitative in nature, making them difficult to quantify. To overcome these difficulties, we utilize text mining and network analysis tools to first mine and quantify the information, and then analyze it in a meaningful way.

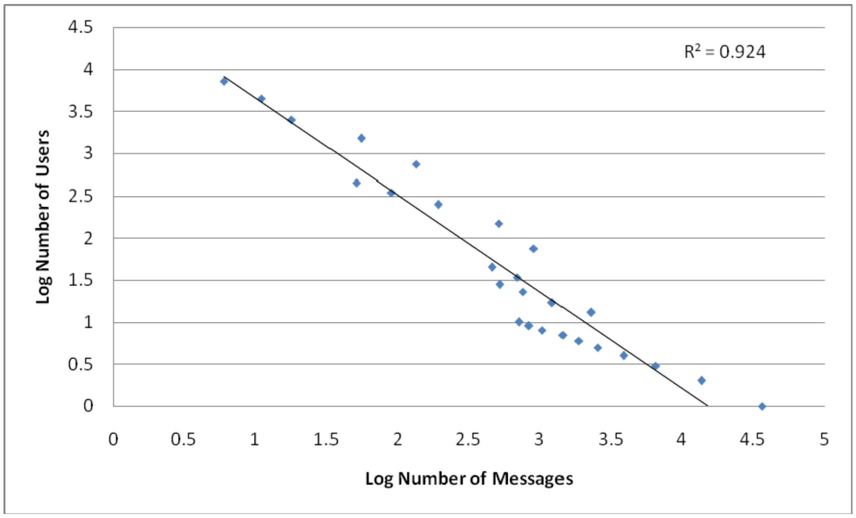
We demonstrate the value of the proposed apparatus in two studies involving sedan cars and diabetes drugs forums. Because the structure and environment of the automotive market is relatively familiar, we use this application to test the face validity of the proposed approach. Indeed, the semantic network derived from the car forum provides a high degree of face validity in terms of the centrality of cars to the discussion and the similarity between cars as assessed by their co-occurrence in the forum's messages. Similarly, analyzing the adverse drug reactions mentioned in the diabetes drugs forums and comparing them with the adverse drug reactions reported in formal media provides an additional assessment for the validity of the proposed approach.

In addition to the demonstrated ability of the proposed approach to capture the adverse drug reactions that are formally reported to be common and severe, we provide some anecdotal evidence for the external validity of the information mined from the sedan cars consumer forum. The National Insurance Crime Bureau has compiled a list of the vehicles most frequently reported stolen in the United States in 2007. According to this report, the three most stolen cars in the United States in 2007 were the Honda Accord, Honda Civic and Toyota Camry. We mined our forum to identify the cars most frequently mentioned with theft-related phrases ("stolen," "steal," and "theft"). Interestingly, the three car models most frequently mentioned together with these terms perfectly match the Crime Bureau's list: the Honda Accord (165 occurrences), the Honda Civic (101 occurrences), and the Toyota Camry (71 occurrences). This type of analysis may suggest that some of the value in mining consumer-generated content lies in what Surowiecki (2004) called the Wisdom of the Crowds. Although that many of the forum's members are unlikely to have predicted correctly the three most stolen cars in the United States, aggregating the information across members provided an accurate answer.

The analysis of the adjectives and nouns commonly mentioned with each car model provides insightful information with respect to the content of the discussion. The investigation of the drivers of co-occurrence of cars in the forum reveals that cars with similar sticker prices, common brands, or common manufacturers are likely to be compared in the forum discussion. In the second study, we further explore specific textual relationships and go beyond mere co-occurrence of terms to investigate adverse drug reactions and the dynamics of the discussion. Because our objective is more descriptive than predictive, we focused on utilizing text mining and network analysis tools to describe the nature of discussion in the forum. Future research might explore the potential of using the proposed approach as a predictive tool. Such an endeavor should consider further how well the forum participants represent the population of consumers at large and the risk of firms manipulating the discussion (Dellarocas 2006).

In text mining the consumer forum, we conducted a census of all the messages in the focal forum. Thus, from a sampling point of view, the analysis reflects well the discussion in the forum. However, one possible concern with analyzing consumer forum data is that the data may reflect mainly the discussion of a few "active" users as opposed to the entire population of the forum's posters. Indeed, 10% (8,620) of the users posted over 80% of the messages in the sedan cars forum; 47% posted only once in the forum. The log-log relationship between number of users and number of messages they post is close to linear ($R^2 = 0.924$, see Figure 9).

Figure 9 - Log Number of Users to Log Number of Messages in the Sedan Cars Forum



To test whether the active users—that is, those generating the majority of the content in the forum—differ from the less active users, we calculated the correlation between the car model co-occurrence matrix generated only by the active users and the matrix generated by the less active users. We defined active users as those who posted at least 10 messages in the forum (11% of the users, who generated 82% of the content). The less active group consists of the 89% who posted between one and nine messages each and accounted for 18% of the content. The correlation between the two matrices is 0.857 (QAP Pseudo p-value < 0.001), suggesting that for the investigated forum, the active and less active users are similar in terms of content generated. Thus our analysis reflects not only the nature of the discussion in the forum, but also the discussion of various users in the forum (Dwyer 2009).

The high face validity of the forums analysis may suggest the proposed approach has the required external validity to reflect not only the opinions and views of the forum members, but also those of the wider population of consumers. Future research could explore this topic more formally. Furthermore, although text mining allows us to minimize recall bias and demand effects commonly present in eliciting information from consumers, views posted on the forums may be biased because the respondents aim for their views to be publicly available on the Web.

Future research could also explore applications of the proposed approach in less established domains than the sedan cars and pharmaceutical drugs in order to provide initial understanding of consumers' discussion in an emerging domain. Furthermore, one could extend the application of proposed text-mining apparatus beyond consumer forums to mining blogs, product reviews, or more formal news articles. In fact, the mining of more formal channels is often easier because the context of the discussion is more organized and the language used tends to follow grammatical standards and rules.

Figure 8 provides a first step in exploring the opportunity of utilizing the real-time stream of data consumer forums provide. For new (or re-positioned) products, one can take a dynamic approach to the analysis of the semantic network generated by the forum discussion. For example, one can use the proposed approach to study dynamics of the discussion pre-, during, and post-launch of a new product.

In summary, we hope the text mining and semantic network analysis presented in this paper provides a first step in exploring the extremely large, rich, and useful consumer data readily available on the Web 2.0.

REFERENCES

Anderson, Ryan J, Kenneth E. Freedland, Ray E. Clouse, and Patrick L Lustman (2001), "The Prevalence of Comorbid Depression in Adults with Diabetes: A Meta-Analysis," *Diabetes Care*, 24 (6), 1069-78.

Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis (2008), Deriving the Pricing Power of Product Features by Mining Consumer Reviews: Working Paper, New York University.

Börner, Katy, Chaomei Chen, and Kevin W. Boyack (2003), "Visualizing Knowledge Domains," *Annual review of information science and technology,* 37, 179-255.

Callon, Michel, John Law, and Arie Rip (1986), *Mapping the dynamics of science and technology: sociology of science in the real world*, Houndmills, Basingstoke: The Macmillan Press Ltd.

Chen, Pei-Yu, Shin-Yi Wu, and Jungsun Yoon (2004), "The Impact of Online Recommendations and Consumer Feedback on Sales." *Proceedings of the International Conference on Information Systems 2004*, pp. 711–24 ICIS.

Chevalier, Judith A. and Dina Mayzlin (2006), "The effect of word of mouth online: Online book reviews," *Journal of Marketing Research*, 43 (3), 345-54.

Das, Sanjiv R. and Mike Y. Chen (2007), "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," *Management Science*, 53 (9), 1375-88.

Dave, Kushal, Steve Lawrence, and David M. Pennock (2003), "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," *Proceedings of the 12th international conference on World Wide Web*, 519-28.

Dellarocas, Chrysanthos, Xiaoquan Michael Zhang, and Neveen Farag Awad (2007), "Exploring the Value of Online Product Ratings in Revenue Forecasting: The Case of Motion Pictures," *Journal of Interactive Marketing*, 21 (4), 23-45.

Dellarocas, Chrysanthos N. (2006), "Strategic manipulation of Internet opinion forums: Implications for consumers and firms," *Management Science*, 52 (10), 1577-93.

Ding, Xiaowen, Bing Liu, and Lei Zhang (2009), "Entity Discovery and Assignment for Opinion Mining Applications," Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2009, 1125-34.

Dörre, Jochen, Peter Gerstl, and Roland Seiffert (1999), "Text mining: finding nuggets in mountains of textual data," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. San Diego, California, United States: ACM.

Dwyer, Paul (2009), "Measuring Interpersonal Influence in Online Conversations," MSI working paper

Edwards, Ralph I., and Jeffery K. Aronson (2000), "Adverse Drug Reactions: Definitions, Diagnosis, and Management," *Lancet*, 356 (9237), 1255-59.

Eliashberg, Jehoshua, Sam K. Hui, and John Z. Zhang (2007), "From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts," *Management Science*, 53 (6), 881.

Eliashberg, Jehoshua, Jedid J. Jonker, Mohanbir S. Sawhney, and Bernard Wierenga (2000), "MOVIEMOD: An Implementable Decision-Support System for Prerelease Market Evaluation of Motion Pictures," *Marketing Science*, 19 (3), 226-43.

Fan, Weiguo, Linda Wallace, Stephanie Rich, and Zhongju Zhang (2006), "Tapping the Power of Text Mining," *Communication of the ACM*, 49 (9), 76-82.

Feldman, Ronen, Moshe Fresko, Jacob Goldenberg, Oded Netzer, and Lyle Ungar (2007), "Extracting Product Comparisons from Discussion Boards," *Proceedings of the Seventh IEEE International Conference on Data Mining, 2007, ICDM 2007*, 469-74.

Feldman, Ronen, Moshe Fresko, Jacob Goldenberg, Oded Netzer, and Lyle Ungar (2008), "Using Text Mining to Analyze User Forums," *Proceedings of the Service Systems and Service Management, 2008 International Conference*, 1-5.

Feldman, Ronen, Suresh Govindaraj, Joshua Livnat and Benjamin Segal (2009) "Management's Tone Change, Post Earnings Announcement Drift and Accruals," *Review of Accounting Studies Journal*, forthcoming.

Feldman, Ronen, Moshe Fresko, Yakkov Kinar, Yehuda Lindell, Orly Liphstat, Martin Rajman, Yonatan Schler, and Oren Zamir (1998), "Text Mining at the Term Level," in *Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery*. New York: Springer-Verlag.

Feldman, Ronen and James Sanger (2006), *The text mining handbook*: NewYork: Cambridge University Press.

Fowdur, Lona, Vrinda Kadiyali, and Vishal Narayan (2009), "The Impact of Emotional Product Attributes on Consumer Demand: An Application to the US Motion Picture Industry," working paper, Cornell University.

Freitag, Dayne (2000), "Machine Learning for Information Extraction in Informal Domains," *Machine Learning*, 39 (2), 169-202.

Girvan, Michelle and Mark E. J. Newman (2002), "Community Structure in Social and Biological Networks," *Proceedings of the National Academy of Sciences*, 99 (12), 7821.

Godes, David and Dina Mayzlin (2004), "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science,* 23 (4), 545-60.

Godes, David and Dina Mayzlin (2009), "Firm-Created Word-of-Mouth Communication: Evidence from a Field Test," *Marketing Science*, forthcoming.

Godes, David, Dina Mayzlin, Yubo Chen, Sanjiv Das, Chrysanthos Dellarocas, Bruce Pfeiffer, Barak Libai, Subrata Sen, Mengze Shi, and Peeter Verlegh (2005), "The Firm's Management of Social Interactions," *Marketing Letters*, 16 (3), 415-28.

Guadagni, Peter M. and John D. C. Little (1983), "A Logit Model of Brand Choice Calibrated on Scanner Data," *Marketing Science*, 2 (3), 203-38.

He, Qin (1999), "Knowledge Discovery through Co-Word Analysis," *Library Trends*, 48 (1), 133-59.

Hu, Minging and Bing Liu (2004), "Mining and summarizing customer reviews," *in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining 2004*, 22-25.

Janis, Irving L. and Leon Mann (1977), *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment*. New York: Free Press.

Kamada, Tomihisa and Satoru Kawai (1989), "An Algorithm for Drawing General Undirected Graphs," *Information Processing Letters*, 31, 7-15.

Krackhardt, David (1988), "Predicting with networks: Nonparametric multiple regression analysis of dyadic data," *Social Networks*, 10 (4), 359-81.

Lee, Thomas (2007), "Constraint-based Ontology Induction from Online Customer Reviews," *Group Decision and Negotiation*, 16 (3), 255-81.

Lee, Thomas Yupoo and Eric Bradlow (2008), "Automatic Construction of Conjoint Attributes and Levels from Online Customer Reviews," Working Paper, Wharton School.

Liu, Bing, Minqing Hu, and Junsheng Cheng (2005), "Opinion observer: analyzing and comparing opinions on the Web," *in Proceedings of the 14th international conference on World Wide Web*. Chiba, Japan. May 10-14.

Liu, Yong (2006), "Word-of-Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of Marketing*, 70 (3).

Mayzlin, Dina (2006), "Promotional Chat on the Internet," *Marketing Science*, 25 (2), 155-63.

McCallum, Andrew and Ben Wellner (2005), "Conditional models of identity uncertainty with application to noun coreference," *Advances in Neural Information Processing Systems*, 17, 905-12.

Oberndorf, Shannon (2000), "When is a Virus a Good Thing?," Catalog Age, 17 (1), 43-44.

Pang, Bo and Lillian Lee (2008), "Opinion Mining and Sentiment Analysis," *Information Retrieval*, 2 (1-2), 1-135.

Pai, Seema and S. Siddarth (2009), "The Impact of Online Buzz on Purchase Decisions: The Case of

Motion Pictures," working paper, Boston University.

Rao, Vithala R. and Darius Jal Sabavala (1981), "Inference of Hierarchical Choice Processes from Panel Data," *The Journal of Consumer Research*, 8 (1), 85-96.

Ray, Soumya and Mark Craven (2001), "Representing Sentence Structure in Hidden Markov Models for Information Extraction." *Proceedings of the Seventeen International Joint Conference on Artificial Intelligence (IJCAI-2001) Seattle, WA (2001)*, 1273-79.

Reichheld, Frederick F. and Thomas Teal (1996), *The loyalty effect: the hidden force behind growth, profits, and lasting value.* Boston: Harvard Business School Press.

Roden Dan (2008), "Principle of Clinical Pharmacology," in *Harrison's Manual of Medicine*, part 1, Chapter 5, editors (Anthony S. Fauci, Eugene Braunwald, Dennis L. Kasper, Stephen L. Hauser, Dan L. Longo, J. Larry Jameson, and Joseph Loscalzo). New York: McGraw-Hill.

Rosa, José Antonio, Jelena Spanjol, and Joseph F. Porac (2004), "Text-based Approaches to Marketing Strategy Research," in Christine Moorman and Donald R. Lehmann (eds.), A*ssessing Marketing Strategy Performance*, Cambridge, MA: Marketing Science Institute (MSI), 185-211.

Rousseau-Anderson, Jacqueline (2008), "American Technographics Consumer Benchmark Survey," Forrester Report.

Rzhetsky, Andrey, Ivan Iossifov, Tomohiro Koike, Michael Krauthammer, Pauline Kra, Mitzi Morris, Hong Yu, Pablo Ariel Dubou,, Wubin Weng, W. John Wilbur, Vasileios Hatzivassiloglou, and Carol Friedman (2004), "GeneWays: A System for Extracting, Analyzing, Visualizing, and Integrating Molecular Pathway Data," *Journal of Biomedical Informatics*, 37 (1), 43-53.

Saiz, Albert and Uri Simonsohn (2007), "Downloading Wisdom from Online Crowds," Working Paper, The Wharton School, University of Pennsylvania.

Schindler, Robert M. and Barbara Bickart (2005), "Published Word of Mouth: Referable, Consumer-Generated Information on the Internet," *Online Consumer Psychology: Understanding And Influencing Consumer Behavior in the Virtual World*, 2, 35-60.

Seshadri Tirunillai and Gerard J. Tellis (2009), "Does Chatter Matter? The Impact of Online Consumer Generated Content on a Firm's Financial Performance." Working Paper, Marshall School of Business.

Shafir, Eldar, Itamar B. Simonson, and Amos B. Tversky (1993), "Reason-based choice," Cognition, 49.

Shin, Hyun S., Dominique M. Hanssens, and Bharath Gajula (2008), "The Impact of Positive vs. Negative Online Buzz on Retail Prices," Working Paper. Los Angeles: Anderson School of Management, UCLA.

Sifry, David (2008), "State of the Blogosphere."

Surowiecki, James (2004), *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Doubleday.

Swanson, Don R. (1988), "Migraine and Magnesium: Eleven Neglected Connections," *Perspectives in Biology and Medicine*, 31 (4), 526-57.

Swanson, Don R. and Neil R. Smalheiser (2001), "Information Discovery from Complementary Literatures: Categorizing Viruses as Potential Weapons," *Journal of the American Society for Information Science and Technology*, 52 (10), 797-812.

Urban, Glen L. and John R. Hauser (2004), "'Listening In' to Find and Explore New Combinations of Customer Needs," *Journal of Marketing*, 68 (2), 72-87.

Verlegh, Peeter C., Celine Verkerk, Mirijam A. Tuk, and Ale Smidts. (2004). “Customers or Sellers? The Role of Persuasion Knowledge in Customer Referral,” *Advances in Consumer Research*, 31, 304-5.

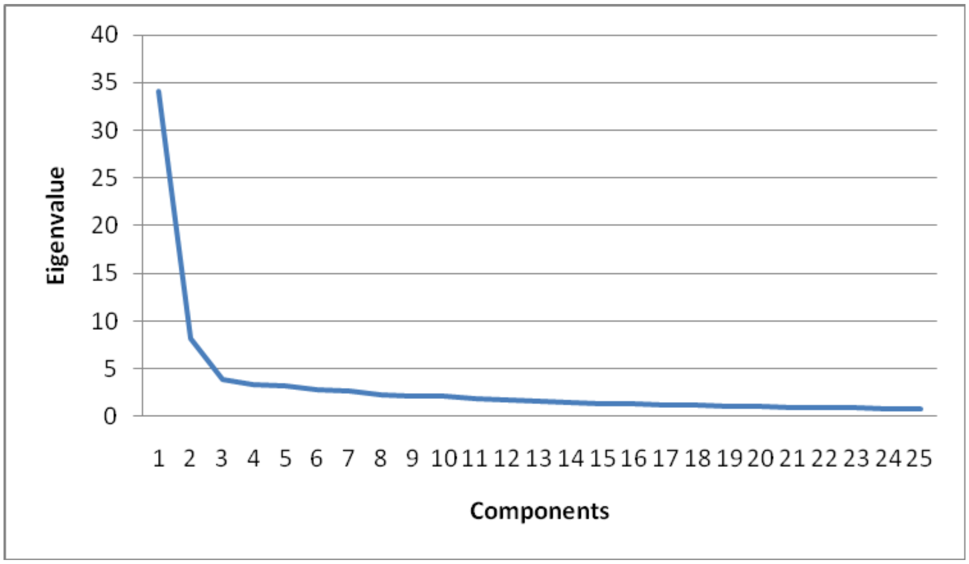
Appendix – Factor Analysis of Terms Most Frequently Mentioned in the Forum

Table A1 – Rotated Component Matrix – Three Factors Solution

Term	Factor 1	Factor 2	Factor 3
automatic	0.2684133	0.167272897	0.513797445
quality	0.427513389	0.361978222	0.521163941
standard	0.49680336	0.243013892	0.537379313
lower	0.485960721	0.250731767	0.538432382
mileage	0.036572491	0.449452909	0.548574211
small	0.473442418	0.065067734	0.551527016
mpg	-0.064388462	0.220402116	0.558649699
door	0.013221097	0.325408636	0.646351952
engine	0.031471638	0.26010346	0.649754098
gas	-0.020838514	0.167226504	0.668771316
fuel	-0.047205762	0.246219576	0.675102217
power	0.16911637	0.094915575	0.701702258
previous	0.495032283	0.537236465	0.077186726
wife	0.06919454	0.539747512	-0.032743914
warranty	0.092168251	0.540080803	0.433783178
sold	0.505139947	0.557649858	0.25726266
love	0.436383638	0.570839794	0.226761714
owners	0.504554289	0.597933852	0.244979838
dealer	0.261006117	0.622890825	0.343088254
old	0.246737806	0.654575006	0.159993735
experience	0.466817416	0.654689316	0.182199134
service	0.228258849	0.672197034	0.29957159
problems	0.082910231	0.697561119	0.281150888
owner	0.176337794	0.703040095	0.139892
own	0.272120718	0.720446334	0.339276103
owned	0.08523827	0.77216885	-0.000103129
purchased	-0.079646173	0.774075393	0.243013287
problem	-0.123782892	0.802608512	0.355070451
miles	-0.227014212	0.862250672	0.13495326
drove	0.506759319	0.215386257	0.261856471
big	0.513133383	0.374194717	0.236566455
dealers	0.527598659	0.16939428	0.349794773
pretty	0.532440989	0.303806731	0.303620783
lease	0.534532388	-0.052263447	-0.077815319
far	0.546151488	0.340532653	0.486441643
value	0.551368363	0.193238576	0.320030059
great	0.571395139	0.427280295	0.358976406
driven	0.581735436	0.188062091	0.263923599
styling	0.589863714	0.119958723	-0.028643369
equipped	0.594602384	-0.003247618	0.227529199
room	0.597103481	0.286124899	0.23726969
brand	0.598924375	0.24082405	0.20742975
larger	0.614961159	0.212783063	0.436448405
design	0.627580778	0.283223709	0.210096352
feel	0.641909952	0.309447872	0.41343176
cost	0.642547238	0.27794094	0.439775496
interior	0.643717472	0.038672307	0.335938481
sedan	0.644446391	0.022122006	0.291064446
reliability	0.645655398	0.451810333	0.059857018
offer	0.65077304	0.15695849	0.18077534
saying	0.652494961	0.14877629	0.249126292
manual	0.656128181	0.174091387	0.247452413

Term	Factor 1	Factor 2	Factor 3
luxury	0.665316148	0.13757115	-0.11685764
higher	0.669650349	0.141462447	0.424085773
performance	0.678056766	-0.035496258	0.234300939
premium	0.685582579	-0.006967972	0.128935246
size	0.686645678	0.278490094	0.026241451
smaller	0.68860204	0.14129586	0.454204233
sport	0.695722278	0.033613945	0.034022825
price	0.712064349	0.061313024	0.360384179
models	0.720255974	0.232403999	0.261732533
available	0.735343511	0.024396445	0.301009774
compared	0.743829787	0.18048389	0.041357115
bigger	0.744858847	0.129902563	0.334348619
best	0.746996356	0.263832319	0.280803228
expensive	0.785752919	0.136585675	0.302975781
handling	0.80440824	0.20531069	0.162856735
class	0.877335491	0.155180063	-0.035986887
generation	0.358420268	0.232037808	-0.116456827
platform	0.254935724	0.070694199	-0.028256094
tires	0.192920211	0.166944342	-0.015543526
reliable	0.282897098	0.40439535	-0.010424993
wait	0.46155203	-0.155394647	0.018276442
mid	0.263793692	0.107847575	0.040042897
coupe	0.466005378	-0.045051714	0.07847856
cyl	0.04664714	0.01762008	0.081537165
transmission	-0.019470164	0.476516719	0.170032465
fun	0.422211319	-0.042832503	0.179818505
heard	0.489736398	0.325232114	0.184053463
money	0.379661062	-0.101563919	0.19613595
cylinder	0.228776784	-0.080123207	0.211979664
torque	0.137099204	-0.278154295	0.239049879
deal	0.4381308	0.289380561	0.24102299
base	0.107722009	0.43835775	0.269765939
suspension	0.413343624	0.293559531	0.272545922
wheel	0.339285183	0.053534345	0.282874231
family	0.159999779	0.289346454	0.288628373
worth	0.433995358	0.331693455	0.293324911
features	0.459649535	0.250175725	0.297216589
rear	0.366507201	0.386590092	0.298665212
speed	0.229565305	0.115427577	0.345907259
loaded	0.21173375	-0.041513003	0.372239081
fit	0.328882042	0.254554304	0.389025159
bad	0.239732744	0.321630907	0.405803302
seats	0.376081108	0.327877136	0.418936959
front	0.191916636	0.417416435	0.444753826
dealership	0.273033653	0.422708902	0.45488407
steering	0.204023617	0.276724569	0.468266891
hp	0.283192633	-0.288032812	0.470006805
leather	0.237492393	0.2531634	0.484637997

Figure A1 – Factor Analysis Scree Plot


Eigenvalue
Components